

PROJECTE FINAL DE CARRERA



Sistema d'estructuració automàtica de llibreries digitals

Autor: Ruben Soriano Nos

Titulació: Enginyer Tècnic en Informàtica de Gestió

Curs: 2005 – 2006

Directors projecte: Dr. Antonio Moreno i Professor David Sánchez

**Universitat Rovira i Virgili (URV)
Escola Tècnica Superior d'Enginyeria (ETSE)
Grup de Sistemes Multi – Agents (GRUSMA)**

ÍNDIX

1- INTRODUCCIÓ	4
1.1- Objectius del projecte.....	6
1.2- Estructuració de la documentació. Capítols i continguts.	7
2- DESCRIPCIÓ DE LA TECNOLOGIA UTILITZADA EN EL PROJECTE	8
2.1- Extracció del text d'una pàgina <i>web</i>	10
2.1.1- Extracció.....	10
2.1.2- Transformació	11
2.1.3- Utilitat en el projecte	11
2.2- Processament del Llenguatge Natural (PLN).....	11
2.2.1- <i>Snowball</i>	13
2.2.1.1- Utilitat en el projecte	13
2.2.2- <i>OpenNLP</i>	14
2.2.2.1- Utilitat en el projecte	15
2.3- GoogleApi	15
2.3.1- Peticions de cerca	15
2.3.2- Utilitat en el projecte	16
2.4- Tecnologia <i>web</i>	16
3- EL SISTEMA	17
3.1- Motor de cerca.....	18
3.1.1- Anàlisi de pàgines <i>web</i>	20
3.1.2- Selecció de subdominis	24
3.1.3- Iteració del sistema.....	28
3.2- Estructures de dades	30
3.2.1- Estructura de dades elemental	30
3.2.2- Resultat de la cerca.....	30
3.2.3- Element de l'estructura.....	30
3.2.4- Estructura de resultats	31
3.2.5- Bases de dades.....	31
3.3- Llibreries digitals	32
3.3.1- Association for Computing Machinery. The Guide.....	33
3.3.2- PubMed	34
3.3.3- IEEE. Computer Society	34
3.3.4- NASA Astrophysics Data System.....	35
3.4- Interfície web.....	36
3.5- Manual d'usuari	38
3.5.1- Pàgina web d'inici	38
3.5.2- Pàgina web que interactua amb el motor de cerca de l'aplicació.....	39

4- JOC DE PROVES	44
4.1- Prova 1	45
4.2- Prova 2	47
4.3- Prova 3	49
4.4- Prova 4	51
4.5- Prova 5	52
4.6- Prova 6	53
4.7- Prova 7	54
4.8- Prova 8	56
4.9- Prova 9	70
4.10- Prova 10	71
4.11- Prova 11	72
5- CONCLUSIONS I TREBALL FUTUR	73

1- INTRODUCCIÓ

Internet és una xarxa mundial amb moltíssima informació sobre molts de temes diferents. Aquesta informació però, es troba molt desestructurada i molts cops, aconseguir el que és busca, pot ocupar molt de temps o simplement pot ser impossible de trobar.

Quan nosaltres volem buscar informació sobre algun tema, gairebé ho fem sempre amb l'ajuda de cercadors tipus *Google* o *Yahoo*. Des de fa pocs anys, el *Google* [1] s'ha tornat un dels cercadors per excel·lència. Aquest sistema busca per tota la *web* pàgines relacionades amb el concepte o conceptes introduïts per l'usuari. Aproximadament, és capaç de buscar entre unes 8000 milions de pàgines *web*, donant una ràpida resposta gràcies a la utilització de *caches*. Com hem dit abans però, la informació que retorna el *Google* és moltíssima i desestructurada, i és l'usuari final el que ha de seleccionar el que realment li interessa.

Per tant, la quantitat d'informació que existeix a la *web* és enorme i disposar d'aplicacions o eines que ens permetin disposar d'aquesta d'una forma més estructurada, és molt necessari.



Figura 1. Estructuració de la informació retornada pel *Clusty*, a partir del terme “cancer”.

Actualment existeixen en la *web* aplicacions que realitzen la funció de cercadors taxonòmics, com poden ser el *Clusty* [2] o *Vivisimo* [3], entre moltes altres. Aquestes aplicacions s'encarreguen de cercar per tota la *web* informació relativa al terme o termes introduïts per l'usuari. La seva principal tasca, que els diferencia dels cercadors tipus *Google*, és que ofereixen com a resultat una jerarquia que mostra estructuradament una sèrie de subdominis o subclasses del concepte introduït per l'usuari, és a dir, es realitza una mena d'estructuració de la informació.

L'avantatge d'aquest sistema és que a partir d'aquesta jerarquia, l'usuari pot veure ràpidament i de forma clara els subdominis relacionats directament amb el terme que ha introduït, de forma que pot ajudar-lo d'una forma significativa, en la cerca de la informació final que ell necessita.

Per posar un exemple, imaginem que es desitja cercar informació sobre el terme “càncer”. Doncs bé, utilitzant una aplicació com *Clusty*, aquesta ens retornarà una taxonomia (jerarquia) amb una sèrie de subdominis com poden ser càncer de pulmó, càncer de pròstata, càncer de pell,..., per exemple, i molts altres tipus d'informació relacionada amb el càncer.

Recordem que una taxonomia és una classificació o estructuració, de tota la informació referent a un concepte determinat, en diferents grups o categories semàntiques.

Quin és el problema principal però, d'utilitzar aquests sistemes? Bé, tot i l'ajuda que ens ofereixen aquestes aplicacions en la construcció de jerarquies a partir de cerques taxonòmiques, encara retornen una quantitat de resultats potencialment molt elevada, i a cops encara desestructurada.

En la *Figura 1*, es mostra el resultat de realitzar la cerca del terme “*cancer*”, en el *Clusty*. Com es pot observar, els resultats són encara molt desestructurats.

Aquests sistemes busquen subdominis del terme introduït per tota la *web*. Això significa que la majoria de vegades, aquests subdominis seran en general de molt baixa qualitat, ja que les fonts de les quals s'han extret són moltes i variades i el tema central d'aquestes pot no coincidir amb el tema del terme a partir del qual es desitja construir la taxonomia.

Deixant de banda els cercadors d'Internet com a cerca de fonts d'informació, existeixen en la *web* moltes bases de dades o llibreries digitals d'articles o ‘*papers*’ sobre temes d'estudi molt variats. Aquestes fonts d'informació acostumen a ser bastant més fiables i estan molt més estructurades que les que existeixen en la resta de la *web*. Per tant, poden ser de gran ajuda per a l'usuari que està buscant informació sobre un tema determinat.

Les llibreries digitals són bases de dades molt grans que contenen multitud d'articles de diferents autors, anys, temes,... Aquestes llibreries acostumen a disposar d'un motor de cerca intern que és capaç de buscar aquests articles segons una sèrie de paràmetres diferents, com poden ser l'autor o autors, l'any de la seva publicació, l'editorial,...

Llavors, és aquí on entra en joc l'aplicació que s'ha desenvolupat en aquest projecte.

La nostra aplicació, de funcionament similar als cercadors tipus *Clusty*, s'encarrega d'estructurar la informació o recursos existents en una llibreria digital. Aquesta informació farà referència a un o més conceptes que l'usuari ha d'haver introduït prèviament.

Ha de quedar ben clar que l'aplicació no té per funció la construcció de taxonomies, sinó l'estructuració en forma de taxonomia, de la informació existent en una llibreria digital o repositori sobre un terme determinat. En el cas que la llibreria sigui suficientment gran i amb una bona cobertura per al terme en concret que s'ha introduït, el resultat generat pel sistema potser si es podria considerar una taxonomia del domini.

Llavors, quina és la principal diferència entre la nostra aplicació i les altres? Doncs bé, la principal diferència són les fonts d'informació de les quals s'extrauran els subdominis.

És a dir, no es tracta de buscar per tota la *web*, sinó que ara els subdominis s'extrauran de repositoris d'articles o llibreries digitals, de les quals ja hem parlat en un dels paràgrafs anteriors. Es tracta d'una eina especialment dissenyada per a estructurar automàticament aquestes llibreries.

D'aquesta forma es vol aconseguir una major qualitat dels subdominis aconseguits en el procés. Si per exemple, ens interessa cercar el terme “*càncer*”, perquè buscar en la immensitat de la *web*, podent buscar solament per una llibreria digital que conté sobre 16 milions d'articles relacionats amb la medicina? Potencialment, els resultats obtinguts

seran molt millors, ja que tot i buscar solament en una font d'informació, aquesta serà fiable i de bona qualitat.

A part de la qualitat de les fonts d'informació, el procés d'estructuració d'aquesta serà relativament ràpid, ja que s'analitzaran directament abstracts d'articles.

Un altre aspecte important a tenir en compte sobre aquest sistema, és que pot ajudar als gestors de llibreries electròniques a estructurar automàticament els seus articles a partir d'un domini determinat.

I ja per finalitzar, s'hauria de destacar que tot i que el disseny i implementació del sistema s'ha realitzat des de zero, s'ha utilitzat com a precedent, l'experiència generada en la realització del projecte fi de carrera de Laura Prieto Rebollo [4], i del projecte de tesi del professor David Sánchez [5], director del present projecte.

En el següent apartat es descriuran detalladament els objectius que es pretenen aconseguir en el projecte.

1.1- Objectius del projecte

A continuació s'enumeren els objectius principals que es pretenen aconseguir en aquest projecte.

- Construir un motor de cerca que sigui capaç d'extreure una sèrie de subdominis concrets d'un terme o termes introduïts per l'usuari, tenint com a fonts d'informació les llibreries digitals.
- El motor de cerca ha de ser capaç de generar una estructuració de la informació i dels recursos, en forma taxonòmica a partir dels subdominis aconseguits anteriorment en la llibreria digital. La complexitat d'aquesta estructura, és a dir, el número total de subdominis que la compondran, dependrà de si el terme clau introduït per l'usuari té molta o poca relació amb el tema dels articles que componen la llibreria.
- L'aplicació ha de funcionar sobre més d'un repositori d'articles diferent, de forma que l'usuari podrà escollir entre més d'una llibreria digital per a estructurar la informació.
- Finalment serà necessària una interfície gràfica de forma que l'usuari pugui interaccionar amb el sistema passant-li els paràmetres necessaris, i obtenint per aquesta mateixa via, els resultats retornats per l'aplicació.

1.2- Estructuració de la documentació. Capítols i continguts.

En aquest document s'intenta explicar amb detall el procés utilitzat per a aconseguir els principals objectius abans mencionats. La tasca no és trivial, i per a aconseguir que el lector entengui tots els passos que es duen a terme, és necessària una bona estructuració de la documentació de forma que els continguts es vagin mostrant pas a pas.

El capítol 2 s'ha reservat per a explicar les tecnologies utilitzades en el nostre sistema. En ell es veuran les eines o llibreries utilitzades així com les característiques més importants de cadascuna. També es farà referència a la utilitat d'aquestes en el disseny de la nostra aplicació.

El capítol 3 és un dels més importants de la documentació i en ell s'explica molt detalladament el sistema en sí. Al principi es fa una petita introducció dels passos que segueix l'aplicació per a portar a terme les seves funcions. I a continuació s'explica a fons cadascun d'aquest passos. Com hem dit abans en els objectius principals, el projecte constarà de dues parts molt importants. Una serà el motor de cerca de l'aplicació i l'altra la interfície *web* a partir de la qual interaccionarà l'usuari. Doncs bé, aquestes dues parts seran explicades exhaustivament en aquest capítol.

En el capítol 4 ens dedicarem a realitzar una sèrie de joc de proves de l'aplicació de forma que podrem observar els resultats que retorna l'aplicació depenent dels paràmetres introduïts per l'usuari.

El capítol 5 contindrà les conclusions del projecte. Aquest capítol contindrà un petit resum del treball, les diverses aplicacions que se li poden donar al projecte, així com possibles treballs futurs que es podrien realitzar per a millorar-lo.

2- DESCRIPCIÓ DE LA TECNOLOGIA UTILITZADA EN EL PROJECTE

En aquest capítol farem referència a la tecnologia utilitzada en la construcció de l'aplicació.

El sistema es divideix en dues parts molt importants, que tot seguit explicarem.

→ La part més important és el **motor de cerca**, el qual s'encarrega de parserjar les pàgines *web* d'articles que retornarà la llibreria digital, seleccionar els possibles subdominis del terme introduït per l'usuari, determinar quins d'aquests subdominis seran finalment bons candidats, i per últim generar l'estructuració de la informació en forma de taxonomia, de forma que l'usuari pugui interactuar amb ella.

Aquesta part ha estat dissenyada íntegrament amb Java (llenguatge orientat a objectes i desenvolupat per Sun Microsystems [6]), degut a que es tracta d'un llenguatge molt potent, molt flexible i enfocat a entorns multiplataforma. A part, moltes de les llibreries de codi lliure disponibles a la *web*, i que també utilitzem en l'aplicació, estan codificades amb Java.

Com que una de les activitats principals del motor de l'aplicació serà parsejar planes *web* que continguin els articles retornats per les llibreries digitals, ens serà necessària una eina que ens permeti realitzar aquesta tasca.

Les pàgines *web* contenen molts tipus d'informació, com poden ser, *banners*, imatges, títols, animacions *flash*, text,... Doncs bé, molta d'aquesta informació no ens interessa a nosaltres. Per dir-ho d'alguna forma, a nosaltres solament ens interessarà el text que hi ha a les pàgines *web*, més concretament el text dels articles. Per tant, en aquest sistema hem decidit utilitzar una eina molt important anomenada *HTMLParser* [7] que ens soluciona aquest tipus de problema, i que a més està inclosa en el *sourceforge* [8], és a dir, es tracta de software de lliure distribució.

Doncs bé, el text que s'extreu amb el *HTMLParser*, s'ha de poder manipular. Per tant, per a la nostra aplicació, serà necessària la utilització de tecnologies que ens permetin tractar aquest text de manera eficaç. Les eines o llibreries que utilitzen aquest tipus de tecnologies s'anomenen eines de Processament de Llenguatge Natural (PLN), amb anglès, *Natural Language Processing (NLP)*.

Més concretament, en la nostra aplicació utilitzarem una eina anomenada *Snowball* [9], que permet convertir una paraula qualsevol al seu corresponent terme arrel. També utilitzarem com a eina de PLN, l'*OpenNLP* [10]. Aquesta llibreria permet realitzar un anàlisi morfològic de cada paraula, de forma que podem saber si es tracta d'un nom, un verb, un adjectiu,... Més avant veurem per a que s'utilitzen aquestes eines en el nostre sistema.

S'han decidit utilitzar aquestes eines perquè desenvolupen la seva activitat correctament i a més, pertanyen al grup de *software* de lliure distribució.

I per últim, una altra de les eines utilitzades en el motor de cerca de l'aplicació serà l'*API* de *Google*. Aquest API es refereix a una llibreria escrita en Java que permet

realitzar cerques a la *web* de la mateixa manera que ho podem fer des de la pàgina principal del *Google*.

→ L'altra, i no pas menys important part del sistema, és la **interfície gràfica**. Una interfície gràfica, amb anglès *Graphic User Interface (GUI)*, és un mètode per a facilitar la interacció de l'usuari amb la computadora (en el nostre cas amb el motor de cerca) a través de la utilització d'imatges i objectes pictòrics (icones, finestres, botons, camps de text, animacions,...), a més del propi text.

Hi ha moltes tecnologies disponibles per a dissenyar interfícies gràfiques d'usuari. Per anomenar-ne algunes d'importants podem fer referència al *Swing* [11], que correspon a una biblioteca gràfica per a Java i que forma part de les *Java Foundation Classes (JFC)*. Uns altres tipus d'interfícies gràfiques, i molt usuals, són les pàgines o aplicacions *web*. Les tecnologies més esteses per a la seva construcció són la utilització d'*HTML*, *PHP*, *ASP*, *JSP*, *servlets*,...

Unes de les tecnologies *web* que s'han utilitzat en el projecte han estat els *JSP*'s i els *servlets*. Aquestes tecnologies poden treballar perfectament de forma conjunta donant uns resultats excel·lents en el disseny, tant d'aplicacions com d'entorns *web*.

El motor del nostre sistema, com hem dit abans, ha estat dissenyat íntegrament amb codi Java. Per tant s'ha decidit construir la interfície gràfica amb un llenguatge, que a més de ser dinàmic, ens permeti també interactuar de forma senzilla amb el motor de cerca, i finalment s'ha escollit el llenguatge *JSP (Java Server Pages)*.

Aprofitant que parlem de les eines que s'utilitzen en el projecte seria interessant anomenar l'editor que s'ha escollit per a la codificació de l'aplicació. Es tracta de "*Eclipse Platform v.3.0.1*" [12]. S'ha decidit utilitzar aquest editor perquè és molt potent, molt flexible, està molt ben estructurat, permet agregar llibreries d'una forma molt intuïtiva i minimitza molt la feina del programador. A més, permet expandir les seves funcions, amb la possibilitat d'agregar-li *plug-in*'s.

I ja per últim, solament quedaria anomenar el *Jakarta Tomcat v.4.1.31* [13]. El *Jakarta Tomcat* no és res més que un programa que realitza la funció de servidor, sobre el qual poden executar-se aplicacions *web*. S'ha escollit aquest servidor per ser gratuït i per la facilitat que dóna a l'hora d'incorporar-li noves aplicacions *web* creades pels programadors. A part, es reinicia ràpidament en el cas en que es produeixin canvis de codificació en les classes Java de l'aplicació.

A continuació, i abans d'endinsar-nos en el disseny del sistema, donarem un cop d'ull a les característiques més importants de cadascuna de les eines utilitzades.

2.1- Extracció del text d'una pàgina web

HTMLParser és una llibreria de Java que s'utilitza per a analitzar *HTML* a temps real. Es tracta de codi *OpenSource* (codi d'accés lliure). El que ha atret del *HTMLParser* a la majoria de programadors ha estat seva simplicitat en el disseny, la velocitat i l'habilitat per a manejar *HTML*.

Les seves dos utilitats fonamentals són l'*extracció* i la *transformació*, i conté filtres, *visitors*, etiquetes pròpies i *JavaBeans* fàcils d'utilitzar. És un paquet ràpid, robust i ben testat.

En general, per a utilitzar el *HTMLParser* sols és necessari ser capaç d'escriure codi en llenguatge Java. Encara que es proporcionen alguns exemples de programes que poden ser molt útils, és més probable que necessitem (o vulguem) crear els nostres propis programes o modificar els previstos per a aconseguir l'aplicació desitjada.

Per a utilitzar la llibreria s'ha d'afegir l'*htmllexer.jar* o *htmlparser.jar* al *classpath* quan compilem i executem. El *htmllexer.jar* proporciona accés a baix nivell a *strings* genèrics, comenta i etiqueta nodes d'una pàgina de forma lineal, plana i seqüencial. El *htmlparser.jar*, que inclou les classes de l'*htmllexer.jar*, proporciona accés a una pàgina com una seqüència d'etiquetes diferenciades que contenen nodes amb *strings*, comentaris i altres etiquetes.

El *parser* tracta d'anivellar etiquetes d'obertura amb etiquetes de tancament per presentar l'estructura de la pàgina, mentre que el *lexer* simplement separa nodes. Si la nostra aplicació sols requereix coneixement estructural modest de la pàgina i tracta primordialment amb nodes individuals i aïllats, deuríem considerar la utilització del *lexer*. Però si la nostra aplicació requereix coneixement de la estructura en conjunt de la nostra pàgina, per exemple processament de taules, probablement necessitem el *parser*.

2.1.1- Extracció

L'extracció cobreix tots els programes de recuperació d'informació que no serveixen per a conservar la pàgina font. S'utilitza per a:

- Extracció de text, per a utilitzar-lo, per exemple, com a entrada de bases de dades de motors de cerca de textos.
- Extracció de *links*, per moure's a través de les pàgines *web* o aconseguir direccions *e-mail*.
- *Screen scraping*, per a entrada de dades programades de pàgines *web*.
- Extracció de recursos, recollint imatges i sons.
- Comprovació de *links*, per a assegurar-se que són vàlids.
- Control de *site*, comprovant diferències de pàgines més enllà de diferències simplistes.

Hi ha varies facilitats en el codi base del *HTMLParser* per a ajudar amb l'extracció, incloent filtres, *visitors* i *JavaBeans*.

2.1.2- Transformació

La transformació inclou tot el procés on l'entrada i la sortida són pàgines d'*HTML*.

- Reescriptura d'*URL*'s, modificant alguns o tots els *links* d'una pàgina.
- Captura de *site*, movent contingut de la *web* al disc local.
- Censura, eliminant paraules i frases d'ofensa de les pàgines.
- Neteja d'*HTML*, corregint pàgines errònies.
- Eliminació d'anuncis (banners), eliminant *URL*'s que fan referència a publicitat.
- Conversió a XML, transformant pàgines *web* existents a *XML*.

2.1.3- Utilitat en el projecte

Utilitzarem aquesta eina, com hem dit abans, per a parsejar les pàgines *web* d'articles que ens retornin les llibreries digitals. La extracció del text es fa a partir de la *URL* de la plana *web* que es vol parsejar.

Mitjançant una classe *TextExtractingVisitor* i utilitzant un mètode d'extracció, obtindrem tot el text en forma plana de forma que el podrem emmagatzemar en un *string* per a la seva manipulació posterior.

2.2- Processament del Llenguatge Natural (PLN)

El 'Processament del Llenguatge Natural (PLN)', o '*Natural Language Processing (NLP)*' amb anglès, és una subdisciplina de la Intel·ligència Artificial i de la branca d'enginyeria de la lingüística computacional. El PLN s'ocupa de la formulació i investigació de mecanismes eficaços computacionalment per a la comunicació entre persones o entre persones i màquines per mitjà dels llenguatges naturals. El PLN no s'encarrega de la comunicació per mitjà de llenguatges naturals d'una forma abstracta, sinó de dissenyar mecanismes per a comunicar-se que siguin eficaços computacionalment, és a dir, que es puguin realitzar per mitjà de programes que executin o simulin la comunicació. Els models aplicats s'enfoquen, no sols a la comprensió del llenguatge per sí, sinó a aspectes generals cognitius humans i a la organització de la memòria. El llenguatge natural sols serveix de mitjà per a estudiar aquests fenòmens.

L'arquitectura bàsica dels sistemes PLN és la següent. Seguirem com a oració d'exemple "el gatocomepescado":

- Reconeixement de l'idioma.
- Segmentació de paraules, oracions i seccions:
→ "el gato come pescado"
- Anàlisi morfològic (*tagging*):
 - Detecció de paraules especials.
 - Assignació d'etiquetes a aquestes paraules:
→ Determinant: "el". Nom: "gato". Verb: "come". Nom: "pescado".

- Anàlisi sintàctic (*chunking* / *parsing*).
 - Constituents bàsics o *chunks*:
 - Grup nominal: “el gato”. Grup verbal: “come”. Grup nominal: “pescado”.
 - Estructura oració:
 - SN (subjecte): “el gato”. SV (predicat): “come pescado”.
- Anàlisi semàntic
 - Lèxic:
 - Entitat -> ésser viu -> animal: “el gato”.
 - Acció -> voluntària: “come”
 - Entitat -> inanimat -> natural || Entitat -> ésser viu -> animal -> vertebrat -> peix: “pescado”
 - Preposicional:
 - $\exists \text{ gato } (X) \wedge \exists \text{ pescado } (Y) \wedge \text{come } (X,Y)$

A continuació explicarem una eina de PLN anomenada *Snowball*. Abans però, hem de comentar dos conceptes amb els que està molt relacionada: recuperació de dades i *Stemming*.

→ Recuperació de dades

La recuperació de dades (*IR*) és essencialment una qüestió de decidir quins documents dins d'una col·lecció s'han de recuperar per a satisfer la necessitat d'informació d'un usuari. Aquesta necessitat d'informació es representa per una consulta o un perfil, i conté un o més termes de cerca, a més, d'una certa informació addicional com graus d'importància. Per tant, la decisió de la recuperació s'agafa comparant els termes de la consulta amb els termes índex (les paraules o frases importants) que apareixen en el propi document. La decisió pot ser binària (recuperar / desestimar), o pot implicar la estimació del grau d'importància que té el document amb relació a la consulta.

Desafortunadament, les paraules que apareixen en documents i en consultes tenen a sovint moltes variables morfològiques. Així els parells de termes tals com *computing* i *computation* no seran reconeguts com a equivalents sense un cert tipus de procés de llenguatge natural (NLP).

→ *Stemming*

En la majoria dels casos, les variants morfològiques de paraules tenen interpretacions semàntiques similars i es poden considerar com a equivalents per al propòsit d'aplicacions d'*IR*. Per aquesta raó, s'han desenvolupat un nombre d'*stemming Algorithms* o *stemmers*, que tracten de reduir una paraula a un *stem* o arrel. Així, els termes clau d'una consulta o d'un document són representats per *stems* i no per les paraules originals. Això no sols significa que diverses variants d'un terme es poden reduir a una sola forma representativa, també redueix el tamany del diccionari, és a dir, el nombre de termes diferents que es necessiten per a representar un conjunt de documents. Un tamany menor del diccionari dona lloc a un estalvi de l'espai d'emmagatzematge i del temps de processament.

Per als propòsits d'IR, no importa generalment si els *stems* generats són autèntiques paraules o no (així, *computation* podria ser reduït a *comput*) a condició de que diverses paraules amb el mateix significat base es redueixin a la mateixa forma i que les paraules amb significats diferents es mantinguin separades. Un algoritme que tracta de convertir una paraula a la seva arrel lingüísticament correcta (*compute* en aquest cas) es sol anomenar un *lemmatizer*.

Els *stemmers* i *lemmatizers* també s'utilitzen extensament dins el camp de la lingüística computacional.

Un exemple algoritme d'*stemmer* és el de *Porter stemming algorithm* (o *Porter stemmer*), creat per Martin Porter en 1980 [14]. És un procés per a eliminar les terminacions morfològiques més comunes de les paraules amb anglès. El seu ús principal és com a part d'un procés de normalització de termes que es fa normalment quan es creen sistemes de recuperació de dades.

L'algoritme s'ha utilitzat, cotitzat i adaptat extensament en els últims 20 anys. Desafortunadament, abunden variants seves que reclamen ser implementacions vertaderes, i això pot causar confusió.

2.2.1- *Snowball*

Snowball és una eina que serveix per a definir algoritmes d'*stemming* i a partir del qual es poden generar ràpidament programes *stemmer* en ANSI o Java.

Snowball es va crear per dos raons principals:

- La falta d'algoritmes d'*stemming* fàcilment disponibles per a idiomes diferents de l'anglès.
- La falta de promoció d'implementacions exactes del *Porter stemming algorithm*.

En aquest projecte s'ha utilitzat una versió precompilada dels *stemmers* d'*Snowball*. Per a veure com funciona *Snowball*, utilitzarem el següent exemple, on es mostra l'anàlisi de dos oracions:

→ Anàlisi de: "The quick brown fox jumped over the lazy dogs"

→ *SnowballAnalyzer*: [quick] [brown] [fox] [jump] [over] [lazi] [dog]

2.2.1.1- Utilitat en el projecte

Utilitzarem l'*Snowball* per a comprovar si dos paraules pertanyen a la mateixa família, és a dir, tenen la mateixa arrel. D'aquesta manera evitarem construir un arbre que contingui subdominis (subclasses) que signifiquin el mateix.

Per exemple, imaginem que el sistema troba els termes '*lung cancer*' i '*lungs cancer*'. Doncs bé, sense l'eina *Snowball*, s'introduirien com a bones les subclasses '*lung*' i '*lungs*', cosa que no hauria de passar, ja que en realitat es refereixen a la mateixa subclasse, l'únic que una és el plural de l'altra.

2.2.2- OpenNLP

OpenNLP és un centre d'organització per a projectes *open source*, relacionat en el 'natural language processing (NLP)'. El seu paper principal és animar i facilitar la col·laboració d'investigadors i programadors en tals projectes.

OpenNLP també té una varietat d'eines Java basades en eines *NLP*, les quals porten a cap la detecció d'oracions, 'tokenization', 'pos-tagging', 'chunking', 'parsing' i detecció de nom – entitat.

Per a utilitzar l'eina *OpenNLP* en una aplicació, prèviament s'han hagut de carregar a memòria els dos fitxers objectes següents:

- "*EnglishTok.bin*": s'encarrega de separar les paraules d'una frase en *tokens* independents. Aquest fitxer es carrega en una variable anomenada *EnglishTokenizerME*.
- "*EnglishPOS.bin*": s'encarrega de classificar sintàcticament els *tokens* obtinguts anteriorment. *POS (part-of-speech tagging)* → etiquetat de les parts d'una frase. Aquest fitxer es carrega en una variable anomenada *EnglishPOSTaggerME*.

A continuació es mostra una llista de *tags* en ordre alfabètic i les parts d'una oració corresponents a aquestes etiquetes:

- 1- CC Coordinating conjunction
- 2- CD Cardinal number
- 3- DT Determiner
- 4- EX Existential there
- 5- FW Foreign word
- 6- IN Preposition or subordinating conjunction
- 7- JJ Adjective
- 8- JJR Adjective_ comparative
- 9- JJS Adjective_ superlative
- 10- LS List item marker
- 11- MD Modal
- 12- NN Noun_ singular or mass
- 13- NNS Noun_ plural
- 14- NNP Proper noun_ singular
- 15- NNPS Proper noun_ plural
- 16- PDT Predeterminer
- 17- POS Possessive ending
- 18- PRP Personal pronoun
- 19- PRP_ Possessive pronoun
- 20- RB Adverb
- 21- RBR Adverb_ comparative
- 22- RBS Adverb_ superlative
- 23- RP Particle
- 24- SYM Symbol
- 25- TO to
- 26- UH Interjection

- 27- VB Verb_ base form
- 28- VBD Verb_ past tense
- 29- VBG Verb_ gerund or present participle

2.2.2.1- Utilitat en el projecte

Aquesta eina s'ha utilitzat en el projecte per a classificar sintàcticament les paraules.

Més concretament ens interessa que els subdominis escollits pel sistema siguin noms o adjectius, és a dir, no ens interessen els verbs com a possibles candidats a subclasses. En la llista anterior de les *tags*, s'han remarcat amb color groc les etiquetes dels termes que no es descartaran, és a dir, termes que seran seleccionats com a possibles candidats a subdominis.

2.3- GoogleApi

En el *Google* es poden fer cerques de dues formes diferents:

- Una forma és l'habitual que utilitza la majoria de la gent, que és des de la pròpia interfície *web* del *Google*, la qual busca per *caches* que s'actualitzen molt regularment.
- L'altra forma és mitjançant l'*API* de *Google* [15]. Aquesta eina permet realitzar cerques de conceptes de forma similar a l'anterior, i amb la possibilitat també d'utilitzar paràmetres de cerca. Aquesta possibilitat de cerca, però, utilitza una via diferent a l'anterior. També busca per *caches* repartides arreu del món. A aquestes però, no se'ls dona tanta prioritat com a les anteriors i són actualitzades amb menys regularitat, donant d'aquesta forma resultats menys actualitzats.

Un altre aspecte a tenir en compte és que per a utilitzar l'*API* de *Google* des d'un programa en Java, per exemple, s'ha d'utilitzar una clau (proporcionada per *Google*), la qual et permet realitzar fins a un màxim de 1000 cerques diàries.

A continuació s'expliquen algunes de les seves funcionalitats.

2.3.1- Peticions de cerca

Una petició de cerca envia el terme a buscar més un conjunt de paràmetres en forma d'*string* al servei de *Google Web Api*, i rep un conjunt de resultats de la cerca. Els resultats de la cerca s'obtenen de l'índex de *Google*, el qual disposa d'aprox. uns 2 milions de pàgines *web*.

→ Format de la petició de cerca (Paràmetres de la cerca)

La taula següent mostra tots els possibles paràmetres que es poden usar en una petició de cerca i descriu com aquests paràmetres modificaran els resultats de la cerca:

Nom	Descripció
key	Proporcionada per <i>Google</i> , es requereix per a l'accés al seu servei. <i>Google</i> usa la clau per a l'autenticació i el registre.
q	Especifica el terme (<i>query</i>) que es desitja buscar.
start	Índex a partir del qual es desitja que es retornen els resultats.
maxResults	Número màxim de resultats desitjats per cerca.
filter	Activa o desactiva filtres automàtics per als resultats, els quals oculten resultats molt similars i resultats que provenen del mateix servidor <i>web</i> . Els filtres s'encarreguen de millorar la experiència final sobre <i>Google</i> , però per a la teva aplicació pots preferir desactivar-los.
restrict	Restringeix les cerques a un subconjunt de l'índex de la <i>web</i> del <i>Google</i> , com pot ser a un país concret o a un tema relacionat amb <i>Linux</i> , per exemple.
safeSearch	Un valor booleà, el qual permet filtrar continguts adults en els resultats de la cerca.
lr	Restricció del llenguatge: Restringeix la cerca a documents amb una o més llengües determinades.

Per a més informació sobre l'API de *Google*, llegir-se el document “*APIs_Reference.html*”, el qual està present en el propi “*googleapi.zip*”, que es pot descarregar des de la pàgina oficial del *Google*, en l'apartat d'API's.

2.3.2- Utilitat en el projecte

L'API de *Google* s'utilitza en el projecte per a l'obtenció d'estadístics.

És a dir, utilitzarem l'API per a realitzar cerques de termes, de forma que a nosaltres solament ens interessaran el nombre de resultats trobats, no els resultats en sí.

2.4- Tecnologia web

Tal com hem dit en la introducció d'aquest capítol segon, la tecnologia *web* utilitzada en la nostra aplicació per a crear la interfície gràfica serà a partir de *JSP*'s i *servlets*.

Els *JSP* (*Java Server Pages*) són fitxers de codi que poden contenir tant codi *HTML*, com codi *Java*. Permeten la construcció dinàmica d'aplicacions *web*.

Els *servlets* són fitxers de codi, a l'igual que els *JSP*'s, que s'utilitzen també per a la programació d'aplicacions *web*. Per dir-ho d'alguna forma, els *JSP*'s tenen un alt contingut de codi *HTML* envers de codi *Java*. En canvi els *Servlets* són fitxers dissenyats majoritàriament en codi *Java* que fan la funció de processar la informació que reben dels *JSP*'s i que acaben mostrant també en fitxers *JSP*'s.

3- EL SISTEMA

Aquest capítol és el més important pel que fa a la documentació del projecte, i en ell s'explicarà detalladament el disseny de l'aplicació.

El sistema, a grans trets, es divideix en tres parts. Aquestes tres parts corresponen ni més ni menys que al motor de l'aplicació (també anomenat en aquest projecte, motor de cerca), a les estructures de dades (E.D.), i a la interfície gràfica d'usuari, i seran explicades de forma molt detallada en aquest capítol.

Es començarà explicant el nucli de l'aplicació, així com els paràmetres necessaris per al seu funcionament.

El següent pas serà fer referència a les estructures de dades utilitzades pel programa per a emmagatzemar les dades.

I ja per últim, s'acabarà parlant de les parts que formaran la interfície gràfica d'usuari, així com de cadascun dels components que la formen (camp de text, botons, imatges...), construïda, com hem dit abans, per *JSP's* i *servlets*.

3.1- Motor de cerca

En aquest apartat es parlarà de la part més important del sistema, el nucli o motor de cerca de l'aplicació.

En la *Figura 2*, es mostra un esquema gràfic on es poden veure, de forma general, els passos més importants que es duen a terme en el motor de cerca.

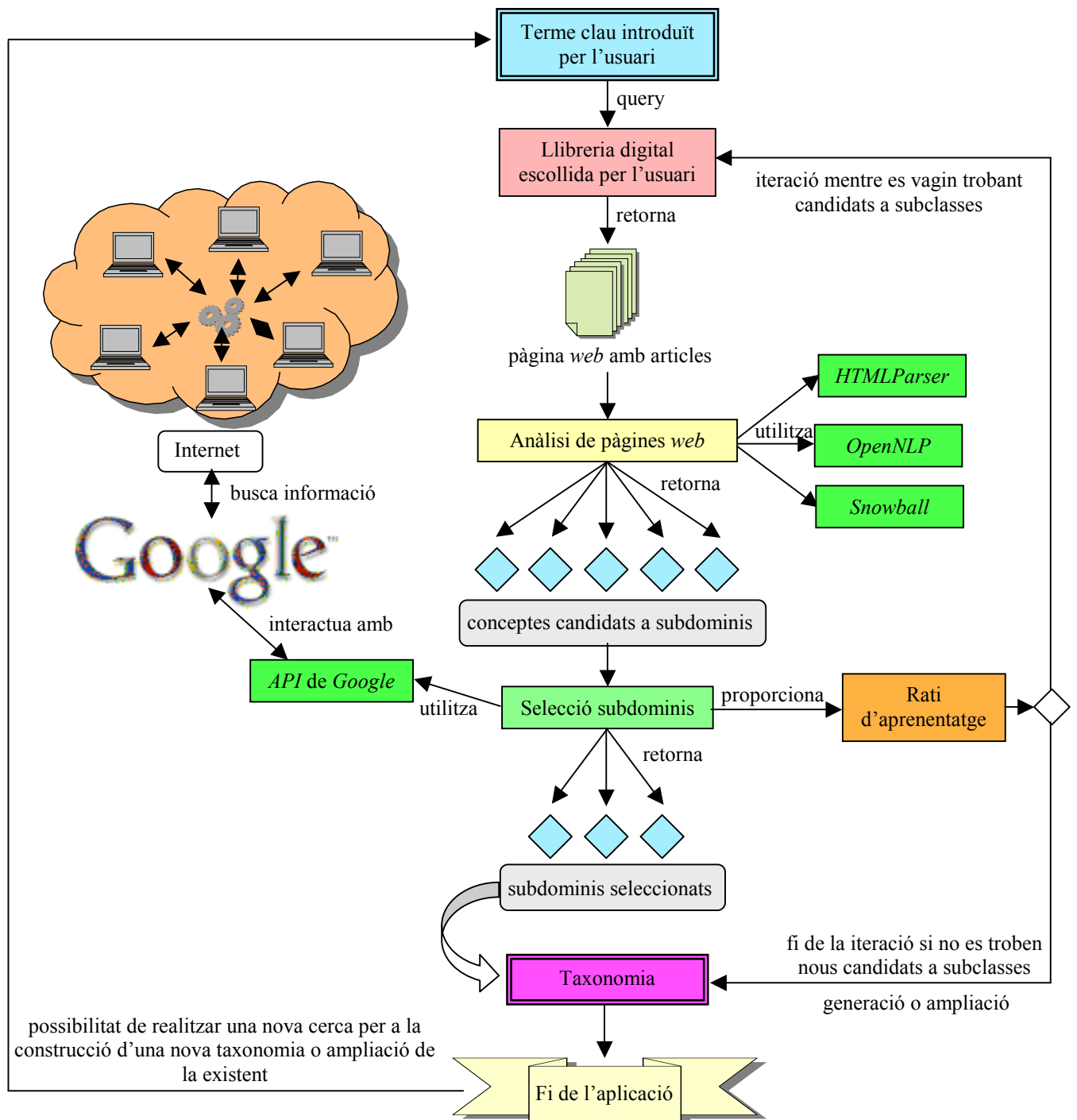


Figura 2. Passos generals que es duen a terme en el motor de cerca de l'aplicació.

Com es pot veure en la *Figura 2*, el motor de cerca de l'aplicació segueix una sèrie de passos des de que l'usuari introdueix el terme clau fins que es genera l'estructuració de la informació en forma de taxonomia. A continuació es descriuran de forma general cadascun d'aquests passos:

- 1- L'usuari introdueix al sistema el domini a partir del qual s'estructurarà la informació.
El terme clau, del qual encara no havíem parlat, ha de ser un terme escrit amb llengua anglesa. Això és així per dues raons. Una perquè el sistema ha estat dissenyat per a funcionar correctament amb dominis escrits amb llengua anglesa. La segona raó és que els articles obtinguts de les llibreries digitals utilitzades al sistema estan íntegrament escrits amb anglès.
- 2- A partir del terme clau, es realitza una *query* a la base de dades o llibreria digital que l'usuari ha escollit anteriorment.
- 3- La llibreria digital retornarà una pàgina *web* amb una sèrie d'articles relacionats amb el terme o termes introduïts per l'usuari.
- 4- A continuació, i a partir del procediment d'anàlisi de pàgines *web*, inclòs en el motor de cerca, s'analitzarà aquesta plana *web* amb l'ajuda de l'*HTMLParser* i de les tecnologies de processament de llenguatge natural (PLN) explicades anteriorment en el capítol 2 (*OpenNLP* i *Snowball*). El resultat d'aquest anàlisi *web*, serà una sèrie de conceptes candidats a subdominis del terme clau.
- 5- Seguidament, mitjançant el procediment de selecció de subdominis, també inclòs en el motor de cerca, es realitzarà la selecció de candidats finals a partir dels possibles candidats retornats per l'anàlisi de la pàgina *web*. Aquest procediment utilitzarà l'*API* de *Google* per a cercar informació a la *web* i ens retornarà una sèrie d'estadístics, la funció dels quals, explicarem més avant.
- 6- El procediment anterior de selecció de subdominis retornarà un rati d'aprenentatge, a partir del qual el sistema decidirà si s'ha de continuar analitzant planes *web* d'articles de la llibreria digital i així trobar més possibles candidats a subclasses, o simplement el procés ha de finalitzar i s'ha de generar finalment l'estructuració de la informació en forma de taxonomia a partir dels subdominis seleccionats com a bons.
- 7- Ja per acabar, i un cop construïda la taxonomia, l'aplicació dóna la oportunitat a l'usuari de fer una nova cerca i d'escollir una llibreria digital diferent si es vol. Si en canvi, el que es desitja és ampliar la taxonomia actual, l'usuari podrà realitzar noves cerques a partir de conceptes que aquesta conté a la seva estructura.

Un cop explicat molt per sobre els passos del funcionament del motor de cerca del sistema, procedirem a explicar detalladament els dos procediments principals que aquest conté: l'anàlisi de pàgines *web* i la selecció de subdominis. Seguidament hi ha un apartat que parla de la iteració del sistema, on s'expliquen detalladament els passos que segueix l'aplicació per a decidir si ha de continuar o no parsejant més planes *web* d'articles, per a extreure'n o no més quantitat de subdominis.

3.1.1- Anàlisi de pàgines web

En aquest apartat s'explicarà detalladament un dels procediments més importants del motor de cerca, l'anàlisi de pàgines web. La funció principal d'aquest procediment és analitzar les pàgines web amb articles i seleccionar una sèrie de candidats possibles a subdominis del concepte introduït per l'usuari.

Com hem vist en l'esquema anterior, la llibreria digital escollida per l'usuari retorna una pàgina web amb una sèrie d'articles relacionats amb el terme o termes introduïts per aquest.

Doncs bé, el primer que es fa en aquest procediment és parsejar aquesta pàgina web d'articles a partir de la seva URL passada per paràmetre. Per a aquesta tasca s'utilitza l'eina *HtmlParser*, descrita en l'apartat 2.1 i es fa servir el mètode d'extracció. El resultat final equival a una cadena de text que conté totes les paraules que hi havia a la pàgina web, de forma que es pot emmagatzemar a una variable de tipus *String*. Recordem que l'eina *HTMLParser* és capaç d'extreure tot el text d'una pàgina web prescindint de la resta d'informació innecessària, com poden ser fotografies, *banners*,...

En aquesta aplicació, potser a l'usuari solament li interessi estructurar la informació d'una llibreria digital, a partir d'un sol terme, com per exemple "*cancer*". Però el sistema ha estat dissenyat per a funcionar correctament també en més d'un terme, com per exemple "*breast cancer*". Això s'aconsegueix de la següent forma.

Si per exemple, l'usuari ha introduït "*breast cancer*", el que es fa és substituir l'espai (' ') per un *underscore* ('_'). De forma que el terme introduït per l'usuari canvia a "*breast_cancer*". Seguidament es substitueix en tot l'*string* retornat pel parsejador els termes "*breast cancer*" pel terme "*breast_cancer*". D'aquesta forma quan busquem entre tot aquest text, a més de cercar els termes "*breast cancer*" (cosa més difícil), podrem simplement realitzar la cerca buscant el terme "*breast_cancer*".

Un altre punt molt important dins el procediment d'anàlisi de les pàgines web, és l'eina '*StringTokenizer*'. Aquesta classe Java ens permetrà dividir tot l'*string* retornat abans pel parsejador en *tokens* o parts més petites de text. A continuació es veu un exemple de com funciona (els *strings* continguts entre claudàtors són els *tokens*):

- *String* extret d'un article de la B.D. *PubMed* [16] abans d'aplicar-li '*StringTokenizer*':

```
"in the normal organs (such as heart) of CT90-BikDD-treated animals."
```

→ *String* després d'aplicar-li '*StringTokenizer*' per defecte:

```
StringTokenizer st = new StringTokenizer(contingut_pagina_web);
```

```
[in][ ][the][ ][normal][ ][organs][ ][(such)][ ][as][ ][heart)][ ][of][  
[ ][CT90-BikDD-treated][ ][animals.]
```

→ *String* després d'aplicar-li '*StringTokenizer*' amb els paràmetres següents:

```
StringTokenizer st = new StringTokenizer  
(contingut_pagina_web, " .,:+~*/=%$~#@&<>{}[]()|\\\"\\n\\t!;?_`", true);
```

```
[in][ ][the][ ][normal][ ][organs][ ][ ]([)][ ][such][ ][as][ ][ ][heart][ )][ ][
][of][ ][CT90][ -][BikDD][ -][treated][ ][animals][ .]
```

Com es pot observar en el primer cas, la divisió en *tokens* de l'*string* es fa cada cop que es troba un espai (' '), ja que és el procediment per defecte de '*StringTokenizer*'. En canvi, en el segon cas la divisió en *tokens* de l'*string*, es fa cada cop que es troba un dels següents caràcters: [] [.] [,] [;] [:] [+] [-] [*] [/] [=] [%] [\$] [~] [#] [@] [&] [<] [>] [{] [}] [[]] [()] [||] [\\] [\"] [\\n] [\\t] [!] [;] [?] [_]. El paràmetre '*true*' significa que aquests caràcters es transformaran també en *tokens*.

Un cop es té tota la pàgina *web* transformada en *tokens*, ja es pot començar en la part més important del procés d'anàlisi de la *web*, el qual, com hem dit abans, ens retornarà una sèrie de candidats possibles a subdominis.

Recordem que el que ens interessa són especialitzacions del terme clau que ha estat introduït per l'usuari. Tant aquests especialitzacions com el terme clau, estaran escrites amb llengua anglesa.

Per tant, el fet que estigui tot amb llengua anglesa, implica que el terme que especifica vagi sempre davant del terme especificat. Per exemple, imaginem el parell de termes “*lung cancer*”. Doncs bé, en aquest exemple el terme especificat serà “*cancer*”, mentre que el terme que especifica serà “*lung*”.

Això significa que el procés d'analitzar la pàgina *web*, es centrarà solament amb els termes situats immediatament abans del terme clau introduït per l'usuari.

El procediment és sempre el següent:

- 1- Es llegeix un *token*, el qual s'emmagatzema en una variable anomenada '*prevWord*'. Els *tokens* [], [\n], [\t], que corresponen a un espai, a un salt de línia i a un tabulador, respectivament, es descarten.
- 2- A continuació es llegeix el següent *token* que s'emmagatzema en una variable anomenada '*word*'. Llavors es comprova si aquest *token* coincideix amb el terme que ha introduït l'usuari, és a dir, el terme clau. En el cas que no coincideixin, el *token* de la variable '*word*' passarà a la variable '*prevWord*' i en '*word*' s'emmagatzemarà el següent *token*, i així continuarà el procediment fins que no es trobi una coincidència entre un *token* llegit i el terme introduït per l'usuari.
- 3- Quan es troba el cas en que el *token* emmagatzemat en '*word*' coincideix amb el terme clau introduït per l'usuari, es duen a terme una sèrie de passos per a comprovar si el terme o *token* emmagatzemat a '*prevWord*' pot ser una possible subclasse del terme clau.
Per exemple, imaginem que el terme introduït per l'usuari (terme clau) ha estat "*cancer*" i resulta que en '*word*' hi ha el *token* "*cancer*". Immediatament es procedirà a analitzar el *token* existent en '*prevWord*', que podria ser un simple caràcter (per exemple, una ','), o una subclasse bona (per exemple, el terme "*breast*").

A continuació s'expliquen detalladament els passos (filtres) seguits per a analitzar el possible subdomini contingut en *'prevWord'* i decidir si s'accepta o no. A partir d'ara, a aquest possible subdomini l'anomenarem *'prevWord'*.

- 1- Es comprova que la *'prevWord'* no sigui massa curta, és a dir, si té menys de dos caràcters es desestimarà. Amb aquest filtre ens desfem dels signes de puntuació ja que tots tenen longitud d'un caràcter.
- 2- Es comprova que la *'prevWord'* no sigui una *stopWord*. Les *stopWords*, de les quals encara no havíem parlat, són termes, en aquest cas escrits amb llengua anglesa, tals com proposicions, articles definits, articles indefinits,... Per exemple: a, about, above, across, after, afterwards, again, against,... Aquestes paraules, donat que no tenen contingut semàntic, no poden ser de cap manera un subdomini, per tant, es desestimen.
- 3- Com s'ha dit abans, els subdominis són especialitzacions del terme clau introduït per l'usuari. Això implica que aquests termes hagin de ser per força noms o adjectius, ja que aquestes categories gramaticals són les que realment aporten un significat al terme clau.
Llavors, en aquest filtre es comprova que la *'prevWord'* es tracti solament d'un nom o un adjectiu. D'aquesta manera s'evitaran els verbs com a possibles subclasses. L'eina utilitzada per a realitzar aquesta tasca s'anomena *OpenNLP* i està explicada amb anterioritat en l'apartat 2.2.2. A continuació es mostren uns exemples:

- Ex.: "*breast cancer*" ("*breast*" → subclasse bona, es tracta d'un nom).
- Ex.: "*invasive cancer*" ("*invasive*" → subclasse bona, es tracta d'un adjectiu).
- Ex.: "*finding cancer*" ("*finding*" → subclasse dolenta, es tracta d'un verb).

- 4- En aquest filtre es comprova que l'arrel morfològica de la *'prevWord'* no coincideixi amb cap de les arrels de les subclasses que ja s'han donat per bones. D'aquesta forma s'evita tenir més d'una subclasse amb el mateix significat. L'eina utilitzada per a realitzar aquest tasca és l'*Snowball* i està explicada amb anterioritat en l'apartat 2.2.1. A continuació en veiem un exemple:

- Ex.: "*lung cancer*" ("*lung*" → subclasse bona, es tracta d'un nom).
- Ex.: "*lungs cancer*" ("*lungs*" → subclasse bona, es tracta d'un nom).

El sistema donaria per bones les dues subclasses. Això seria incorrecte, ja que com podem veure, la subclasse "*lungs*" és simplement el plural de la subclasse "*lung*".

Qualsevol terme contingut a *'prevWord'* que no compleixi cadascun dels quatre filtres anteriors, es desestimarà i no serà considerat com a un possible candidat a subclasse. En la *Figura 3*, es mostren els passos generals que realitza el procediment d'anàlisi de pàgines web.

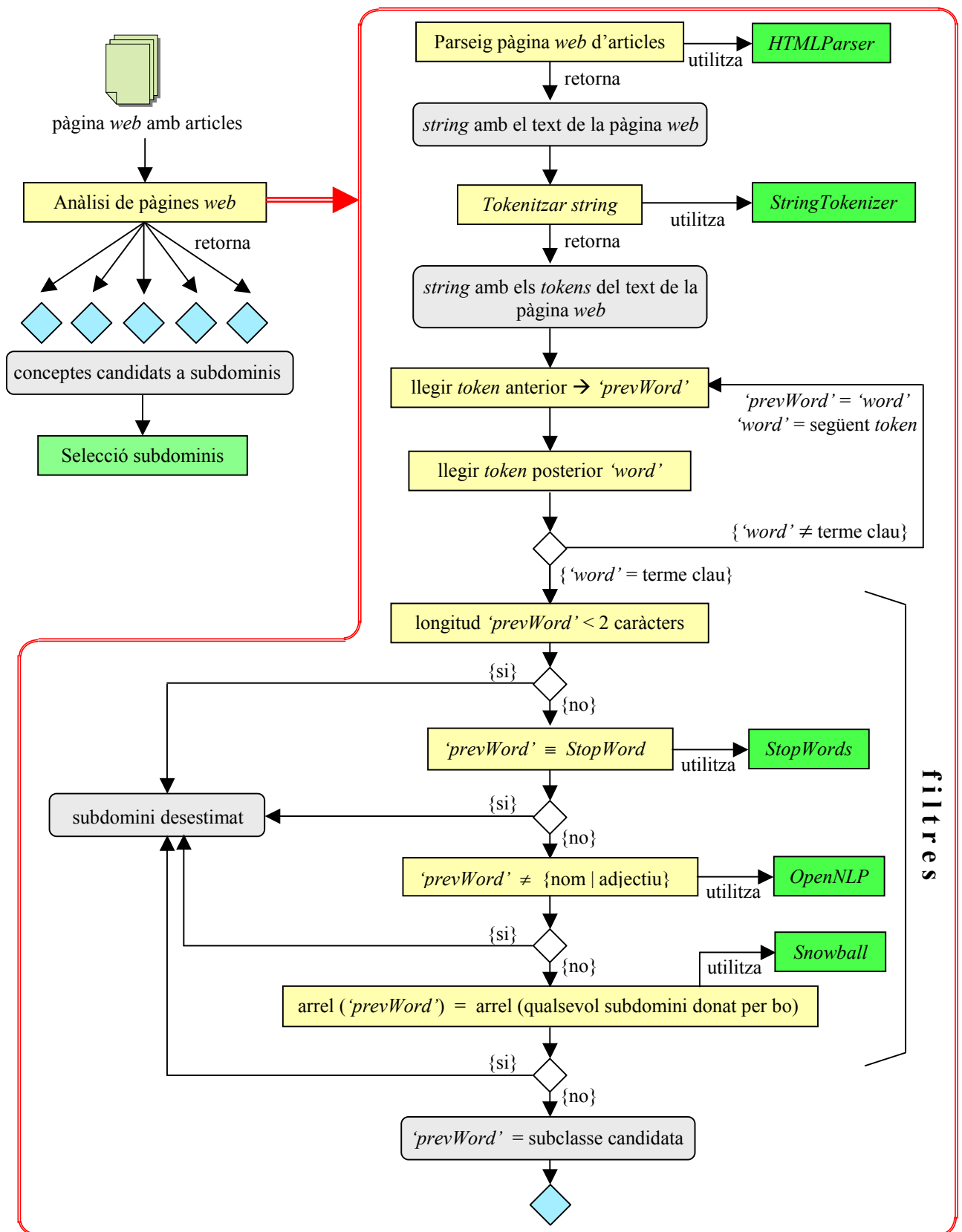


Figura 3. Passos seguits pel procediment d'anàlisi de pàgines web. En aquesta figura es poden observar els filtres que ha de superar un terme per a poder ser considerat un possible candidat a subdomini.

Doncs aquests són els quatre passos o filtres inicials que es segueixen per a decidir si el *token* existent en la variable *'prevWord'* es pot considerar o no com a una subclasse del terme clau. Tot i així, encara no es garantitza l'acceptació final d'aquesta possible subclasse.

Per tant, el següent i últim pas per a decidir la validesa d'un possible subdomini del terme clau serà utilitzar l'*API* de *Google* per a la construcció d'estadístics. En l'apartat següent s'explicarà detalladament el procediment de la selecció final de subclasses, el qual, juntament amb el procediment d'anàlisi de pàgines *web*, és un dels més importants del motor de cerca.

3.1.2- Selecció de subdominis

En aquest apartat s'explicarà detalladament un altre dels procediments importants del motor de cerca, la selecció de subdominis. La funció principal d'aquest procediment serà seleccionar els subdominis més rellevants dels possibles candidats que s'han escollit en el procediment anterior d'anàlisi de pàgines *web*.

Els candidats escollits en aquest procediment, seran els subdominis del terme clau que ha introduït l'usuari, i que finalment s'utilitzaran per a generar l'estructuració de la informació en forma de taxonomia.

Per a seleccionar els candidats més rellevants, s'utilitzaran uns estadístics que s'obtidran a partir de cerques al *Google* mitjançant el seu *API*. Més exactament aquests estadístics corresponen al número de resultats que retorna l'*API* com a conseqüència de realitzar la cerca d'un terme determinat.

Perquè utilitzar estadístics procedents del *Google*? Hi ha dues raons principals:

- Es poden obtenir molt ràpidament enviant una *query* a través de l'*API*.
- Aquests estadístics fan referència a la informació existent en tot Internet, el qual consta aproximadament d'un 8000 milions de pàgines *web*. Per tant, és millor utilitzar estadístics procedents del *Google*, abans de fer-ho, per exemple, de la mateixa llibreria digital de la qual s'ha extret el candidat a subclasse.

Llavors, perquè ens interessa saber el número de resultats (estadístics) que retorna el *Google* a partir de realitzar en ell la cerca d'un terme o termes determinats?

Posem per cas que l'usuari ha introduït com a paraula clau el terme "*cancer*", i imaginem que el procediment anterior d'analitzar pàgines *web* ha trobat el terme "*breast*" com a una de les possibles subclasses candidates. Doncs bé, aquest últim terme es passarà per paràmetre al procediment de selecció de candidats, el qual procedirà de la següent manera:

- 1- Es busca el terme "*breast*" (especialització) en el *Google* mitjançant l'*API* de *Google*. A nosaltres sols ens interessa el número de pàgines que es retornen (x) que fan referència a aquest terme.

- 2- Es busca el terme “*breast cancer*” (especialització + terme clau) en el *Google* mitjançant l’API de *Google*. A nosaltres sols ens interessa el número de pàgines que es retornen (y) que fan referència a aquests termes.
- 3- Un cop tenim els resultats retornats per l’API de *Google*, es fa la relació entre les dues cerques i es divideix y / x . Aquest nombre serà sempre menor que 1. La idea és buscar primer el terme “*breast*” i després els termes “*breast cancer*”. D’aquesta forma, segons els resultats que ens torni la segona cerca envers la primera, ens estarà informant inevitablement de si la relació que existeix entre aquests dos termes és forta (resultat s’aproxima més a 1) o més aviat fluixa (resultat s’aproxima menys a 1).

Un cop arribat en aquest punt, ja sabem perquè utilitzem estadístics, i perquè aquests s’obtenen del *Google* i no d’un altre lloc. La següent pregunta que ens podem fer a continuació és: com influirà el valor de la relació obtinguda en el punt 3 en el desenvolupament d’aquest procediment?

Bé, per a explicar-ho s’haurà de fer referència a un paràmetre necessari per a l’estructuració de la informació en forma de taxonomia, i que l’usuari ha d’haver introduït anteriorment. Ens referim a la precisió de la cerca. Aquest paràmetre pot tenir tres valors diferents: 0.001, 0.0001 i 0.00001. Aquests valors signifiquen una precisió elevada, una precisió mitja i una precisió baixa, respectivament, en el procés de selecció dels candidats finals per a la construcció de la jerarquia.

Per exemple, una precisió elevada (‘0.001’), ens assegura una selecció de candidats bastant bons. En canvi, amb una precisió baixa (‘0.00001’) obtindrem més candidats, però amb una qualitat mitja més baixa.

Explicat això, ja podem continuar amb el punt 4 i últim del procediment de selecció de subdominis.

- 4- Si la relació de les dues cerques obtinguda en el punt anterior és major que la precisió introduïda per l’usuari, la possible subclasse es donarà per bona, en canvi, si és menor, es desestimarà.

Perquè s’utilitzen aquests valors “empírics” i no uns altres? Bé, s’han realitzat una sèrie de proves, i s’han observat els candidats seleccionats dels que no ho han estat. A partir de la qualitat que s’obtenia dels candidats seleccionats depenent del valor de la precisió de cerca escollit, s’ha decidit finalment optar per aquests tres valors.

→ A continuació veiem un exemple per a una subclasse candidata que es seleccionarà com a bona:

- Cerca del terme “*breast*” a l’API del *Google* retorna un total de 183.000.000 de resultats.
- Cerca del terme “*breast cancer*” a l’API del *Google* retorna un total de 96.200.000 resultats.
- Posem per cas que el rati (precisió cerca) escollit per l’usuari és de 0.001.

- Com que la relació entre els resultats de les dues cerques ($96.200.000 / 183.000.000$) és de 0.5256, major que el rati (0.001), la subclasse “*breast*” de la classe “*cancer*” es donarà per bona i serà inclosa en l’estructuració final de la informació.

→ A continuació veiem un exemple per a una subclasse candidata que no es seleccionarà com a bona:

- Cerca del terme “*individual*” a l’API del *Google* retorna un total de 1.710.000.000 de resultats (x).
- Cerca del terme “*individual cancer*” a l’API del *Google* retorna un total de 137.000 resultats (y).
- Posem per cas que el rati (precisió cerca) escollit per l’usuari és de 0.001.
- Com que la relació entre els resultats de les dues cerques ($137.000 / 1.710.000.000$) és de 0.00008, menor que el rati (0.001), la subclasse “*individual*” de la classe “*cancer*” no es donarà per bona i serà descartada.

A continuació, en la *Figura 4*, es pot veure un esquema que mostra gràficament el funcionament del procediment de selecció de subdominis. Es posa per exemple el procés de selecció del subdomini “breast”, tenint com a paraula clau el terme “cancer”.

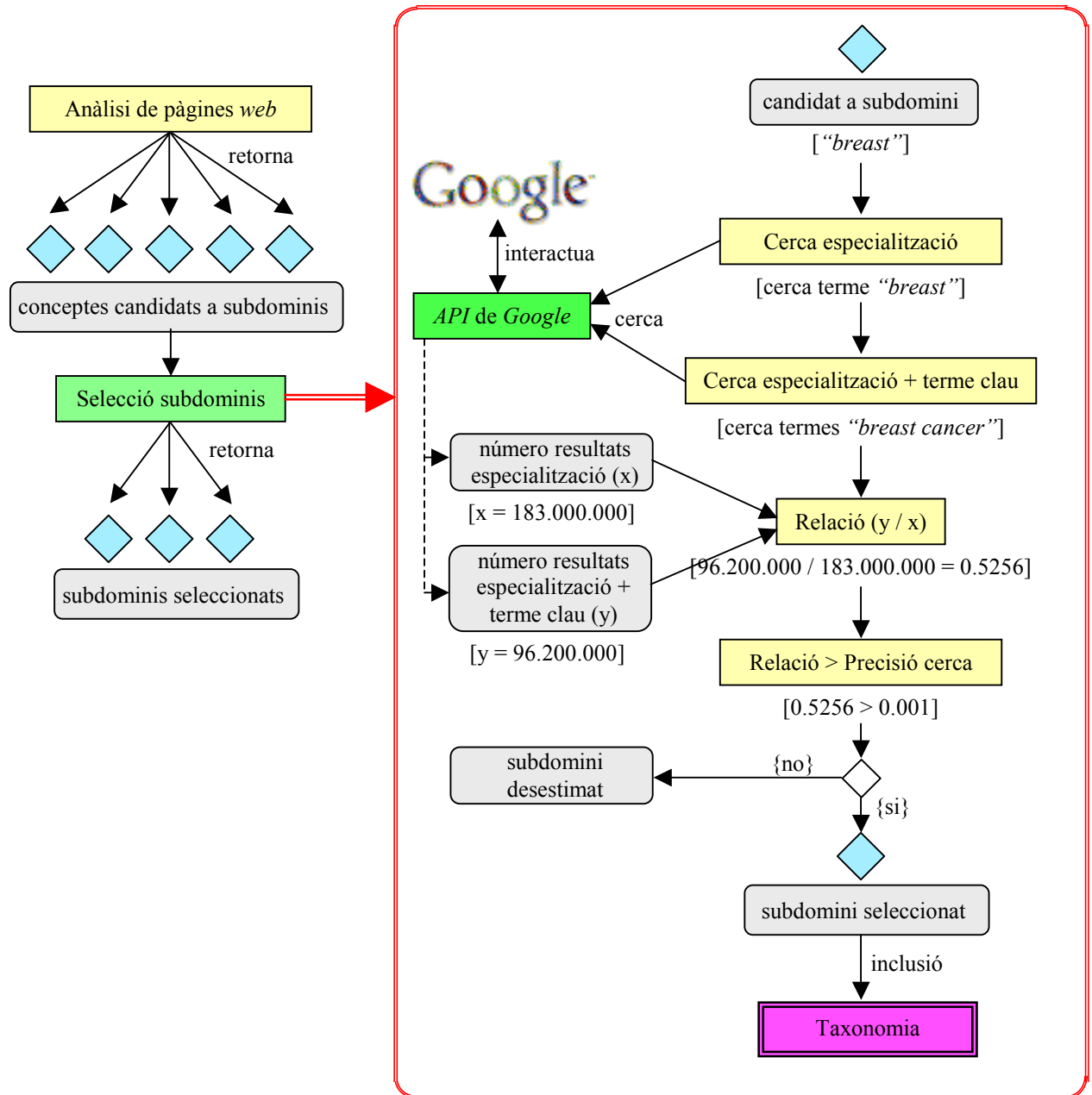


Figura 4. Passos generals que segueix el procediment de selecció de subdominis.

3.1.3- Iteració del sistema

Arribats en aquest punt de la documentació, sabem que el motor de cerca està format per dos procediments molt importants.

Un s'encarrega de parsejar les pàgines *web* que contenen articles relacionats amb el terme introduït per l'usuari, i a extreure'n d'aquestes, les possibles subclasses que potser acabin formant part de l'estructuració d'informació de la llibreria digital.

L'altre procediment s'encarrega de decidir, amb l'ajuda d'estadístics retornats pel *Google*, quines de les possibles subclasses seleccionades anteriorment passaran a formar part de les branques de la taxonomia que acabarà mostrant-se a l'usuari.

En la introducció de l'apartat 3.1, hem explicat les tasques principals que desenvolupa el motor de cerca. En un dels seus punts hem indicat que el procediment de selecció de subdominis retorna el rati d'aprenentatge. A partir d'aquest rati el sistema decideix si s'han de continuar analitzant pàgines *web* d'articles de la llibreria digital i així trobar més possibles candidats a subclasses, o simplement, el procés ha de finalitzar i s'ha de generar finalment la jerarquia a partir dels subdominis seleccionats com a bons.

El *learning rate*, o rati d'aprenentatge, és una relació del número de subclasses considerades com a bones, segons el procediment de selecció de subdominis, entre el número total de subclasses que s'han trobat mitjançant el procediment d'anàlisi de pàgines *web*.

Arribats en aquest punt hem de fer referència a un altre dels paràmetres que l'usuari ha d'haver introduït prèviament per a l'execució de l'aplicació. Ens referim a l'amplada de la cerca. Aquest paràmetre pot tenir quatre valors diferents: 80%, 60%, 50% i 40%. Aquests valors signifiquen una amplada de la cerca senzilla, mitjana, complexa i molt complexa, respectivament. Què significa i per a que s'utilitza aquest paràmetre?

Doncs bé, això significa ni més ni menys, que una amplada de cerca senzilla, ens retornarà un nombre més aviat petit de subdominis amb un temps relativament petit. En canvi, en una amplada de cerca molt complexa, el sistema ens pot arribar a retornar un gran nombre de subdominis, això sí, utilitzant una gran quantitat de temps.

Aquest paràmetre és molt important degut a que dona l'oportunitat a l'usuari de decidir l'exhaustivitat del procés de cerca.

Si per exemple l'usuari disposa de poc temps, pot escollir l'opció d'una amplada de cerca senzilla. Amb aquesta opció, el sistema li retornarà una taxonomia amb dos minuts aprox., que contindrà algunes de les especialitzacions que té el terme que ha introduït.

En canvi, si per exemple a l'usuari no li importa el temps que trigui el sistema, pot escollir la opció d'una amplada de la cerca molt complexa. D'aquesta forma, la taxonomia retornada contindrà moltíssimes de les especialitzacions del terme que ha introduït l'usuari.

A continuació, s'explicarà detalladament i amb l'ajuda d'un exemple, el funcionament d'aquest procés.

- Imaginem que el procediment d'anàlisi de pàgines *web* analitza els 20 primers articles retornats per una de les llibreries digitals del sistema.
- Dels termes analitzats com a possibles subclasses, 10 són bons, mentre que 6 no han estat acceptats. Llavors tenim que el total de subclasses és de 16, mentre que el total de subclasses seleccionades és de 10.
- Si fem la relació ens dona un rati d'aprenentatge de 62.5% $((10 / 16) * 100)$.
- Llavors, que l'aplicació continuï buscant més subclasses o no, dependrà del valor de l'amplada de la cerca introduït per l'usuari.
Per exemple, si l'usuari volia una cerca més aviat superficial i ha escollit l'opció d'una amplada de cerca senzilla (80%), l'aplicació no continuarà buscant perquè $62.5\% < 80\%$.
En canvi, si per exemple l'usuari ha escollit l'opció d'una amplada de cerca mitjana (60%), l'aplicació continuarà buscant més subclasses possibles en els pròxims 20 articles que la llibreria digital retornarà, ja que $62.5\% > 60\%$.
- Imaginem ara que en els pròxims 20 articles analitzats, el sistema troba 4 subdominis bons i 12 de dolents. Si tenim en compte d'abans que el total de subclasses era de 16 i el total de subclasses seleccionades era de 10, ara tindrem que el total de subclasses serà de 32 i el total de subclasses seleccionades serà de 14.
- Si fem la relació ens dona un rati d'aprenentatge de 43.75% $((14 / 32) * 100)$.
- En el cas que l'usuari hagués escollit l'opció d'una amplada de cerca 'Mitjana' (60%), l'execució de l'aplicació finalitzaria ja que $43.75\% < 60\%$.

Per tant, és el mateix sistema, el qual a partir del valor del paràmetre d'amplada de la cerca introduït per l'usuari, decideix quan ha d'aturar-se de parsejar articles. Llavors, per dir-ho d'alguna manera, podríem dir que el motor de cerca utilitza Intel·ligència Artificial, ja que ell sol és capaç de prendre certes decisions.

3.2- Estructures de dades

Les estructures de dades (E.D.) són sempre una part molt important en qualsevol aplicació o programa informàtic. Un bon disseny de les E.D. pot suposar un estalvi de memòria important i un accés ràpid a aquestes. Avui en dia, l'alta velocitat i la gran capacitat de memòria RAM de les computadores fa que aquests aspectes no siguin del tot importants. Això no vol dir però, tenir que donar menys importància al seu disseny.

A continuació descriurem les E.D. utilitzades en l'aplicació. Totes elles són classes escrites en Java i cadascuna de les variables definides dintre d'elles tenen la seva funció *set* (escriure valor) i *get* (llegir valor) respectiva.

3.2.1- Estructura de dades elemental

Aquest apartat fa referència a la E.D. anomenada '*ClassAndWebs*'.

Es tracta de l'E.D. més elemental del sistema i en ella s'emmagatzema la subclasse i el rati retornat per l'*API* de *Google* de la relació entre aquesta subclasse i la seva classe.

3.2.2- Resultat de la cerca

Aquest apartat fa referència a la E.D. anomenada '*ResultatBusqueda*'.

En aquesta E.D. s'emmagatzemarà informació, com el seu nom indica, referent al resultat d'una cerca realitzada per l'usuari. Exactament es guardarà:

- El nom del terme introduït per l'usuari.
- La *URL* de la B.D. que retorna els articles resultants de fer la cerca amb el terme introduït per l'usuari.
- Les subclasses bones que han estat seleccionades pel procediment de selecció de subdominis.
- El rati retornat per l'*API* de *Google*. Per exemple, si la cerca al sistema ha estat dels termes "*breast cancer*", en aquesta variable s'emmagatzemarà la relació entre el nombre de resultats de "*breast*" i el nombre de resultats de "*breast cancer*".

3.2.3- Element de l'estructura

Aquest apartat fa referència a la E.D. anomenada '*ElementEstructura*'.

Aquesta E.D. emmagatzemarà la informació mínima necessària per a poder dibuixar, més tard, una petita part de la taxonomia. En ella es guardarà la següent informació:

- El terme de la classe o subclasse.
- El rati o relació entre aquesta subclasse i la seva classe.
- El nivell. Aquesta variable fa referència a la profunditat del terme en l'estructura de la taxonomia. Per exemple, en el cas que l'usuari introdueixi el terme "*cancer*", aquest tindrà un nivell amb valor 0, ja que es refereix a l'arrel de la jerarquia. Totes les subclasses que es trobin a partir d'aquest terme tindran nivell 1 (ex.: "*breast cancer*"). Les subclasses que es trobin per a qualsevol de les subclasses anteriors tindran nivell 2 (ex.: "*metastatic breast cancer*"), i així successivament.

3.2.4- Estructura de resultats

Aquest apartat fa referència a la E.D. anomenada '*EstructuraResultats*'.

Aquesta E.D. és molt important i utilitza les dues E.D. anteriors ('*ResultatBusqueda*' i '*ElementEstructura*'). Com el seu nom indica, en ella s'emmagatzemaran tots els resultats de les cerques que es vagin generant en el sistema. Està composta per tres variables i un procediment recursiu. Les variables són:

- El terme arrel de la jerarquia. Aquest és el que introdueix l'usuari quan comença una nova cerca. Per exemple, "*cancer*".
- Una *Hashtable* anomenada '*hashCodeResultats*' que contindrà els resultats de les cerques que es vagin realitzant. Més concretament contindrà objectes del tipus '*ResultatBusqueda*'.
- Un *ArrayList* anomenada '*estructura*' que s'encarregarà de guardar objectes del tipus '*ElementEstructura*'.

Per últim, aquesta classe Java està formada per un procediment recursiu anomenat '*construirEstructura*' que s'encarrega d'inicialitzar la variable '*estructura*', que més tard s'utilitzarà per a dibuixar la taxonomia.

3.2.5- Bases de dades

Aquest apartat fa referència a la E.D. continguda en la classe '*BaseDades*'.

Aquesta E.D., com el seu nom indica, es dedica a guardar informació concreta de cadascuna de les quatre B.D. o llibreries digitals existents al sistema. Concretament s'emmagatzema:

- El nom de la llibreria digital.
- La *URL* de la pàgina *web* principal o pàgina d'inici de la llibreria digital.
- El nom d'un fitxer d'imatge on hi ha el logotip de la llibreria digital.

3.3- Llibreries digitals

Com hem anat dient durant aquesta documentació, el sistema es dedica a l'extracció d'especialitzacions d'un terme clau introduït per l'usuari a partir d'articles que retornen les llibreries digitals que el sistema disposa.

En aquest apartat es donarà informació de cadascuna d'aquestes llibreries digitals, així com de les característiques que han de tenir aquestes.

Les quatre bases de dades o llibreries digitals d'articles electrònics són:

- *Association for Computing Machinery. The Guide* [17]
- *PubMed*
- *IEEE. Computer Society* [18]
- *NASA Astrophysics Data System* [19]

Perquè s'han decidit escollir aquestes llibreries i no unes altres? Bé, aquests repositoris d'articles compleixen una sèrie de requisits que a continuació s'especifiquen:

- Ha de ser un repositori d'articles o publicacions científiques.
- Ha de tenir una gran quantitat d'articles, ja sigui de diversos dominis científics o d'un de sol.
- Les llibreries han de disposar d'un buscador de forma que es puguin realitzar *query's* al seu motor de cerca intern.
- S'han de poder parsejar fàcilment les pàgines d'articles retornades per la llibreria, de forma que es puguin extreure els candidats a subclasses.
- Que doni informació dels articles de forma "resumida", com els *abstracts*, vistes prèvies o, al menys el títol complet.

Les quatre llibreries digitals seleccionades compleixen cadascun dels requisits anteriors, per tant s'ha decidit integrar-les en el sistema.

Un cop hem arribat en aquest punt, ens podem preguntar: com realitzem la cerca del terme introduït en la llibreria digital seleccionada per l'usuari? Bé, gairebé totes les B.D. d'articles segueixen un procediment estàndard, que és el d'enviar una *URL*, amb una sèrie de paràmetres, al seu propi motor de cerca.

Entre aquests paràmetres podem trobar: el terme a cercar (paràmetre indispensable), el número d'articles per pàgina a retornar pel motor de cerca de la B.D., l'índex a partir del qual retornar els articles, el format dels articles retornats (text pla, *html*, *XML*,...),...

A continuació mostrem un exemple de la *URL* utilitzada per a cercar articles a la llibreria digital *PubMed*:

→ <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=search&db=pubmed&term=cancer&doptcmdl=citation&inputpage=1>

Com podem observar, les *URL's* tenen dues parts. La primera fa referència a l'adreça del servidor *web* de la B.D. La segona part fa referència als paràmetres i va sempre després del símbol '?'. En aquesta *URL* trobem paràmetres tals com:

- ‘*db*’: es refereix a la B.D. on es desitja buscar, en aquest cas *PubMed*.
- ‘*term*’: es refereix al terme que es desitja cercar i que introdueix l’usuari.
- ‘*doptcmdl*’: fa referència al format del resultat. En aquest cas els resultats seran en forma de *citations*.
- ‘*inputpage*’: és també un paràmetre molt important i es refereix a la pàgina a partir de la qual es retornen els articles. Per exemple, tenint en compte que el *PubMed* sols retorna 20 articles per plana, si *inputpage* val 3, el primer article de la plana retornada correspondrà a l’article amb índex 61.

No explicarem les altres tres *URL*’s de les altres B.D., ja que són molt semblants, llevat de que poden haver-hi més o menys paràmetres i utilitzar noms diferents per a aquests.

A continuació es mostra una informació molt superficial sobre cadascun del repositoris que s’han integrat en el sistema.

3.3.1- Association for Computing Machinery. The Guide

“*The Association for Computing Machinery (ACM)*”, mostra recursos que avancen la informàtica com a una ciència i com a una professió. *ACM* proporciona el *Digital Library*, que serveix als seus membres i als professionals informàtics amb publicacions molt punteres, i recursos de professionals.

“*The Guide*” és una col·lecció de cites bibliogràfiques i *abstracts* de treballs publicats per *ACM* i altres editors. Per a tots els treballs publicats en *ACM* es disposa de *citations* incloses en “*The Guide*” com a un petit resum de tot l’article.

“*The Guide*” disposa de més de 750.000 *citations* de més de 3000 editors, incloent *ACM*, i a més:

- llibres.
- articles de revistes.
- tesis doctorals.
- màsters i informes tècnics.

Tots els professionals d’*ACM*, estudiants, i membres de “*Special Interest Group (SIG)*” tenen un accés total a “*The Guide*” com a membres beneficiaris. Es requereix un compte gratuït a *ACM* per a accedir-hi.

Els que no siguin membres d’*ACM* o el públic en general, tenen un accés més restringit a l’hora de realitzar les cerques o navegar pels articles.

Per a més informació sobre la B.D. “*Association for Computing Machinery. The Guide*”, es pot visitar la seva pàgina web.

3.3.2- PubMed

PubMed, disponible via “*NCBI Entrez Retrieval System*”, és una font gratuïta que va ser desenvolupada per “*National Center for Biotechnology Information (NCBI)*” a la “*National Library of Medicine (NLM)*”, situat a “*National Institutes of Health (NIH)*”. *Entrez* és un sistema basat en cerques de text, usat a *NCBI* per a serveis i inclou B.D. com: *PubMed*, *Nucleotide i Protein Sequences*, *Protein Structures*, *Complete Genomes*, *Taxonomy*, *OMIM*, i molts més. *PubMed* va ser dissenyada per a donar accés a articles de la literatura biomèdica. *Linkout* dona accés al text complet d'articles a llocs *web* de publicacions i altres fonts relacionades. *PubMed* també dona accés mitjançant *links*, a altres fonts de la biologia molecular d'*Entrez*.

Els editors participants en *PubMed* envien electrònicament els seus articles a *NCBI*. Si l'editor té un lloc *web* que ofereix el text complet dels seus articles, *PubMed* proporciona *links* a aquests llocs *web* i també als recursos biològics, informació de sanitat per al consumidor, eines d'investigació,...

PubMed permet buscar entre més de 16 milions de cites bibliogràfiques i *abstracts* en el camp de la medicina, infermeria, odontologia, medicina veterinària, sistema d'assistència mèdica, i ciències pre-clíniques.

Per a més informació sobre la B.D. “*PubMed*”, es pot visitar la seva pàgina *web*.

3.3.3- IEEE. Computer Society

Amb prop de 100.000 membres, la *IEEE Computer Society* és la organització més puntera en el món pel que fa als professionals de la informàtica. Fundada en 1946, és la més gran de les 39 societats de la IEEE (*Institute of Electrical and Electronics Engineers*).

La missió d'*IEEE Computer Society* és el proveïment d'informació tècnica, serveis a la comunitat i serveis personalitzats en el món dels professionals de la informàtica.

La societat es dedica en l'avenç de la teoria, la pràctica i l'aplicació de la informàtica i el processament tecnològic de la informació. A través de les seves conferències, aplicacions relacionades, revistes d'investigació orientada, seccions dedicades a estudiants o particulars, campus d'aprenentatge a distància, comitès tècnics, i grups de treball estàndards, la societat promou un actiu intercanvi d'informació, idees, i innovació tecnològica entre els seus membres. A més, la societat manté llaços amb el “*US Computing Sciences Accreditation Board*” i “*Accreditation Board for Engineering and Technology*”, controlant i avaluant les normes acreditatives curriculars.

Amb el 40% aprox. dels membres vivint i treballant fora dels EEUU, la *Computer Society* acull comunicació internacional, cooperació, i intercanvi d'informació. Per a satisfer les necessitats dels membres convenientment i eficientment, la societat manté una oficina de servei a Tokyo (Japó); una oficina de publicacions en Los Alamitos (Califòrnia); i la seva seu central en Washington, DC.

Per a més informació sobre la B.D. “*IEEE. Computer Society*”, es pot visitar la seva pàgina *web*.

3.3.4- NASA Astrophysics Data System

El “*Astrophysics Data System (ADS)*” és un projecte fundat per la *NASA* el qual manté tres B.D. bibliogràfiques que contenen més de 4.7 milions d’articles: “*Astronomy and Astrophysics*”, “*Physics*” i “*ArXiv e-prints*”. El cos principal de les dades de l’*ADS* consisteix en els articles bibliogràfics, els quals són accessibles a través del nostre “*Abstract Service query forms*”, i articles sencers de la literatura astronòmica, els quals poden ser visitats a través del nostre “*Browse Interface*”. Integrat en aquestes B.D., el *ADS* dona accés a la riquesa de les fonts externes, incloent articles electrònics, catàlegs de dades i arxius. Nosaltres actualment disposem de *links* sobre uns 4.8 milions d’articles mantinguts pels nostres col·laboradors.

Fixeu-vos que tots els *abstracts* i articles en l’*ADS* tenen drets d’autor, i el seu ús es lliure solament com a ús personal. Per a més informació, llegiu detalladament la nostra pàgina *Terms and Conditions*, la qual regula l’ús dels nostres recursos.

La importància del paper d’*ADS* en el suport de la comunitat científica ha estat reconeguda per societats i individuals.

Per a més informació sobre la B.D. “*NASA Astrophysics Data System*”, es pot visitar la seva pàgina *web*.

3.4- Interfície web

En aquest apartat s'explicarà detalladament la interfície gràfica d'usuari. Més exactament, la interfície *web*, ja que com hem dit en la introducció del capítol 2 (descripció de la tecnologia utilitzada en el projecte), la interacció de l'usuari amb l'aplicació es realitzarà mitjançant una aplicació *web* construïda amb *JSP's* i *servlets*.

S'ha decidit realitzar una pàgina *web* que embolcalli de forma transparent l'accés a les llibreries digitals abans mencionades. És a dir, la interfície *web* es tracta d'una meta – interfície, la funció de la qual, és proporcionar a l'usuari una forma senzilla i similar d'interactuar amb els quatre repositoris d'articles.

El fet que l'usuari hagi d'introduir les dades a partir d'una aplicació *web* construïda amb *JSP's* implica que aquestes dades, quan s'envien (*submit*) al motor de cerca de l'aplicació, aquest les haurà de recuperar mitjançant un *servlet*.

Això implica haver tingut de realitzar uns canvis en l'estructura del sistema. I és que el motor de cerca de l'aplicació s'ha tingut d'incrustar dintre d'un *servlet* per a poder llegir les dades que l'usuari introdueix via *JSP*.

Els *JSP's* i els *servlets* estan contínuament connectats, és a dir, els *servlets* s'encarreguen del control i el processament de les dades, mentre que els *JSP's* s'encarreguen més aviat de l'adquisició i presentació d'aquestes.

És a dir, l'usuari introdueix dades mitjançant el *JSP*, i aquestes són enviades al *servlet* per a ser processades. Doncs bé, l'espai utilitzat per a l'intercanvi d'aquestes dades s'anomena sessió. Per dir-ho d'alguna forma la sessió és com un contenidor on tant el *JSP* com el *servlet* poden emmagatzemar i llegir variables en forma de nom – valor.

Abans de continuar parlant de la interfície *web*, hauríem de fer referència als diferents fitxers de que consta l'aplicació i classificar-los segons la seva funció. L'aplicació està formada per tres tipus diferents de fitxers. Els *JSP's*, els *Servlets* i les Classes.

Les Classes són fitxers escrits en Java i en aquesta aplicació corresponen bàsicament a les estructures de dades que s'utilitzen.

→ *JSP's*:

- *Pàgina web d'inici ('index.jsp')*: Fa referència a la pàgina de benvinguda de l'aplicació on s'escull la llibreria digital.
- *Pàgina web* que interactua amb el motor de cerca de l'aplicació (*'searcher.jsp'*): És la pàgina a través de la qual l'usuari introduirà els paràmetres de cerca i obtindrà també els resultats.

→ Servlets:

- ‘*SearcherServlet.java*’: És el fitxer més important del sistema. En ell hi ha el motor de cerca de l’aplicació que s’encarrega de parsejar els articles de la B.D., obtenir les possibles subclasses, seleccionar les que siguin adequades i generar l’estructura d’informació en forma de taxonomia que més tard es mostrarà en la pàgina *web* del motor de cerca.
- ‘*BaseDadesServlet.java*’: Aquest fitxer es correspon també a un *servlet* i està estrictament lligat a la ‘Pàgina *web* d’inici’. S’encarrega de guardar a la sessió el nom de la B.D. escollit per l’usuari de forma que pugui ser utilitzat posteriorment pel fitxer ‘*SearcherServlet.java*’.

→ Classes:

- ‘*ClassAndWebs.java*’, ‘*ElementEstructura.java*’, ‘*EstructuraResultats.java*’, ‘*BaseDades.java*’ i ‘*ResultatBusqueda.java*’: són fitxers corresponents a les estructures de dades utilitzades pel sistema per a organitzar tota la informació.
- ‘*StringComparator.java*’: Classe Java que s’encarrega de comparar dos *strings* passats per paràmetre. Aquesta classe s’utilitza en el fitxer ‘*SearcherServlet.java*’ per a ordenar alfabèticament les llistes de subclasses.

En el fitxer de tipus *servlet* ‘*SearcherServlet*’ hi ha dos procediments principals que contenen tots els altres:

- `public void init (ServletConfig conf) throws ServletException;`
aquest procediment s’executa solament el primer cop que es crida el *servlet*. En ell es carreguen a memòria tres fitxers. Un d’ells és el que conté les *stopWords*, de les quals ja hem parlat anteriorment. Els altres dos fitxers (‘*EnglishTok.bin*’ i ‘*EnglishPOS.bin*’) ja els hem mencionat abans en l’apartat 2.2.2, i són utilitzats per l’eina *OpenNLP* per a classificar sintàcticament les paraules.
- `protected void doPost (HttpServletRequest req, HttpServletResponse resp) throws ServletException, IOException;`
aquest procediment s’executa sempre que es crida el *servlet* i conté el motor de cerca de l’aplicació. Per tant, també conté els dos procediments més importants del motor de cerca: l’anàlisi de pàgines *web* i la selecció de subdominis.

Com hem vist en la estructuració de fitxers, n’hi ha dos que són del tipus *JSP*. Doncs aquests dos fitxers corresponen a les dues parts en que s’ha decidit dividir la interfície *web*. Per una banda tenim la pàgina *web* d’inici, i per l’altra, la pàgina *web* que l’usuari utilitzarà per a interactuar amb el sistema.

En el següent i últim apartat del capítol 3 s’inclourà un manual d’usuari.

3.5- Manual d'usuari

Aquest apartat pretén solucionar qualsevol dubte que l'usuari pugui tenir sobre el funcionament de la interfície *web*, així com dels paràmetres utilitzats per a generar la estructuració de la informació en forma de taxonomia.

3.5.1- Pàgina web d'inici

La tasca principal de la pàgina *web* d'inici, a part de donar la benvinguda a l'usuari, permetrà l'opció d'escollir una d'entre les quatre llibreries digitals que l'aplicació disposa i que ja s'han explicat en l'apartat anterior. La *Figura 5* mostra l'aspecte que té aquesta part de la interfície *web*.

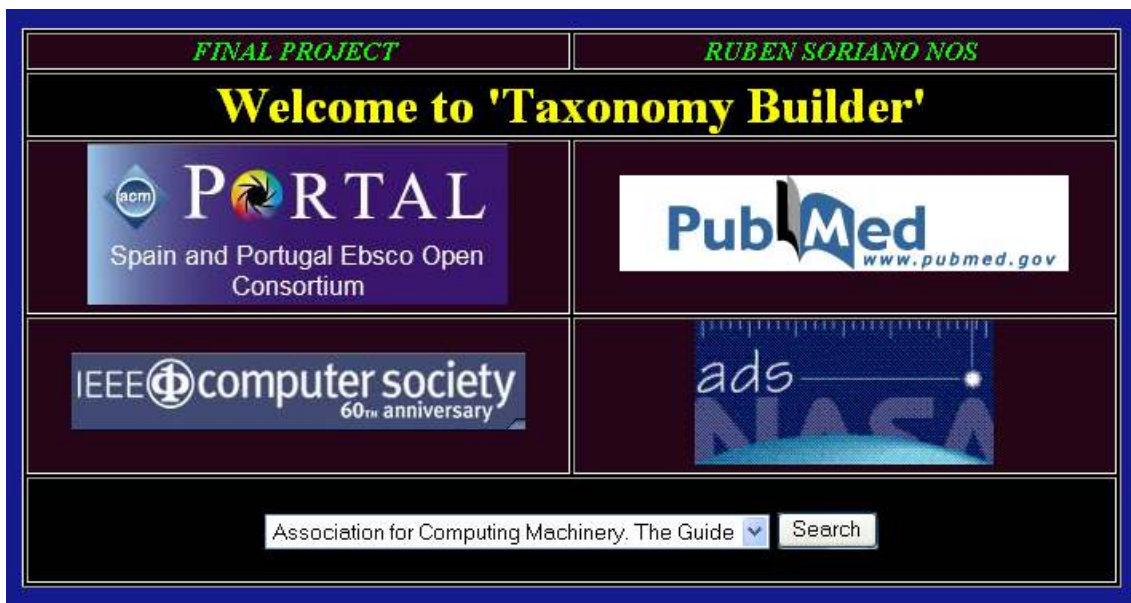


Figura 5. Vista principal de la pàgina *web* d'inici.

3.5.2- Pàgina web que interactua amb el motor de cerca de l'aplicació



Figura 6. Vista principal de la pàgina web que interactua amb el motor de cerca de l'aplicació

La pàgina web que interactua amb el motor de cerca de l'aplicació és l'eina que l'usuari utilitzarà per a enviar les dades al sistema i també per a veure els seus resultats. Es tracta d'una taula *HTML* on cada cel·la té una funció determinada.

En la *Figura 6* es pot veure l'aspecte que té aquesta part de la interfície web. Per explicar-la detalladament s'ha dividit en cinc parts diferents que a continuació s'expliquen detalladament:

- 1- Correspon al logotip de la B.D. escollida en la pàgina web d'inici.
- 2- Aquesta cel·la està formada per un formulari amb tres camps, els quals són els suficients i necessaris per a que l'usuari pugui realitzar la cerca.
 - *Search word:* en aquest quadre de text s'introduirà el terme clau a cercar, a partir del qual es construirà la taxonomia.

Recordem que el concepte ha d'estar escrit amb llengua anglesa per raons que ja s'han explicat anteriorment en l'apartat 3.1 del motor de cerca.

Aquest quadre de text sols estarà activat per a introduir el terme arrel a partir del qual es construirà la jerarquia. Més endavant, l'usuari sols podrà interactuar amb els termes d'aquesta jerarquia per a poder realitzar noves cerques.

- *Search width*: Hi ha quatre opcions a escollir en aquest *select*: ‘*Senzilla*’, ‘*Mitjana*’, ‘*Complexa*’ i ‘*Molt_Complexa*’. La opció escollida per defecte serà ‘*Senzilla*’.

La funció d’aquest paràmetre ha estat explicada molt detalladament en l’apartat 3.1.3 que parla sobre la iteració del sistema.

Com és lògic d’imaginar, però, depenent del valor que s’esculli es realitzarà una cerca més extensa (‘*Molt_Complexa*’) on es parsejaran molts articles de la B.D., o una cerca més superficial (‘*Senzilla*’) on es parsejaran menys articles de la B.D.

- *Search precision*: Hi ha tres opcions a escollir en aquest *select*: ‘*0.001*’, ‘*0.0001*’ i ‘*0.00001*’. La opció escollida per defecte serà ‘*0.001*’.

La funció d’aquest camp ja ha estat explicada força detalladament en l’apartat 3.1.2 que parla sobre la selecció de subdominis.

El valor ‘*0.001*’ correspondria a una cerca de precisió elevada, el valor ‘*0.0001*’ correspondria a una cerca de precisió mitja i el valor ‘*0.00001*’ correspondria a una cerca de precisió baixa. Si la precisió és elevada (‘*0.001*’), els candidats que es seleccionin com a subclasses seran molt bons, és a dir, d’una alta qualitat. En canvi, per a una precisió baixa (‘*0.00001*’), els candidats que es seleccionin com a subclasses tindran, com a mitja, una qualitat més baixa, encara que la quantitat retornada serà major.

3- En aquesta cel·la s’ubiquen tres botons que explicarem a continuació:

- ‘*Confirm Search*’: Aquest botó fa referència al formulari anterior i és el botó *submit* que envia les dades introduïdes per l’usuari en l’apartat 2 al motor de cerca del sistema.
- ‘*Save Search*’: Aquest botó solament estarà activat si ja s’ha realitzat una cerca anteriorment. Com el seu nom indica s’utilitza per a emmagatzemar la cerca realitzada en un fitxer amb format *htm*, és a dir, s’emmagatzema la estructuració de la informació en forma de taxonomia generada pel sistema.

Aquesta opció pot ser interessant per a salvar resultats, els quals, el sistema ha trigat molt temps en generar.

- ‘*New Search*’: Aquest botó es pot utilitzar en qualsevol moment i prémer-lo significa dues coses.

Primerament el que es fa es resetejar les estructures de dades dels resultats obtinguts fins al moment, és a dir, es partirà des de zero per a començar amb una altra cerca.

La segona funció del botó és anar a la pàgina *web* d’inici un altre cop, de forma que es doni opció a l’usuari de poder escollir una altra llibreria digital diferent.

- 4- L'espai d'aquesta cel·la està destinat a representar gràficament la jerarquia de la informació estructurada que s'ha generat pel sistema a partir del terme clau introduït per l'usuari.

Quan entrem per primera vegada a la pàgina *web* que interactua amb el motor de cerca de l'aplicació, després d'haver escollit una llibreria digital en la pàgina *web* d'inici, aquesta cel·la se'ns presentarà buida, ja que no s'ha realitzat encara cap cerca, i per tant encara no hi ha cap jerarquia a dibuixar.

En canvi, quan realitzem una cerca, el sistema dibuixarà la taxonomia corresponent en aquest espai.

La estructuració de la informació es representa per nivells. És a dir, el terme arrel introduït per l'usuari es mostra en un nivell de profunditat de valor 0. Els subdominis trobats per a aquest terme arrel, es representaran amb una tabulació que correspondrà a un nivell de profunditat de valor 1, i així successivament.

En la jerarquia generada, a la part dreta de cada subdomini, hi haurà un nombre decimal que representarà la relació entre aquest terme i el terme al qual especialitza. Aquest tema l'hem explicat detalladament en l'apartat 3.1.2 que parla sobre la selecció de subdominis.

Per últim, dir que els termes de l'estructura s'escriuen en forma d'hiper-enllaç i que quan cliquem sobre qualsevol d'un d'ells, aquest apareixerà automàticament en el camp '*Search word*' de la segona cel·la per si es vol realitzar una cerca d'aquest, i ampliar així, els nivells de l'estructura.

- 5- I per últim, l'espai de la última cel·la està dedicat a la presentació de pàgines *web*. Es poden trobar els següents casos:

- Quan entrem per primera vegada a la pàgina *web* que interactua amb el motor de cerca de l'aplicació, després d'haver escollit una llibreria digital, en aquest *frame* se'ns mostrarà la pàgina *web* d'inici d'aquesta llibreria.
- Quan realitzem una cerca, se'ns mostrarà una pàgina *web* amb els articles que retorna la llibreria digital sobre la que estem buscant, els quals tindran relació amb el terme introduït per l'usuari.
- Com hem dit abans, els termes de l'estructura de l'arbre estan escrits en forma d'hiper-enllaç. D'aquesta forma quan es cliqui sobre qualsevol d'un d'ells, automàticament se'ns mostrarà la pàgina *web* amb els articles que retorna la llibreria digital sobre la que estem buscant, els quals tindran relació amb el terme que s'ha clicat.

Llavors, un cop explicats els components de la interfície *web*, com reacciona el *servlet* on hi ha el motor de cerca, quan es fa el *submit* des de la pàgina *web* que interactua amb el motor de cerca de l'aplicació?

Bé, quan l'usuari introdueix les dades en el formulari de la pàgina *web* que interactua amb el motor de cerca de l'aplicació i prem el botó '*Confirm Search*', es produeix un

submit i el control de l'aplicació passa del *JSP* al *servlet*. Aquest últim ja és capaç de llegir aquestes dades des de la sessió, de forma que el motor de cerca comença la seva tasca.

Primer que tot es comprova quin botó del formulari s'ha premut. En el cas en s'hagi premut '*New search*', el que es fa és esborrar l'estructura de dades que conté la jerarquia generada, de forma que aquesta queda lliure per a començar a introduir resultats d'una nova cerca. El control de l'aplicació torna una altra vegada a la pàgina *web* d'inici, de forma que es dona la possibilitat a l'usuari de poder escollir una altra llibreria digital per a realitzar una nova cerca.

En el cas que s'hagi premut el botó de '*Save search*', el que es fa és generar un fitxer '*htm*' amb l'estructuració de la informació representada gràficament.

En el cas que s'hagi premut el botó de '*Confirm Search*', poden succeir dues coses.

En el cas que l'usuari no hagi introduït cap terme clau a cercar, el control de l'aplicació passa un altre cop a la pàgina *web* que interactua amb el motor de cerca de l'aplicació com si no hagués succeït res.

En el cas que l'usuari sí hagi introduït algun terme, el primer que es fa és controlar si aquest terme ja havia estat cercat anteriorment de forma que no puguin existir cerques repetides en l'estructura de dades. Si és així el control de l'aplicació passa un altre cop a la pàgina *web* que interactua amb el motor de cerca de l'aplicació com si no hagués succeït res. En cas contrari es continuarà amb el desenvolupament de l'aplicació.

En la *Figura 7*, es mostra l'aspecte de la pàgina *web* que interactua amb el motor de cerca de l'aplicació un cop s'ha realitzat una cerca. Com es pot veure, el sistema retorna, a partir de la interfície *web*, l'estructuració de la informació que s'ha generat.



Figura 7. Pàgina *web* que interactua amb el motor de cerca de l'aplicació, on es pot veure l'estructuració de la informació generada.

En la *Figura 8*, i ja per a finalitzar, es mostrarà un esquema que ens permetrà veure gràficament com estan interconnectats tots els fitxers explicats abans en l'apartat 3.4 i quins paràmetres hi intervenen.

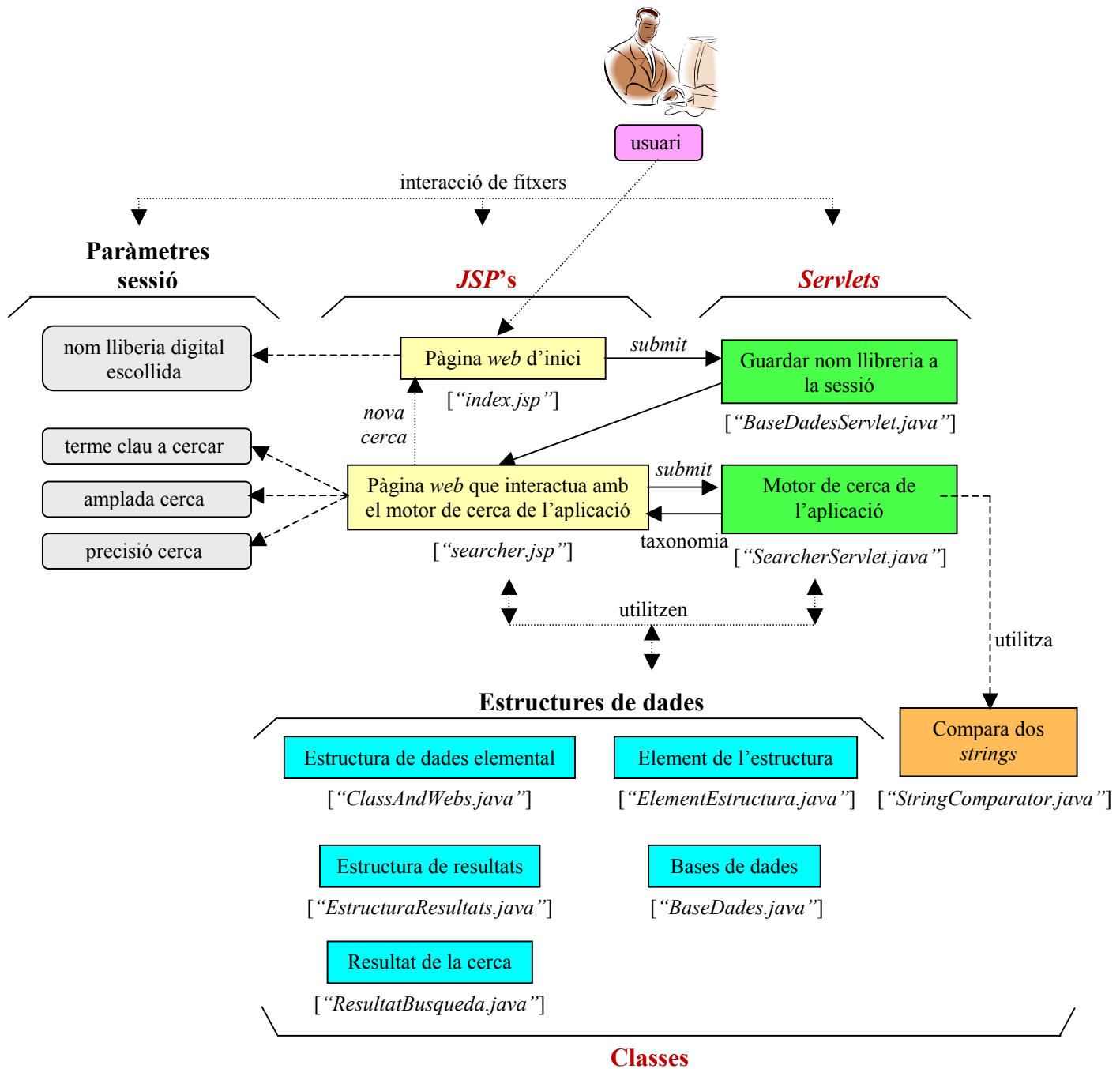


Figura 8. Esquema gràfic on es mostra com estan interconnectats tots els fitxers del sistema, així com els paràmetres que hi intervenen.

4- JOC DE PROVES

En aquest capítol es realitzaran una sèrie de proves de funcionament del sistema. Més exactament, s'observaran les estructures d'informació (taxonomies) generades per l'aplicació depenent dels paràmetres introduïts per l'usuari.

Recordem que el sistema necessita quatre paràmetres bàsics per al seu funcionament: el terme clau a partir del qual es construirà la taxonomia, la llibreria digital escollida, l'amplada de la cerca i la precisió de la cerca.

Si es fixa el paràmetre referit al terme clau i es varien els altres tres, les estructures d'informació referents a una llibreria determinada seran diferents. Per exemple:

- Depenent del terme clau introduït per l'usuari i de la llibreria digital escollida, el nombre de subdominis que formaran la jerarquia final pot ser major o menor. Per exemple, si el concepte introduït és “*cancer*”, i la B.D. escollida és ‘*PubMed*’, és molt probable que el número de subclasses seleccionades sigui més gran que si s’hagués escollit la B.D. *NASA ‘Astrophysics Data System’*. Això es degut a que la primera llibreria conté articles relacionats amb la medicina, en canvi, la segona conté articles de temes més aviat aerospacials i astronòmics.
- Un altre aspecte a tenir en compte, és l'amplada de la cerca. Si s’ha escollit una amplada més aviat extensa (‘*Complexa*’ o ‘*Molt complexa*’), es retornaran un nombre més gran de subdominis seleccionats. En canvi, per a una amplada de la cerca més aviat superficial (‘*Senzilla*’ o ‘*Mitjana*’), el número de subdominis retornats serà molt menor.
- I ja per finalitzar, l’últim paràmetre que ens variarà els resultats, serà la precisió de la cerca. Una precisió elevada (‘*0.001*’) ens assegurarà una selecció de subdominis d’una qualitat major, és a dir, amb una gran relació amb el domini al qual especifiquen (terme clau). En canvi, una precisió baixa (‘*0.00001*’), ens retornarà més subdominis que en el cas anterior, però amb una qualitat mitja inferior, és a dir, n’hi haurà alguns que tindran poca relació amb el domini que especifiquen.

Doncs bé, un cop s’ha comentat com intervenen els valors dels paràmetres en la generació de l’estructuració de la informació en forma de taxonomia, és el moment de veure alguns exemples del funcionament del sistema. És el que s’anomenarà el joc de proves.

4.1- Prova 1

Llibreria digital:	<i>PubMed</i>
Terme clau:	<i>"cancer"</i>
Amplada cerca:	<i>Senzilla (80)</i>
Precisió cerca:	<i>Elevada (0.001)</i>
Temps aproximat	4.5 minuts

Candidat a subdomini	Nombre resultats	subdomini + terme clau	Nombre resultats	Relació	Decisió
gastric	31600000	gastric cancer	2710000	0.0857	subdomini seleccionat
human	1950000000	human cancer	2320000	0.0011	subdomini seleccionat
pancreatic	18000000	pancreatic cancer	7100000	0.3944	subdomini seleccionat
prostate	62400000	prostate cancer	36100000	0.5785	subdomini seleccionat
ovarian	24100000	ovarian cancer	13900000	0.5767	subdomini seleccionat
different	0				subdomini no seleccionat
breast	146000000	breast cancer	103000000	0.7054	subdomini seleccionat
comprehensive	1010000000	comprehensive cancer	5560000	0.0055	subdomini seleccionat
oesophageal	1690000	oesophageal cancer	238000	0.1408	subdomini seleccionat
hypopharyngeal	202000	hypopharyngeal cancer	65100	0.3222	subdomini seleccionat
maccallum	1620000	maccallum cancer	138000	0.0851	subdomini seleccionat
cellular	208000000	cellular cancer	995	0.000004783	subdomini no seleccionat
metastatic	7410000	metastatic cancer	937000	0.1264	subdomini seleccionat
Pàgina 1 (20 articles analitzats) / LearningRate = 84.61538461538461 > 80 --> continua l'aplicació					
colon	66700000	colon cancer	15300000	0.2293	subdomini seleccionat
brown	748000000	brown cancer	54400	0.00007272	subdomini no seleccionat
colorectal	17700000	colorectal cancer	12200000	0.6892	subdomini seleccionat
lung	123000000	lung cancer	33400000	0.2715	subdomini seleccionat
natl	27300000	natl cancer	2610000	0.0956	subdomini seleccionat
renal	41600000	renal cancer	805000	0.0193	subdomini seleccionat
monitoring	645000000	monitoring cancer	23600	0.000036589	subdomini no seleccionat
bladder	25300000	bladder cancer	3970000	0.1569	subdomini seleccionat
Pàgina 2 (40 articles analitzats) / LearningRate = 80.95238095238095 > 80 --> continua l'aplicació					
anderson	255000000	anderson cancer	3560000	0.0139	subdomini seleccionat
affect	479000000	affect cancer	111000	0.00023173	subdomini no seleccionat
residual	41700000	residual cancer	69900	0.0016	subdomini seleccionat
invasive	72300000	invasive cancer	385000	0.0053	subdomini seleccionat
esophageal	8660000	esophageal cancer	1830000	0.2113	subdomini seleccionat
bmc	32300000	bmc cancer	170000	0.0052	subdomini seleccionat
clinical	537000000	clinical cancer	1630000	0.0030	subdomini seleccionat
acute	109000000	acute cancer	9630	0.000088348	subdomini no seleccionat
Pàgina 3 (60 articles analitzats) / LearningRate = 79.3103448275862 < 80 --> finalitza l'aplicació					

Com es pot veure en aquesta prova, els subdominis seleccionats són més aviat pocs, ja que s'ha realitzat amb una amplada de cerca senzilla i amb una precisió elevada.

Aquests paràmetres fan però, que els subdominis retornats siguin en general d'una alta qualitat. Com a exemple d'aquests tenim: *gastric, pancreatic, prostate, ovarian, breast, oesophageal, hypopharyngeal, colon, colorectal, lung, renal, bladder* i *esophageal*. Tots ells corresponen a tipus de càncer determinats. Hi ha d'altres però, que s'han seleccionat i que no corresponen a tipus de càncer, com pot ser el terme *anderson* entre d'altres. Aquest fa referència a una associació d'ajuda contra el càncer.

Com es pot veure, no s'han seleccionat termes com: *different, brown, monitoring, affect*, els quals tindrien molt poca, o cap relació amb el concepte "*cancer*".

En la *Figura 9*, es pot veure l'estructura d'informació final que retornaria l'aplicació mitjançant la interfície gràfica, utilitzant els paràmetres abans esmentats. S'hi pot observar que els subdominis seleccionats estan ordenats alfabèticament i que a la seva dreta hi ha un nombre (amb un valor de 0 a 1), que especifica la relació que hi ha entre aquest subdomini i el terme al qual especialitza.

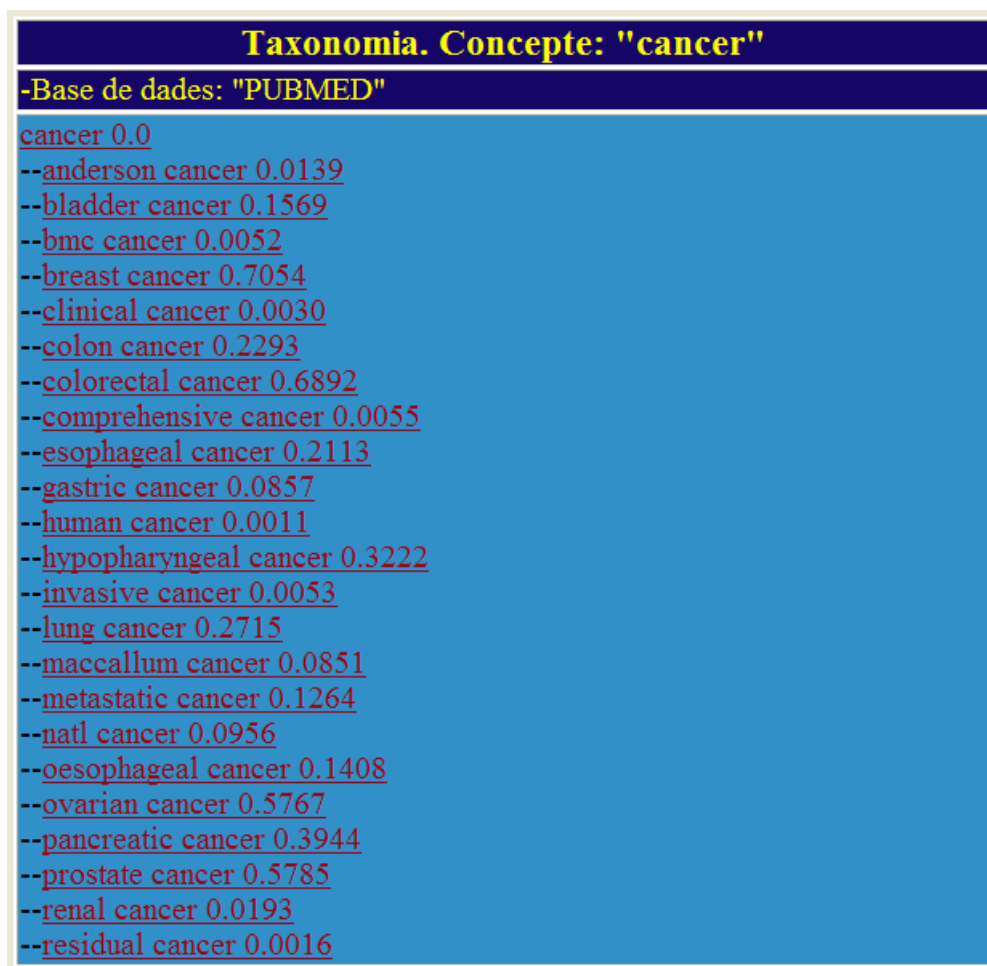


Figura 9. Taxonomia construïda a partir del terme "*cancer*".

4.2- Prova 2

En aquesta prova hem decidit ampliar la taxonomia anterior a partir del terme “*breast cancer*”.

Llibreria digital:	<i>PubMed</i>
Terme clau:	<i>"breast cancer"</i>
Amplada cerca:	<i>Senzilla (80)</i>
Precisió cerca:	<i>Elevada (0.001)</i>
Temps aproximat	2 minuts

Candidat a subdomini	Nombre resultats	subdomini + terme clau	Nombre resultats	Relació	Decisió
human	0				subdomini no seleccionat
invasive	77400000	invasive breast cancer	578000	0.0074	subdomini seleccionat
metastatic	7410000	metastatic breast cancer	1130000	0.1524	subdomini seleccionat
positive	448000000	positive breast cancer	378000	0.00084375	subdomini no seleccionat
risk	1120000000	risk breast cancer	107000	0.000095535	subdomini no seleccionat
contralateral	2430000	contralateral breast cancer	64600	0.0265	subdomini seleccionat
adjuvant	5400000	adjuvant breast cancer	58400	0.0108	subdomini seleccionat
stage	688000000	stage breast cancer	610000	0.00088662	subdomini no seleccionat
recurrent	25600000	recurrent breast cancer	718000	0.028046875	subdomini seleccionat
responsive	101000000	responsive breast cancer	31400	0.00031089	subdomini no seleccionat
Pàgina 1 (20 articles analitzats) / LearningRate = 50 < 80 --> finalitza l'aplicació					

Com es pot observar, utilitzant els mateixos paràmetres que la prova anterior, pel que fa a l'amplada i la precisió de la cerca, se'n retornen menys subdominis. És a dir, a mesura que les especialitzacions són més concretes, es troben menys termes candidats.

La Figura 10 mostra el resultat de la taxonomia de la *Prova 1* anterior construïda a partir del terme “cancer”, i ampliada en aquesta prova a partir del terme “breast cancer”.

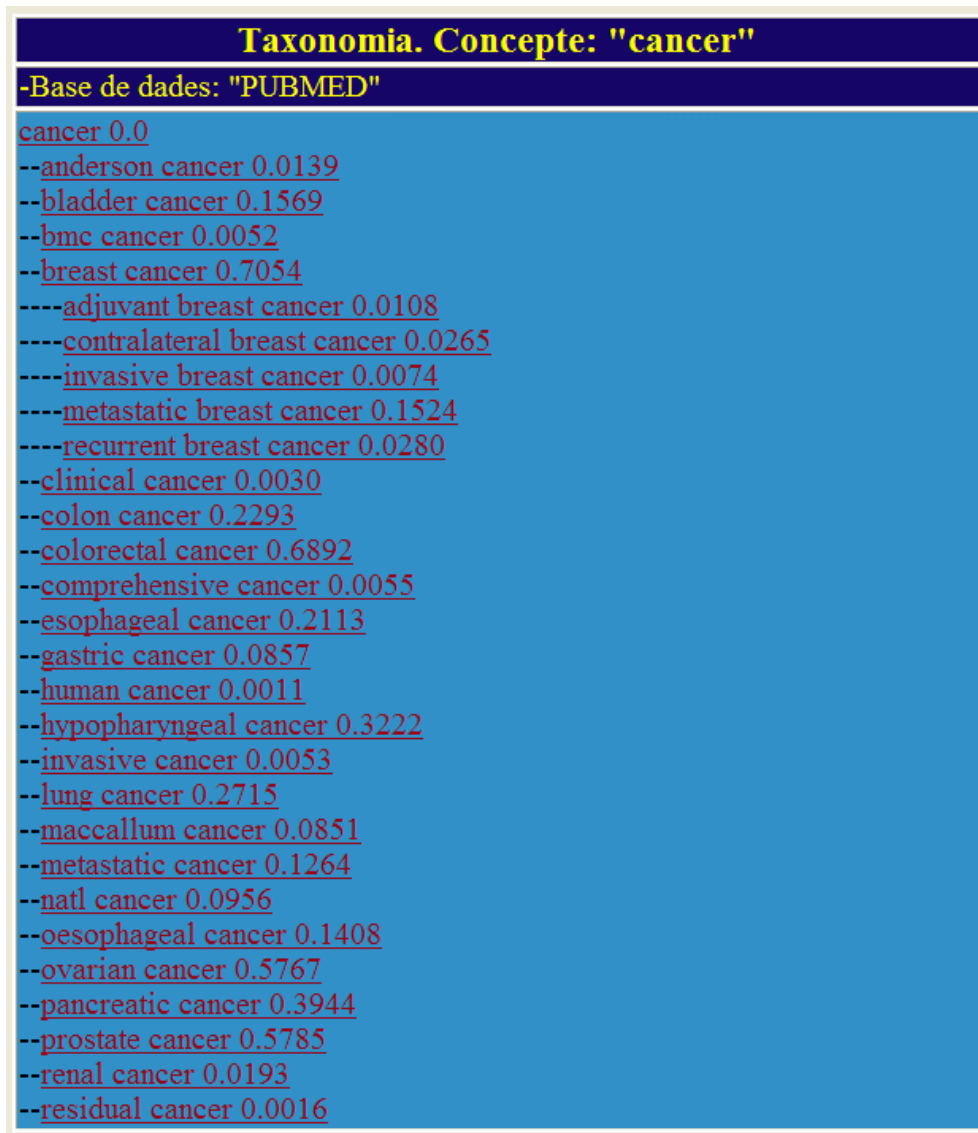


Figura 10. Taxonomia construïda a partir del terme "*cancer*" i ampliada a partir del terme "*breast cancer*".

4.3- Prova 3

Aquesta prova és molt similar a la primera. Solament varia el valor del paràmetre de la precisió de la cerca. En aquesta cerca, la selecció de subdominis no és tan restrictiva, i per tant, el número de subdominis que compondran la taxonomia pot ser major.

Llibreria digital:	<i>PubMed</i>
Terme clau:	<i>"cancer"</i>
Amplada cerca:	<i>Senzilla (80)</i>
Precisió cerca:	<i>Mitja (0.0001)</i>
Temps aproximat	6 minuts

Candidat a subdomini	Nombre resultats	subdomini + terme clau	Nombre resultats	Relació	Decisió
gastric	31600000	gastric cancer	2230000	0.0705	subdomini seleccionat
human	1950000000	human cancer	2890000	0.0014	subdomini seleccionat
pancreatic	17000000	pancreatic cancer	7100000	0.4176	subdomini seleccionat
prostate	62400000	prostate cancer	30400000	0.4871	subdomini seleccionat
ovarian	18600000	ovarian cancer	13100000	0.7043	subdomini seleccionat
different	2110000000	different cancer	132000	0.000062559	subdomini no seleccionat
breast	146000000	breast cancer	83700000	0.5732	subdomini seleccionat
comprehensive	860000000	comprehensive cancer	8380000	0.0097	subdomini seleccionat
oesophageal	1800000	oesophageal cancer	226000	0.1255	subdomini seleccionat
hypopharyngeal	202000	hypopharyngeal cancer	65100	0.3222	subdomini seleccionat
maccallum	1310000	maccallum cancer	138000	0.1053	subdomini seleccionat
cellular	256000000	cellular cancer	971	0.000003792	subdomini no seleccionat
metastatic	10600000	metastatic cancer	937000	0.0883	subdomini seleccionat
Pàgina 1 (20 articles analitzats) / LearningRate = 84.61538461538461 > 80 --> continua l'aplicació					
colon	86800000	colon cancer	19700000	0.2269	subdomini seleccionat
brown	565000000	brown cancer	54400	0.000096283	subdomini no seleccionat
colorectal	17700000	colorectal cancer	9520000	0.5378	subdomini seleccionat
lung	123000000	lung cancer	33400000	0.2715	subdomini seleccionat
natl	37200000	natl cancer	2610000	0.0701	subdomini seleccionat
renal	42300000	renal cancer	805000	0.0190	subdomini seleccionat
monitoring	542000000	monitoring cancer	24300	0.000044833	subdomini no seleccionat
bladder	25300000	bladder cancer	3970000	0.1569	subdomini seleccionat
Pàgina 2 (40 articles analitzats) / LearningRate = 80.95238095238095 > 80 --> continua l'aplicació					

anderson	327000000	anderson cancer	3560000	0.0108	subdomini seleccionat
affect	566000000	affect cancer	111000	0.00019611	subdomini seleccionat
residual	55100000	residual cancer	69900	0.0012	subdomini seleccionat
invasive	72300000	invasive cancer	385000	0.0053	subdomini seleccionat
esophageal	6190000	esophageal cancer	1830000	0.2956	subdomini seleccionat
bmc	32300000	bmc cancer	177000	0.0054	subdomini seleccionat
clinical	537000000	clinical cancer	1630000	0.0030	subdomini seleccionat
acute	109000000	acute cancer	991	0.000009091	subdomini no seleccionat
Pàgina 3 (60 articles analitzats) / LearningRate = 82.75862068965517 > 80 --> continua l'aplicació					
skin	478000000	skin cancer	14000000	0.0292	subdomini seleccionat
rectosigmoid	222000	rectosigmoid cancer	10500	0.0472	subdomini seleccionat
total	1350000000	total cancer	116000	0.000085925	subdomini no seleccionat
childhood	216000000	childhood cancer	2120000	0.0098	subdomini seleccionat
understanding	1070000000	understanding cancer	161000	0.00015046	subdomini seleccionat
eradicate	17300000	eradicate cancer	35000	0.0020	subdomini seleccionat
neck	99500000	neck cancer	3510000	0.0352	subdomini seleccionat
thyroid	26700000	thyroid cancer	4370000	0.1636	subdomini seleccionat
blood	515000000	blood cancer	445000	0.00086407	subdomini seleccionat
Pàgina 4 (80 articles analitzats) / LearningRate = 84.21052631578947 > 80 --> continua l'aplicació					
cervical	34200000	cervical cancer	13000000	0.3801	subdomini seleccionat
rectal	13700000	rectal cancer	2580000	0.1883	subdomini seleccionat
national	0				subdomini no seleccionat
taussig	993000	taussig cancer	46300	0.0466	subdomini seleccionat
good	0				subdomini no seleccionat
family	2060000000	family cancer	151000	0.0000733	subdomini no seleccionat
Pàgina 5 (100 articles analitzats) / LearningRate = 79.54545454545455 < 80 --> finalitza l'aplicació					

Veiem que el terme *affect*, no ha estat seleccionat en la primera prova. En canvi, sí ho ha estat en aquesta, degut a que la precisió de cerca és més baixa. També serà més baixa, però, qualitat dels termes seleccionats.

La selecció d'aquest terme ha fet pujar el learning rate en acabar d'analitzar la pàgina 3, de manera que s'han analitzat 40 articles més que en la *Prova 1*.

4.4- Prova 4

Aquesta prova és igual que la primera, solament que s'ha escollit una altra llibreria digital (amb articles relacionats amb la informàtica). L'objectiu d'aquesta prova és veure que la construcció d'una taxonomia a partir d'un terme que no està relacionat amb el tema central dels articles de la llibreria, implica que els subdominis retornats pel sistema seran menors. En la *Prova 1*, per exemple, es retornaven molts més subdominis que en aquesta tot i utilitzar els mateixos paràmetres, ja que s'havia utilitzat *PubMed*, llibreria referent amb articles de medicina.

Llibreria digital:	<i>IEEE. Computer Society</i>
Terme clau:	<i>"cancer"</i>
Amplada cerca:	<i>Senzilla (80)</i>
Precisió cerca:	<i>Elevada (0.001)</i>
Temps aproximat	1 minut

Candidat a subdomini	Nombre resultats	subdomini + terme clau	Nombre resultats	Relació	Decisió
p53	9010000	p53 cancer	30400	0.0033	subdomini seleccionat
prostate	50800000	prostate cancer	36100000	0.7106	subdomini seleccionat
informative	150000000	informative cancer	148	0.000000986	subdomini no seleccionat
multiple	1250000000	multiple cancer	63600	0.00005088	subdomini no seleccionat
accurate	549000000	accurate cancer	94300	0.0001717	subdomini no seleccionat
breast	179000000	breast cancer	83700000	0.4675	subdomini seleccionat
precursory	305000	precursory cancer	22	0.000072131	subdomini no seleccionat
ovarian	24100000	ovarian cancer	13100000	0.5435	subdomini seleccionat
scheme	299000000	scheme cancer	197	0.000000658	subdomini no seleccionat
finding	906000000	finding cancer	79400	0.000087637	subdomini no seleccionat
Pàgina 1 (20 articles analitzats) / LearningRate = 40 < 80 --> finalitza l'aplicació					

Com es pot observar, solament s'han seleccionat quatre subdominis, i d'aquests, un amb una relació molt baixa (*p53*).

4.5- Prova 5

Aquesta prova és similar a l'anterior, encara que ara s'ha escollit la llibreria anomenada "Association for Computing Machinery. The Guide", la qual tracta també d'articles informàtics.

Llibreria digital:	<i>Association for Computing Machinery. The Guide</i>
Terme clau:	"cancer"
Amplada cerca:	<i>Senzilla</i> (80)
Precisió cerca:	<i>Elevada</i> (0.001)
Temps aproximat	1 minut

Candidat a subdomini	Nombre resultats	subdomini + terme clau	Nombre resultats	Relació	Decisió
prostate	59500000	prostate cancer	36600000	0.6151	subdomini seleccionat
national	0				subdomini no seleccionat
elevated	66600000	elevated cancer	39400	0.00059159	subdomini no seleccionat
illinois	763000000	illinois cancer	15400	0.000020183	subdomini no seleccionat
colorectal	18200000	colorectal cancer	8820000	0.4846	subdomini seleccionat
iowa	441000000	iowa cancer	77800	0.00017641	subdomini no seleccionat
gastric	37300000	gastric cancer	2780000	0.0745	subdomini seleccionat
expression	420000000	expression cancer	62300	0.00014833	subdomini no seleccionat
diagnose	75100000	diagnose cancer	97800	0.0013	subdomini seleccionat
classify	44400000	classify cancer	700	0.000015765	subdomini no seleccionat
breast	184000000	breast cancer	103000000	0.5597	subdomini seleccionat
hutchinson	42200000	hutchinson cancer	1430000	0.0338	subdomini seleccionat
state	0				subdomini no seleccionat
Pàgina 1 (20 articles analitzats) / LearningRate = 46.15384615384615 < 80 --> finalitza l'aplicació					

4.6- Prova 6

Aquesta prova, amb el mateix objectiu que la *Prova 4* i la *Prova 5*, s'ha realitzat amb la llibreria "*NASA Astrophysics Data System*", que conté articles majoritàriament aeroespacials i d'aeronàutica.

Llibreria digital:	<i>NASA Astrophysics Data System</i>
Terme clau:	"cancer"
Amplada cerca:	<i>Senzilla</i> (80)
Precisió cerca:	<i>Elevada</i> (0.001)
Temps aproximat	1 minut

Candidat a subdomini	Nombre resultats	subdomini + terme clau	Nombre resultats	Relació	Decisió
human	0				subdomini no seleccionat
breast	184000000	breast cancer	105000000	0.5706	subdomini seleccionat
lions	75900000	lions cancer	652	0.00000859	subdomini no seleccionat
colorectal	18200000	colorectal cancer	8820000	0.4846	subdomini seleccionat
regional	1220000000	regional cancer	1580000	0.0012	subdomini seleccionat
german	922000000	german cancer	430000	0.0004663	subdomini no seleccionat
prostate	59300000	prostate cancer	36600000	0.6172	subdomini seleccionat
comprehensive	998000000	comprehensive cancer	7560000	0.0075	subdomini seleccionat
colon	87000000	colon cancer	19900000	0.2287	subdomini seleccionat
generation	977000000	generation cancer	15800	0.000016171	subdomini no seleccionat
Pàgina 1 (20 articles analitzats) / LearningRate = 60 < 80 --> finalitza l'aplicació					

4.7- Prova 7

Llibreria digital:	<i>IEEE Computer Society</i>
Terme clau:	<i>"cancer"</i>
Amplada cerca:	<i>Molt Complexa (40)</i>
Precisió cerca:	<i>Baixa (0.00001)</i>
Temps aproximat	4 minuts

Candidat a subdomini	Nombre resultats	subdomini + terme clau	Nombre resultats	Relació	Decisió
p53	9090000	p53 cancer	30000	0.0033	subdomini seleccionat
prostate	53500000	prostate cancer	32800000	0.6130	subdomini seleccionat
informative	242000000	informative cancer	146	0.0006033	subdomini no seleccionat
multiple	1650000000	multiple cancer	66400	0.00004024	subdomini seleccionat
accurate	435000000	accurate cancer	94100	0.0002163	subdomini seleccionat
breast	192000000	breast cancer	112000000	0.5833	subdomini seleccionat
precursory	345000	precursory cancer	27	0.0000782	subdomini seleccionat
ovarian	18300000	ovarian cancer	15700000	0.8579	subdomini seleccionat
scheme	289000000	scheme cancer	206	0.000000712	subdomini no seleccionat
finding	715000000	finding cancer	52900	0.0000739	subdomini seleccionat
Pàgina 1 (20 articles analitzats) / LearningRate = 80 > 40 --> continua l'aplicació					
national	0				subdomini no seleccionat
cervical	31600000	cervical cancer	18900000	0.5981	subdomini seleccionat
heterogeneous	51600000	heterogeneous cancer	239	0.00000463	subdomini no seleccionat
clinical	556000000	clinical cancer	1990000	0.0035	subdomini seleccionat
Pàgina 2 (40 articles analitzats) / LearningRate = 71.42 > 40 --> continua l'aplicació					
colorectal	18200000	colorectal cancer	8820000	0.4846	subdomini seleccionat
lung	94600000	lung cancer	44200000	0.4672	subdomini seleccionat
skin	450000000	skin cancer	13400000	0.0297	subdomini seleccionat
thyroid	24700000	thyroid cancer	5030000	0.2036	subdomini seleccionat
gynecologic	13500000	gynecologic cancer	575000	0.0425	subdomini seleccionat
comprehensive	1020000000	comprehensive cancer	11200000	0.0109	subdomini seleccionat
Pàgina 3 (60 articles analitzats) / LearningRate = 80 > 40 --> continua l'aplicació					
analysing	23100000	analysing cancer	162	0.00000701	subdomini no seleccionat
bladder	23900000	bladder cancer	6490000	0.2715	subdomini seleccionat
expression	322000000	expression cancer	59900	0.000186	subdomini seleccionat
Pàgina 4 (80 articles analitzats) / LearningRate = 78.26 > 40 --> continua l'aplicació					

unknown	297000000	unknown cancer	13200	0.0000444	subdomini seleccionat
hutchinson	42900000	hutchinson cancer	4390000	0.1023	subdomini seleccionat
kills	78900000	kills cancer	110000	0.0013	subdomini seleccionat
brca2	1950000	brca2 cancer	999	0.000512	subdomini seleccionat
uk	2100000000	uk cancer	133000	0.0000633	subdomini seleccionat

Pàgina 5 (100 articles analitzats) / LearningRate = 82.14 > 40 --> continua l'aplicació

L'aplicació finalitza perquè s'acaben els articles retonats per la llibreria digital

Aquesta prova, la número 7, és molt interessant. S'ha realitzat sobre el repositori “*IEEE Computer Society*”, que com hem dit abans tracta majoritàriament d'articles d'informàtica. Doncs bé, a l'intentar crear una taxonomia sobre aquesta llibreria a partir del terme “*cancer*”, aquesta ens retorna pocs articles (exactament 100), degut a que aquest terme clau fa referència més aviat a la branca de la medicina, com hem pogut observar en les proves anteriors amb *PubMed*.

Llavors, si realitzem la cerca amb una amplada molt complexa i una precisió molt baixa, succeeix que quan acabem d'analitzar la plana 5 (100 articles), el learning rate encara és major que 40, més concretament, és de 82.14. El sistema, per tant intenta analitzar més planes web d'articles, cosa no possible degut a que ja no n'hi ha més.

Aquest aspecte s'ha tingut en compte a l'hora de dissenyar l'aplicació, per tant el sistema finalitzarà correctament en cas de que intenti analitzar pàgines *web* retornades pel repositori, les quals, no tinguin articles.

4.8- Prova 8

Aquesta prova és la més llarga i més exhaustiva de totes les que s'han realitzat en aquest capítol del joc de proves. Això es degut a dues raons molt clares. Una, que el valor de l'amplada de la cerca és molt complexa. I l'altra, que es vol estructurar la informació a partir del terme “*cancer*”, sobre la llibreria referent amb articles de medicina, “*PubMed*”. Per tant, és evident que es trobaran i seleccionaran molts candidats a formar part de la taxonomia final.

Llibreria digital:	<i>PubMed</i>
Terme clau:	" <i>cancer</i> "
Amplada cerca:	<i>Molt complexa</i> (40)
Precisió cerca:	<i>Elevada</i> (0.001)
Temps aproximat	25 minuts

Candidat a subdomini	Nombre resultats	subdomini + terme clau	Nombre resultats	Relació	Decisió
neck	103000000	neck cancer	2190000	0.0212	subdomini seleccionat
blood	504000000	blood cancer	746000	0.0014	subdomini seleccionat
childhood	237000000	childhood cancer	4300000	0.0181	subdomini seleccionat
tumor	74800000	tumor cancer	117000	0.0015	subdomini seleccionat
breast	145000000	breast cancer	85000000	0.5862	subdomini seleccionat
national	0				subdomini no seleccionat
tract	55400000	tract cancer	415000	0.0074	subdomini seleccionat
duct	34400000	duct cancer	169000	0.0049	subdomini seleccionat
human	0				subdomini no seleccionat
lung	121000000	lung cancer	36400000	0.3008	subdomini seleccionat
colon	84600000	colon cancer	20900000	0.2470	subdomini seleccionat
colorectal	18200000	colorectal cancer	8820000	0.4846	subdomini seleccionat
Pàgina 1 (20 articles analitzats) / LearningRate = 83.33 > 40 --> continua l'aplicació					

Invasive	79400000	invasive cancer	386000	0.0048	subdomini seleccionat
swedish	245000000	swedish cancer	144000	0.000587	subdomini no seleccionat
gastric	44400000	gastric cancer	2830000	0.0637	subdomini seleccionat
prostate	63600000	prostate cancer	43400000	0.6823	subdomini seleccionat
cervical	31600000	cervical cancer	18900000	0.5981	subdomini seleccionat
nonlung	901	nonlung cancer	84	0.0932	subdomini seleccionat
endometrial	6810000	endometrial cancer	3120000	0.4581	subdomini seleccionat
alberta	198000000	alberta cancer	105000	0.00053	subdomini no seleccionat
hutchinson	42900000	hutchinson cancer	1600000	0.0372	subdomini seleccionat
regional	1250000000	regional cancer	1390000	0.0011	subdomini seleccionat
adjuvant	7300000	adjuvant cancer	17500	0.0023	subdomini seleccionat
tailor	83800000	tailor cancer	406	0.00000484	subdomini no seleccionat
gynecologic	13500000	gynecologic cancer	575000	0.0425	subdomini seleccionat
valley	597000000	valley cancer	31400	0.0000525	subdomini no seleccionat
Pàgina 2 (40 articles analitzats) / LearningRate = 76.92 > 40 --> continua l'aplicació					
anderson	325000000	anderson cancer	5830000	0.0179	subdomini seleccionat
rare	478000000	rare cancer	328000	0.000686	subdomini no seleccionat
rectal	14100000	rectal cancer	1820000	0.1290	subdomini seleccionat
niigata	4860000	niigata cancer	10400	0.0021	subdomini seleccionat
shikoku	4250000	shikoku cancer	9540	0.0022	subdomini seleccionat
Pàgina 3 (60 articles analitzats) / LearningRate = 77.41 > 40 --> continua l'aplicació					
metropolitan	208000000	metropolitan cancer	268	0.00000128	subdomini no seleccionat
comprehensive	865000000	comprehensive cancer	11200000	0.0129	subdomini seleccionat
farber	17400000	farber cancer	5670000	0.3258	subdomini seleccionat
bone	172000000	bone cancer	2790000	0.0162	subdomini seleccionat
bates	52900000	bates cancer	209	0.00000395	subdomini no seleccionat
lurie	5610000	lurie cancer	14300	0.0025	subdomini seleccionat
moffitt	4230000	moffitt cancer	359000	0.0848	subdomini seleccionat
hoed	3190000	hoed cancer	42700	0.0133	subdomini seleccionat
netherlands	504000000	netherlands cancer	281000	0.000557	subdomini no seleccionat
pharyngeal	1790000	pharyngeal cancer	71500	0.0399	subdomini seleccionat
family	0				subdomini no seleccionat
ovarian	26200000	ovarian cancer	15700000	0.5992	subdomini seleccionat
opsonize	33600	opsonize cancer	1	0.0000287	subdomini no seleccionat
pittsburgh	276000000	pittsburgh cancer	179000	0.000648	subdomini no seleccionat
Pàgina 4 (80 articles analitzats) / LearningRate = 71.1111 > 40 --> continua l'aplicació					
uterine	13800000	uterine cancer	3470000	0.2514	subdomini seleccionat
common	1830000000	common cancer	1180000	0.000644	subdomini no seleccionat
testicular	11200000	testicular cancer	5800000	0.5178	subdomini seleccionat
oesophageal	1930000	oesophageal cancer	243000	0.1259	subdomini seleccionat
cervix	7440000	cervix cancer	178000	0.0239	subdomini seleccionat
readable	192000000	readable cancer	19	0.000000098	subdomini no seleccionat
nano	185000000	nano cancer	17200	0.0000929	subdomini no seleccionat
rectum	7230000	rectum cancer	86900	0.0120	subdomini seleccionat
Pàgina 5 (100 articles analitzats) / LearningRate = 69.81 > 40 --> continua l'aplicació					

pancreatic	13700000	pancreatic cancer	5800000	0.4233	subdomini seleccionat
kettering	13700000	kettering cancer	4170000	0.3043	subdomini seleccionat
novel	283000000	novel cancer	215000	0.000759	subdomini no seleccionat
maxillary	1960000	maxillary cancer	1030	0.000525	subdomini no seleccionat
identical	219000000	identical cancer	127	0.000000579	subdomini no seleccionat
type	0				subdomini no seleccionat
Pàgina 6 (120 articles analitzats) / LearningRate = 66.10 > 40 --> continua l'aplicació					
gynaecological	1600000	gynaecological cancer	126000	0.0787	subdomini seleccionat
links	0				subdomini no seleccionat
indigenous	98800000	indigenous cancer	18300	0.000185	subdomini no seleccionat
Pàgina 7 (140 articles analitzats) / LearningRate = 64.51 > 40 --> continua l'aplicació					
bladder	32000000	bladder cancer	3810000	0.1190	subdomini seleccionat
penile	5900000	penile cancer	1250000	0.2118	subdomini seleccionat
nonprostate	733	nonprostate cancer	273	0.3724	subdomini seleccionat
increases	352000000	increases cancer	67300	0.000191	subdomini no seleccionat
confirm	412000000	confirm cancer	590	0.00000143	subdomini no seleccionat
ureteral	903000	ureteral cancer	11400	0.0126	subdomini seleccionat
german	921000000	german cancer	455000	0.000494	subdomini no seleccionat
Pàgina 8 (160 articles analitzats) / LearningRate = 63.76 > 40 --> continua l'aplicació					
resistant	117000000	resistant cancer	70100	0.000599	subdomini no seleccionat
university	0				subdomini no seleccionat
kills	78900000	kills cancer	110000	0.0013	subdomini seleccionat
bmc	33700000	bmc cancer	324000	0.0096	subdomini seleccionat
thyroid	24700000	thyroid cancer	5030000	0.2036	subdomini seleccionat
park	834000000	park cancer	1010000	0.0012	subdomini seleccionat
original	1880000000	original cancer	44800	0.0000238	subdomini no seleccionat
Pàgina 9 (180 articles analitzats) / LearningRate = 63.15 > 40 --> continua l'aplicació					
incurable	7540000	incurable cancer	92800	0.01230	subdomini seleccionat
destroy	190000000	destroy cancer	257000	0.0013	subdomini seleccionat
unusual	257000000	unusual cancer	935	0.00000363	subdomini no seleccionat
mammary	7670000	mammary cancer	275000	0.0358	subdomini seleccionat
Pàgina 10 (200 articles analitzats) / LearningRate = 63.74 > 40 --> continua l'aplicació					
Pàgina 11 (220 articles analitzats) / LearningRate = 63.74 > 40 --> continua l'aplicació					
ingram	30100000	ingram cancer	194000	0.0064	subdomini seleccionat
major	1770000000	major cancer	192000	0.000108	subdomini no seleccionat
clinical	447000000	clinical cancer	1990000	0.0044	subdomini seleccionat
metastatic	11100000	metastatic cancer	1320000	0.1189	subdomini seleccionat
Pàgina 12 (240 articles analitzats) / LearningRate = 64.28 > 40 --> continua l'aplicació					

liver	89800000	liver cancer	4270000	0.0475	subdomini seleccionat
skin	450000000	skin cancer	13400000	0.0297	subdomini seleccionat
laryngeal	2920000	laryngeal cancer	454000	0.1554	subdomini seleccionat
larynx	5460000	larynx cancer	80200	0.0146	subdomini seleccionat
suppress	47500000	suppress cancer	21900	0.000461	subdomini no seleccionat
suppresses	11000000	suppresses cancer	908	0.0000825	subdomini no seleccionat
bc	238000000	bc cancer	442000	0.0018	subdomini seleccionat
Pàgina 13 (260 articles analitzats) / LearningRate = 64.83 > 40 --> continua l'aplicació					
esophageal	8730000	esophageal cancer	3290000	0.3768	subdomini seleccionat
adult	561000000	adult cancer	244000	0.000434	subdomini no seleccionat
older	746000000	older cancer	48700	0.0000652	subdomini no seleccionat
quality	1940000000	quality cancer	241000	0.000124	subdomini no seleccionat
unique	1040000000	unique cancer	15700	0.000015	subdomini no seleccionat
dependent	324000000	dependent cancer	42900	0.000132	subdomini no seleccionat
Pàgina 14 (280 articles analitzats) / LearningRate = 61.85 > 40 --> continua l'aplicació					
rational	124000000	rational cancer	625	0.00000504	subdomini no seleccionat
primary	1050000000	primary cancer	319000	0.000303	subdomini no seleccionat
aids	420000000	aids cancer	391000	0.00093	subdomini no seleccionat
cell	1070000000	cell cancer	2680000	0.0025	subdomini seleccionat
oropharyngeal	890000	oropharyngeal cancer	103000	0.1157	subdomini seleccionat
Pàgina 15 (300 articles analitzats) / LearningRate = 60.78 > 40 --> continua l'aplicació					
new	0				subdomini no seleccionat
minority	287000000	minority cancer	33700	0.000117	subdomini no seleccionat
moroccan	20500000	moroccan cancer	19	0.000000926	subdomini no seleccionat
elena	39800000	elena cancer	18900	0.000474	subdomini no seleccionat
oral	335000000	oral cancer	3160000	0.0094	subdomini seleccionat
edinburgh	174000000	edinburgh cancer	705	0.00000405	subdomini no seleccionat
Pàgina 16 (320 articles analitzats) / LearningRate = 58.33 > 40 --> continua l'aplicació					
purdue	98200000	purdue cancer	19300	0.000196	subdomini no seleccionat
net	0				subdomini no seleccionat
nasopharynx	799000	nasopharynx cancer	13400	0.0167	subdomini seleccionat
Pàgina 17 (340 articles analitzats) / LearningRate = 57.65 > 40 --> continua l'aplicació					
schiffler	177000	schiffler cancer	615	0.0034	subdomini seleccionat
tulane	21600000	tulane cancer	24600	0.0011	subdomini seleccionat
Pàgina 18 (360 articles analitzats) / LearningRate = 58.40 > 40 --> continua l'aplicació					
hebei	4710000	hebei cancer	147	0.0000312	subdomini no seleccionat
Pàgina 19 (380 articles analitzats) / LearningRate = 57.89 > 40 --> continua l'aplicació					

renal	31200000	renal cancer	995000	0.03189	subdomini seleccionat
residual	54900000	residual cancer	68600	0.0012	subdomini seleccionat
Pàgina 20 (400 articles analitzats) / LearningRate = 58.62 > 40 --> continua l'aplicació					
bowel	28400000	bowel cancer	1270000	0.0447	subdomini seleccionat
Pàgina 21 (420 articles analitzats) / LearningRate = 58.97 > 40 --> continua l'aplicació					
Pàgina 22 (440 articles analitzats) / LearningRate = 58.97 > 40 --> continua l'aplicació					
chase	154000000	chase cancer	952000	0.0061	subdomini seleccionat
detects	56200000	detects cancer	15500	0.000275	subdomini no seleccionat
anal	38800000	anal cancer	742000	0.0191	subdomini seleccionat
recurrent	29400000	recurrent cancer	126000	0.0042	subdomini seleccionat
hunterman	4380000	hunterman cancer	161000	0.0367	subdomini seleccionat
wayne	233000000	wayne cancer	66000	0.000283	subdomini no seleccionat
2breast	16800	2breast cancer	6	0.000357	subdomini no seleccionat
candidate	378000000	candidate cancer	12200	0.0000322	subdomini no seleccionat
Pàgina 23 (460 articles analitzats) / LearningRate = 58.4 > 40 --> continua l'aplicació					
influence	436000000	influence cancer	34200	0.0000784	subdomini no seleccionat
progressive	143000000	progressive cancer	23500	0.000164	subdomini no seleccionat
cotton	161000000	cotton cancer	114000	0.000708	subdomini no seleccionat
testis	7000000	testis cancer	77800	0.0111	subdomini seleccionat
eduthe	1950	eduthe cancer	1	0.000512	subdomini no seleccionat
gu	110000000	gu cancer	648	0.00000589	subdomini no seleccionat
kidney	65400000	kidney cancer	3140000	0.0480	subdomini seleccionat
gastrointestinal	32300000	gastrointestinal cancer	207000	0.0064	subdomini seleccionat
hepatocellular	3690000	hepatocellular cancer	104000	0.0281	subdomini seleccionat
Pàgina 24 (480 articles analitzats) / LearningRate = 57.46 > 40 --> continua l'aplicació					
volunteer	504000000	volunteer cancer	418	0.000000829	subdomini no seleccionat
translational	28300000	translational cancer	61000	0.0021	subdomini seleccionat
winship	2020000	winship cancer	102000	0.0504	subdomini seleccionat
genitourinary	5930000	genitourinary cancer	74900	0.0126	subdomini seleccionat
randolph	54100000	randolph cancer	26600	0.000491	subdomini no seleccionat
baker	187000000	baker cancer	46900	0.00025	subdomini no seleccionat
Pàgina 25 (500 articles analitzats) / LearningRate = 57.14 > 40 --> continua l'aplicació					
ishikawa	23100000	ishikawa cancer	29	0.00000125	subdomini no seleccionat
Pàgina 26 (520 articles analitzats) / LearningRate = 56.73 > 40 --> continua l'aplicació					
overexpressing	2040000	overexpressing cancer	665	0.000325	subdomini no seleccionat
Pàgina 27 (540 articles analitzats) / LearningRate = 56.33 > 40 --> continua l'aplicació					
nonhematopoietic	85600	nonhematopoietic cancer	32	0.000373	subdomini no seleccionat
kimmel	13700000	kimmel cancer	449000	0.0327	subdomini seleccionat
wilmot	7830000	wilmot cancer	51000	0.0065	subdomini seleccionat
Pàgina 28 (560 articles analitzats) / LearningRate = 56.55 > 40 --> continua l'aplicació					

indolent	2690000	indolent cancer	300	0.000111	subdomini no seleccionat
expose	96300000	expose cancer	101	0.00000104	subdomini no seleccionat
possible	0				subdomini no seleccionat
Pàgina 29 (580 articles analitzats) / LearningRate = 55.40 > 40 --> continua l'aplicació					
georgia	552000000	georgia cancer	68200	0.000123	subdomini no seleccionat
different	0				subdomini no seleccionat
inhibits	22400000	inhibits cancer	40800	0.0018	subdomini seleccionat
vitro	53000000	vitro cancer	66700	0.0012	subdomini seleccionat
potential	1130000000	potential cancer	375000	0.000331	subdomini no seleccionat
military	681000000	military cancer	65300	0.0000958	subdomini no seleccionat
corpus	61200000	corpus cancer	14300	0.000233	subdomini no seleccionat
definite	49900000	definite cancer	270	0.00000541	subdomini no seleccionat
predicts	62500000	predicts cancer	9380	0.00015	subdomini no seleccionat
total	2050000000	total cancer	129000	0.0000629	subdomini no seleccionat
Pàgina 30 (600 articles analitzats) / LearningRate = 53.16 > 40 --> continua l'aplicació					
pltesticular	0				subdomini no seleccionat
Pàgina 31 (620 articles analitzats) / LearningRate = 52.83 > 40 --> continua l'aplicació					
stage	720000000	stage cancer	308000	0.000427	subdomini no seleccionat
term	1260000000	term cancer	154000	0.000122	subdomini no seleccionat
biliary	5370000	biliary cancer	29300	0.0054	subdomini seleccionat
Pàgina 32 (640 articles analitzats) / LearningRate = 52.46 > 40 --> continua l'aplicació					
appendiceal	144000	appendiceal cancer	549	0.0038	subdomini seleccionat
itcolorectal	0				subdomini no seleccionat
nscl	580000	nscl cancer	286	0.000493	subdomini no seleccionat
Pàgina 33 (660 articles analitzats) / LearningRate = 52.12 > 40 --> continua l'aplicació					
aichi	9190000	aichi cancer	69500	0.0075	subdomini seleccionat
synthetic	103000000	synthetic cancer	308	0.00000299	subdomini no seleccionat
Pàgina 34 (680 articles analitzats) / LearningRate = 52.09 > 40 --> continua l'aplicació					
care	1670000000	care cancer	377000	0.000225	subdomini no seleccionat
camargo	4660000	camargo cancer	133	0.0000285	subdomini no seleccionat
carvalho	22000000	carvalho cancer	79	0.00000359	subdomini no seleccionat
Pàgina 35 (700 articles analitzats) / LearningRate = 51.17 > 40 --> continua l'aplicació					
crab	38000000	crab cancer	88000	0.0023	subdomini seleccionat
kentucky	398000000	kentucky cancer	27600	0.0000693	subdomini no seleccionat
markey	7290000	markey cancer	48600	0.0066	subdomini seleccionat
treat	360000000	treat cancer	2020000	0.0056	subdomini seleccionat
central	2030000000	central cancer	203000	0.0001	subdomini no seleccionat
iv	555000000	iv cancer	61100	0.00011	subdomini no seleccionat
local	0				subdomini no seleccionat
adoptive	11100000	adoptive cancer	200	0.000018	subdomini no seleccionat
Pàgina 36 (720 articles analitzats) / LearningRate = 50.56 > 40 --> continua l'aplicació					

tongue	68900000	tongue cancer	113000	0.0016	subdomini seleccionat
sas	118000000	sas cancer	11	0.000000093	subdomini no seleccionat
increase	1190000000	increase cancer	156000	0.000131	subdomini no seleccionat
Pàgina 37 (740 articles analitzats) / LearningRate = 50.27 > 40 --> continua l'aplicació					
shandong	16400000	shandong cancer	115	0.00000701	subdomini no seleccionat
Pàgina 38 (760 articles analitzats) / LearningRate = 50.0 > 40 --> continua l'aplicació					
shanghai	97000000	shanghai cancer	33600	0.000346	subdomini no seleccionat
bank	897000000	bank cancer	516	0.000000575	subdomini no seleccionat
Pàgina 39 (780 articles analitzats) / LearningRate = 49.45 > 40 --> continua l'aplicació					
chemotherapeutic	3340000	chemotherapeutic cancer	288	0.0000862	subdomini no seleccionat
le	0				subdomini no seleccionat
Pàgina 40 (800 articles analitzats) / LearningRate = 48.92 > 40 --> continua l'aplicació					
serial	272000000	serial cancer	36	0.000000132	subdomini no seleccionat
taussig	1280000	taussig cancer	47200	0.0368	subdomini seleccionat
greater	1010000000	greater cancer	18700	0.0000185	subdomini no seleccionat
year	0				subdomini no seleccionat
Pàgina 41 (820 articles analitzats) / LearningRate = 48.42 > 40 --> continua l'aplicació					
prefectural	5430000	prefectural cancer	568	0.000104	subdomini no seleccionat
fam	36300000	fam cancer	13400	0.000369	subdomini no seleccionat
dominant	119000000	dominant cancer	10600	0.000089	subdomini no seleccionat
strong	1060000000	strong cancer	15700	0.0000148	subdomini no seleccionat
familial	22300000	familial cancer	256000	0.0114	subdomini seleccionat
Pàgina 42 (840 articles analitzats) / LearningRate = 47.69 > 40 --> continua l'aplicació					
gallbladder	6160000	gallbladder cancer	217000	0.0352	subdomini seleccionat
state	0				subdomini no seleccionat
papillary	2510000	papillary cancer	46800	0.0186	subdomini seleccionat
Pàgina 43 (860 articles analitzats) / LearningRate = 47.97 > 40 --> continua l'aplicació					
lambret	77600	lambret cancer	46	0.000592	subdomini no seleccionat
gung	11600000	gung cancer	21	0.00000181	subdomini no seleccionat
rmt	4360000	rmt cancer	4	0.000000917	subdomini no seleccionat
glottic	172000	glottic cancer	23600	0.1372	subdomini seleccionat
inflammatory	36000000	inflammatory cancer	536	0.0000148	subdomini no seleccionat
Pàgina 44 (880 articles analitzats) / LearningRate = 47.29 > 40 --> continua l'aplicació					
adolescent	61000000	adolescent cancer	63000	0.0010	subdomini seleccionat
columbia	1090000000	columbia cancer	169000	0.000155	subdomini no seleccionat
lymphoid	3900000	lymphoid cancer	12300	0.0031	subdomini seleccionat
karmanos	231000	karmanos cancer	182000	0.7878	subdomini seleccionat
defining	297000000	defining cancer	750	0.00000252	subdomini no seleccionat
overall	837000000	overall cancer	128000	0.000152	subdomini no seleccionat
patient	425000000	patient cancer	70700	0.000166	subdomini no seleccionat
Pàgina 45 (900 articles analitzats) / LearningRate = 47.14 > 40 --> continua l'aplicació					
tasmanian	9990000	tasmanian cancer	242	0.0000242	subdomini no seleccionat
Pàgina 46 (920 articles analitzats) / LearningRate = 46.91 > 40 --> continua l'aplicació					

incident	264000000	incident cancer	28100	0.000106	subdomini no seleccionat
tz	40300000	tz cancer	104	0.00000258	subdomini no seleccionat
pz	28600000	pz cancer	48	0.00000167	subdomini no seleccionat
general	0				subdomini no seleccionat
stomach	54400000	stomach cancer	2690000	0.0494	subdomini seleccionat
Pàgina 47 (940 articles analitzats) / LearningRate = 46.29 > 40 --> continua l'aplicació					
vivo	94100000	vivo cancer	109000	0.0011	subdomini seleccionat
connecticut	330000000	connecticut cancer	945	0.00000286	subdomini no seleccionat
Pàgina 48 (960 articles analitzats) / LearningRate = 46.33 > 40 --> continua l'aplicació					
tanta	20100000	tanta cancer	27	0.00000134	subdomini no seleccionat
Pàgina 49 (980 articles analitzats) / LearningRate = 46.11 > 40 --> continua l'aplicació					
seattle	532000000	seattle cancer	140000	0.000263	subdomini no seleccionat
Pàgina 50 (1000 articles analitzats) / LearningRate = 45.90 > 40 --> continua l'aplicació					
hypopharyngeal	255000	hypopharyngeal cancer	77700	0.3047	subdomini seleccionat
maccallum	2260000	maccallum cancer	146000	0.0646	subdomini seleccionat
cellular	265000000	cellular cancer	9310	0.0000351	subdomini no seleccionat
brown	521000000	brown cancer	55900	0.000107	subdomini no seleccionat
natl	44100000	natl cancer	2720000	0.0616	subdomini seleccionat
Pàgina 51 (1020 articles analitzats) / LearningRate = 46.22 > 40 --> continua l'aplicació					
monitoring	646000000	monitoring cancer	23500	0.0000363	subdomini no seleccionat
affect	636000000	affect cancer	114000	0.000179	subdomini no seleccionat
Pàgina 52 (1040 articles analitzats) / LearningRate = 45.81 > 40 --> continua l'aplicació					
acute	111000000	acute cancer	9740	0.0000877	subdomini no seleccionat
rectosigmoid	233000	rectosigmoid cancer	9300	0.0399	subdomini seleccionat
understanding	939000000	understanding cancer	224000	0.000238	subdomini no seleccionat
eradicate	24300000	eradicate cancer	35400	0.0014	subdomini seleccionat
Pàgina 53 (1060 articles analitzats) / LearningRate = 45.88 > 40 --> continua l'aplicació					
good	0				subdomini no seleccionat
Pàgina 54 (1080 articles analitzats) / LearningRate = 45.68 > 40 --> continua l'aplicació					
mol	55700000	mol cancer	619000	0.0111	subdomini seleccionat
multiple	1650000000	multiple cancer	59000	0.0000357	subdomini no seleccionat
Pàgina 55 (1100 articles analitzats) / LearningRate = 45.72 > 40 --> continua l'aplicació					
attractive	232000000	attractive cancer	106	0.000000456	subdomini no seleccionat
responsive	122000000	responsive cancer	949	0.00000777	subdomini no seleccionat
Pàgina 56 (1120 articles analitzats) / LearningRate = 45.33 > 40 --> continua l'aplicació					
teenage	93000000	teenage cancer	118000	0.0012	subdomini seleccionat
lobe	14600000	lobe cancer	136	0.00000931	subdomini no seleccionat
Pàgina 57 (1140 articles analitzats) / LearningRate = 45.37 > 40 --> continua l'aplicació					

peritoneal	5410000	peritoneal cancer	79400	0.0146	subdomini seleccionat
mdr	29700000	mdr cancer	531	0.0000178	subdomini no seleccionat
art	0				subdomini no seleccionat
safe	733000000	safe cancer	7310	0.00000997	subdomini no seleccionat
milan	82200000	milan cancer	762	0.00000927	subdomini no seleccionat
Pàgina 58 (1160 articles analitzats) / LearningRate = 44.85 > 40 --> continua l'aplicació					
alters	20300000	alters cancer	79	0.00000389	subdomini no seleccionat
pediatric	88800000	pediatric cancer	945000	0.0106	subdomini seleccionat
stanford	381000000	stanford cancer	57100	0.000149	subdomini no seleccionat
Pàgina 59 (1180 articles analitzats) / LearningRate = 44.71 > 40 --> continua l'aplicació					
hospital	587000000	hospital cancer	456000	0.000776	subdomini no seleccionat
combinatorial	25100000	combinatorial cancer	38	0.00000151	subdomini no seleccionat
harvard	457000000	harvard cancer	109000	0.000238	subdomini no seleccionat
Pàgina 60 (1200 articles analitzats) / LearningRate = 44.17 > 40 --> continua l'aplicació					
recent	0				subdomini no seleccionat
serum	37100000	serum cancer	806	0.0000217	subdomini no seleccionat
memorial	484000000	memorial cancer	89000	0.000183	subdomini no seleccionat
Pàgina 61 (1220 articles analitzats) / LearningRate = 43.65 > 40 --> continua l'aplicació					
genitourin	126000	genitourin cancer	133	0.0010	subdomini seleccionat
pearl	103000000	pearl cancer	23400	0.000227	subdomini no seleccionat
refractory	8470000	refractory cancer	40100	0.0047	subdomini seleccionat
Pàgina 62 (1240 articles analitzats) / LearningRate = 43.92 > 40 --> continua l'aplicació					
itbreast	30	0			subdomini no seleccionat
Pàgina 63 (1260 articles analitzats) / LearningRate = 43.75 > 40 --> continua l'aplicació					
Pàgina 64 (1280 articles analitzats) / LearningRate = 43.75 > 40 --> continua l'aplicació					
render	185000000	render cancer	417	0.00000225	subdomini no seleccionat
individual	1420000000	individual cancer	147000	0.000103	subdomini no seleccionat
validate	134000000	validate cancer	182	0.00000135	subdomini no seleccionat
sustains	8810000	sustains cancer	21	0.00000238	subdomini no seleccionat
various	1310000000	various cancer	133000	0.000101	subdomini no seleccionat
relat	6770000	relat cancer	110000	0.0162	subdomini seleccionat
Pàgina 65 (1300 articles analitzats) / LearningRate = 43.12 > 40 --> continua l'aplicació					
Pàgina 66 (1320 articles analitzats) / LearningRate = 43.12 > 40 --> continua l'aplicació					
semin	4260000	semin cancer	165000	0.0387	subdomini seleccionat
Pàgina 67 (1340 articles analitzats) / LearningRate = 43.34 > 40 --> continua l'aplicació					
anaemic	842000	anaemic cancer	437	0.000519	subdomini no seleccionat
european	1240000000	european cancer	164000	0.000132	subdomini no seleccionat
second	0				subdomini no seleccionat
juravinski	30700	juravinski cancer	25700	0.8371	subdomini seleccionat
index	0				subdomini no seleccionat
Pàgina 68 (1360 articles analitzats) / LearningRate = 42.91 > 40 --> continua l'aplicació					
metaxa	701000	metaxa cancer	269	0.000383	subdomini no seleccionat
Pàgina 69 (1380 articles analitzats) / LearningRate = 42.75 > 40 --> continua l'aplicació					

gujarat	19700000	gujarat cancer	12000	0.000609	subdomini no seleccionat
grade	479000000	grade cancer	28900	0.0000603	subdomini no seleccionat
Pàgina 70 (1400 articles analitzats) / LearningRate = 42.43 > 40 --> continua l'aplicació					
rex	81400000	rex cancer	413	0.00000507	subdomini no seleccionat
pharmacologic	5780000	pharmacologic cancer	33	0.0000057	subdomini no seleccionat
pancreas	15400000	pancreas cancer	204000	0.0132	subdomini seleccionat
Pàgina 71 (1420 articles analitzats) / LearningRate = 42.33 > 40 --> continua l'aplicació					
brain	454000000	brain cancer	2570000	0.0056	subdomini seleccionat
provincial	76700000	provincial cancer	30700	0.0004	subdomini no seleccionat
Pàgina 72 (1440 articles analitzats) / LearningRate = 42.39 > 40 --> continua l'aplicació					
saskatchewan	101000000	saskatchewan cancer	25800	0.000255	subdomini no seleccionat
guide	0				subdomini no seleccionat
rev	173000000	rev cancer	272000	0.0015	subdomini seleccionat
Pàgina 73 (1460 articles analitzats) / LearningRate = 42.29 > 40 --> continua l'aplicació					
periampullary	82200	periampullary cancer	16100	0.1958	subdomini seleccionat
dead	637000000	dead cancer	11600	0.0000182	subdomini no seleccionat
putative	24800000	putative cancer	855	0.0000344	subdomini no seleccionat
boosts	44500000	boosts cancer	35000	0.000786	subdomini no seleccionat
certain	1940000000	certain cancer	163000	0.000084	subdomini no seleccionat
Pàgina 74 (1480 articles analitzats) / LearningRate = 41.90 > 40 --> continua l'aplicació					
sensitivity	106000000	sensitivity cancer	557	0.00000525	subdomini no seleccionat
bivalve	2120000	bivalve cancer	1	0.000000471	subdomini no seleccionat
Pàgina 75 (1500 articles analitzats) / LearningRate = 41.60 > 40 --> continua l'aplicació					
decreases	67400000	decreases cancer	780	0.0000115	subdomini no seleccionat
heteromorphologic	13	heteromorphologic cancer	4	0.3076	subdomini seleccionat
uk	2100000000	uk cancer	133000	0.0000633	subdomini no seleccionat
Pàgina 76 (1520 articles analitzats) / LearningRate = 41.52 > 40 --> continua l'aplicació					
grovarian	59	grovarian cancer	0		subdomini no seleccionat
numerous	470000000	numerous cancer	19900	0.0000423	subdomini no seleccionat
Pàgina 77 (1540 articles analitzats) / LearningRate = 41.23 > 40 --> continua l'aplicació					
hepatic	9200000	hepatic cancer	69300	0.0075	subdomini seleccionat
zhejiang	25900000	zhejiang cancer	263	0.0000101	subdomini no seleccionat
female	479000000	female cancer	123000	0.000256	subdomini no seleccionat
Pàgina 78 (1560 articles analitzats) / LearningRate = 41.15 > 40 --> continua l'aplicació					
lower	867000000	lower cancer	109000	0.000125	subdomini no seleccionat
mediating	10500000	mediating cancer	345	0.0000328	subdomini no seleccionat
Pàgina 79 (1580 articles analitzats) / LearningRate = 40.87 > 40 --> continua l'aplicació					
nodal	6520000	nodal cancer	75	0.0000115	subdomini no seleccionat
gland	19500000	gland cancer	461000	0.0236	subdomini seleccionat
Pàgina 80 (1600 articles analitzats) / LearningRate = 40.93 > 40 --> continua l'aplicació					

beijing	172000000	beijing cancer	619	0.00000359	subdomini no seleccionat
positive	635000000	positive cancer	55900	0.000088	subdomini no seleccionat
nasopharyngeal	1970000	nasopharyngeal cancer	458000	0.2324	subdomini seleccionat
diverse	603000000	diverse cancer	903	0.00000149	subdomini no seleccionat
Pàgina 81 (1620 articles analitzats) / LearningRate = 40.72 > 40 --> continua l'aplicació					
lscottish	695	lscottish cancer	0		subdomini no seleccionat
du	0				subdomini no seleccionat
subsequent	355000000	subsequent cancer	38000	0.000107	subdomini no seleccionat
cardia	1460000	cardia cancer	22100	0.0151	subdomini seleccionat
Pàgina 82 (1640 articles analitzats) / LearningRate = 40.52 > 40 --> continua l'aplicació					
vumc	1350000	vumc cancer	187	0.000138	subdomini no seleccionat
mammographic	666000	mammographic cancer	19	0.0000285	subdomini no seleccionat
knockout	13700000	knockout cancer	22	0.0000016	subdomini no seleccionat
Pàgina 83 (1660 articles analitzats) / LearningRate = 40.12 > 40 --> continua l'aplicació					
endometrioid	158000	endometrioid cancer	369	0.0023	subdomini seleccionat
gastrointest	1200000	gastrointest cancer	11900	0.0099	subdomini seleccionat
novo	103000000	novo cancer	571	0.00000554	subdomini no seleccionat
germany	1150000000	germany cancer	10200	0.00000886	subdomini no seleccionat
Pàgina 84 (1680 articles analitzats) / LearningRate = 40.25 > 40 --> continua l'aplicació					
noncutaneous	23500	noncutaneous cancer	459	0.0195	subdomini seleccionat
cause	1020000000	cause cancer	5100000	0.0050	subdomini seleccionat
prefer	390000000	prefer cancer	171	0.000000438	subdomini no seleccionat
Pàgina 85 (1700 articles analitzats) / LearningRate = 40.50 > 40 --> continua l'aplicació					
elevated	72100000	elevated cancer	41400	0.000574	subdomini no seleccionat
Pàgina 86 (1720 articles analitzats) / LearningRate = 40.37 > 40 --> continua l'aplicació					
Pàgina 87 (1740 articles analitzats) / LearningRate = 40.37 > 40 --> continua l'aplicació					
trigger	138000000	trigger cancer	27500	0.000199	subdomini no seleccionat
squamous	5160000	squamous cancer	51600	0.01	subdomini seleccionat
microinvasive	78700	microinvasive cancer	759	0.0096	subdomini seleccionat
attack	385000000	attack cancer	148000	0.000384	subdomini no seleccionat
Pàgina 88 (1760 articles analitzats) / LearningRate = 40.49 > 40 --> continua l'aplicació					
higher	1610000000	higher cancer	101000	0.0000627	subdomini no seleccionat
Pàgina 89 (1780 articles analitzats) / LearningRate = 40.37 > 40 --> continua l'aplicació					
terminal	267000000	terminal cancer	1110000	0.0041	subdomini seleccionat
intractable	12700000	intractable cancer	12500	0.000984	subdomini no seleccionat
Pàgina 90 (1800 articles analitzats) / LearningRate = 40.43 > 40 --> continua l'aplicació					
radical	151000000	radical cancer	773	0.00000511	subdomini no seleccionat
theagenion	657	theagenion cancer	0		subdomini no seleccionat
neuroendocrine	3270000	neuroendocrine cancer	15300	0.0046	subdomini seleccionat
exocrine	1230000	exocrine cancer	1120	0.00091	subdomini no seleccionat
Pàgina 91 (1820 articles analitzats) / LearningRate = 40.24 > 40 --> continua l'aplicació					

expressing	105000000	expressing cancer	16900	0.00016	subdomini no seleccionat
fatal	130000000	fatal cancer	123000	0.000946	subdomini no seleccionat
predisposes	1750000	predisposes cancer	42	0.000024	subdomini no seleccionat
zealand	686000000	zealand cancer	22600	0.0000329	subdomini no seleccionat
Pàgina 92 (1840 articles analitzats) / LearningRate = 39.75 < 40 --> finalitza l'aplicació					

Bé, abans de comentar res, ja es pot observar la gran quantitat d'*abstracts* d'articles que s'han arribat a tractar. Més concretament, s'han examinat 92 planes web retornades per la llibreria, que corresponen ni més ni menys que a 1840 articles analitzats.

Aquests nombres ens poden semblar molt elevats, però fixem-nos que l'aplicació finalitza perquè el rati d'aprenentatge acaba sent menor que l'establert en un principi pel paràmetre de la cerca. Què significa això? Dons vol dir que la llibreria digital *PubMed*, encara disposa de molts més articles disponibles que tracten sobre el concepte "*cancer*". Per tant, en el cas que el sistema disposés d'una opció de cerca més exhaustiva encara, es podrien arribar a trobar moltes més especialitzacions.

A continuació, es mostra una llista de tots els subdominis seleccionats pel sistema i que passaran a formar part de la estructuració final de la informació. De tots aquests termes seleccionats, s'han remarcat amb color blau i s'han traduït al català, els que es corresponen amb tipus de càncer determinats. S'han remarcat amb color verd els que fan referència a línies o centres d'investigació o lluita contra el càncer. Amb color taronja tenim especialitzacions que tenen alguna relació amb "*cancer*". I per últim, amb color negre, hi ha els termes que es podria dir que no tenen cap relació amb "*cancer*".

- | | |
|---|--|
| <ul style="list-style-type: none"> - adjuvant → teràpia càncer de pit - adolescent - Aichi - anal cancer → càncer anal - Anderson - appendiceal → càncer d'apèndix - Bc - biliary → càncer biliar - bladder → càncer de veixiga - blood → càncer de sang - Bmc - bone → càncer d'os - bowel → càncer d'intestí - brain → càncer de cervell - breast → càncer de pit - cardia → càncer de cor - cause - cell → càncer de cèl·lula - cervical → càncer cervical - cervix → càncer de cèrvix - chase - childhood | <ul style="list-style-type: none"> - clinical - colon → càncer de colon - colorectal → càncer colo - rectal - comprehensive - crab - destroy - duct → càncer de conducte - endometrial - endometrioid - eradicate - esophageal → càncer d'esòfag - familial - Farber - gallbladder → càncer vesicular - gastric → càncer gàstric - gastrointest - gastrointestinal → càncer gastrointestinal - genitourin - genitourinary - gland → càncer de glàndula |
|---|--|

- glottic → càncer de glotis
- gynaecological → càncer ginecològic
- gynecologic → càncer ginecològic
- hepatic → càncer hepàtic
- hepatocellular → càncer hepato – cel·lular
- heteromorphologic → càncer hetero – morfologic
- Hoed
- Huntsman
- Hutchinson
- hypopharyngeal → càncer d'hypo – faringe
- incurable
- Ingram
- inhibits
- invasive
- juravinski → centre ajuda càncer
- Karmanos
- Kettering
- kidney → càncer de ronyó
- kills
- Kimmel
- laryngeal → càncer de laringe
- larynx → càncer de laringe
- liver → càncer de fetge
- lung → càncer de pulmó
- Lurie
- lymphoid → càncer limfàtic
- Maccallum
- mammary → càncer de mama
- Markey
- metastatic
- microinvasive
- Moffitt
- mol
- nasopharyngeal → càncer de nas – faringe
- nasopharynx → càncer de nas – faringe
- Natl
- neck → càncer de coll
- neuroendocrine → càncer neuro – endocrí
- Niigata
- noncutaneous
- nonlung
- nonprostate

- oesophageal → càncer d'esòfag
- oral → càncer oral
- oropharyngeal
- ovarian → càncer ovàric
- pancreas → càncer de pàncreas
- pancreatic → càncer pancreàtic
- papillary
- Park
- Pediatric
- penile → càncer de penis
- periampullary
- peritoneal → càncer peritoneal
- pharyngeal → càncer de faringe
- prostate → càncer de pròstata
- rectal → càncer rectal
- rectosigmoid
- rectum → càncer de recte
- recurrent
- refractory
- Regional
- relat
- renal → càncer renal
- residual
- Rev
- Schiffler
- semin
- Shikoku
- skin → càncer de pell
- squamous
- stomach → càncer d'estómac
- Taussig
- Teenage
- terminal
- testicular → càncer testicular
- testis → càncer de testicles
- thyroid → càncer tiroïdal
- tongue → càncer de llengua
- tract → urinary tract cancer
- Translational
- treat
- Tulane
- tumor → càncer de tumor
- ureteral → càncer d'urèter
- uterine → càncer uterí
- vitro
- vivo
- Wilmot
- Winship

En la llista anterior hi ha un total de 131 termes que han estat seleccionats per a formar part de la estructuració de la llibreria digital en forma de taxonomia. De tots aquests, aproximadament 71, farien referència a tipus concrets de càncer (color blau). Uns 30 més, s'han considerat com a centres, publicacions o línies de recerca que realitzen investigacions en la lluita contra el càncer (color verd). Uns 19 tenen alguna relació directa amb el terme “*cancer*” (color taronja). I per últim, solament destacar, que de totes aquestes especialitzacions, solament 11 es podrien considerar que no tenen cap o molt poca relació amb “*cancer*”.

Aquesta prova ha estat molt exhaustiva, com s'ha dit anteriorment, i el fet d'utilitzar la llibreria digital adient (*PubMed*) per al domini seleccionat, ha fet que el nombre d'especialitzacions retornades sigui molt gran, i que gairebé totes es puguin donar per bones, sinó en una categoria semàntica, en l'altra.

Per tant, en aquesta prova, la estructuració de la informació retornada pel sistema, es podria considerar directament com una taxonomia molt completa del domini “*cancer*”.

4.9- Prova 9

Llibreria digital:	<i>NASA Astrophysics Data System</i>
Terme clau:	<i>"sensor"</i>
Amplada cerca:	<i>Senzilla (80)</i>
Precisió cerca:	<i>Elevada (0.001)</i>
Temps aproximat	2 minuts

Candidat a subdomini	Nombre resultats	subdomini + terme clau	Nombre resultats	Relació	Decisió
pressure	481000000	pressure sensor	2940000	0.0061	subdomini seleccionat
silicon	188000000	silicon sensor	120000	0.000638	subdomini no seleccionat
fibre	57800000	fibre sensor	26200	0.000453	subdomini no seleccionat
wireless	955000000	wireless sensor	3410000	0.0035	subdomini seleccionat
motion	473000000	motion sensor	3580000	0.0075	subdomini seleccionat
current	0				subdomini no seleccionat
displacement	37500000	displacement sensor	182000	0.0048	subdomini seleccionat
gmr	3650000	gmr sensor	14600	0.0040	subdomini seleccionat
mechanical	317000000	mechanical sensor	30800	0.0000971	subdomini no seleccionat
power	0				subdomini no seleccionat
fluidic	758000	fluidic sensor	202	0.000266	subdomini no seleccionat
hoc	102000000	hoc sensor	149000	0.0014	subdomini seleccionat
multiplatform	3520000	multiplatform sensor	55	0.0000156	subdomini no seleccionat
platform	988000000	platform sensor	13700	0.0000138	subdomini no seleccionat
netted	8530000	netted sensor	10300	0.0012	subdomini seleccionat
smart	609000000	smart sensor	251000	0.000412	subdomini no seleccionat
mems	21900000	mems sensor	42600	0.0019	subdomini seleccionat
predictive	45500000	predictive sensor	330	0.00000725	subdomini no seleccionat
capture	390000000	capture sensor	812	0.00000208	subdomini no seleccionat
economical	58800000	economical sensor	546	0.00000928	subdomini no seleccionat
Pàgina 1 (30 articles analitzats) / LearningRate = 40 < 80 --> finalitza l'aplicació					

En aquesta prova s'ha escollit la llibreria "*NASA Astrophysics Data System*", ja utilitzada anteriorment per a la generació de la taxonomia a partir del terme "*cancer*". Ara però, s'ha realitzat la cerca a partir del terme "*sensor*".

4.10- Prova 10

Llibreria digital:	<i>Association for Computing Machinery. The Guide</i>
Terme clau:	<i>"programming"</i>
Amplada cerca:	<i>Senzilla (80)</i>
Precisió cerca:	<i>Elevada (0.001)</i>
Temps aproximat	2 minuts

Candidat a subdomini	Nombre resultats	subdomini + terme clau	Nombre resultats	Relació	Decisió
extreme	450000000	extreme programming	3540000	0.0078	subdomini seleccionat
pair	268000000	pair programming	383000	0.0014	subdomini seleccionat
teaching	1010000000	teaching programming	151000	0.000149	subdomini no seleccionat
natural	1240000000	natural programming	36200	0.0000291	subdomini no seleccionat
interface	971000000	interface programming	181000	0.000186	subdomini no seleccionat
bilevel	418000	bilevel programming	36400	0.0870	subdomini seleccionat
nonlinear	56100000	nonlinear programming	762000	0.0135	subdomini seleccionat
opengl	23300000	opengl programming	250000	0.0107	subdomini seleccionat
application	0				subdomini no seleccionat
textual	67400000	textual programming	24200	0.000359	subdomini no seleccionat
implicit	64100000	implicit programming	844	0.0000131	subdomini no seleccionat
various	1390000000	various programming	385000	0.000276	subdomini no seleccionat
logic	300000000	logic programming	5560000	0.0185	subdomini seleccionat
procedural	60800000	procedural programming	314000	0.0051	subdomini seleccionat
new	0				subdomini no seleccionat
parallel	287000000	parallel programming	2120000	0.0073	subdomini seleccionat
modern	911000000	modern programming	306000	0.000335	subdomini no seleccionat
generic	254000000	generic programming	468000	0.0018	subdomini seleccionat
Pàgina 1 (20 articles analitzats) / LearningRate = 50 < 80 --> finalitza l'aplicació					

En aquesta prova s'ha escollit la llibreria “*Association for Computing Machinery. The Guide*”, ja utilitzada anteriorment per a la generar la taxonomia a partir del terme “*cancer*”. Ara però, s'ha realitzat la cerca a partir del terme “*programming*”.

4.11- Prova 11

Llibreria digital:	<i>IEEE. Computer Society</i>
Terme clau:	<i>"language"</i>
Amplada cerca:	Molt complexa (40)
Precisió cerca:	<i>Elevada</i> (0.001)
Temps aproximat	2 minuts

Candidat a subdomini	Nombre resultats	subdomini + terme clau	Nombre resultats	Relació	Decisió
assembly	336000000	assembly language	4220000	0.0125	subdomini seleccionat
machine	723000000	machine language	1150000	0.0015	subdomini seleccionat
programming	998000000	programming language	79000000	0.0791	subdomini seleccionat
user	0				subdomini no seleccionat
input	682000000	input language	361000	0.000529	subdomini no seleccionat
algebraic	30200000	algebraic language	96700	0.0032	subdomini seleccionat
target	821000000	target language	2820000	0.0034	subdomini seleccionat
english	0				subdomini no seleccionat
study	1980000000	study language	189000	0.0000954	subdomini no seleccionat
statistic	40800000	statistic language	131	0.00000321	subdomini no seleccionat
Pàgina 1 (20 articles analitzats) / LearningRate = 50 > 40 --> continua l'aplicació					
proposes	82500000	proposes language	868	0.0000105	subdomini no seleccionat
sign	0				subdomini no seleccionat
visual	789000000	visual language	2220000	0.0028	subdomini seleccionat
Pàgina 2 (40 articles analitzats) / LearningRate = 46.1538 > 40 --> continua l'aplicació					
parameterized	8580000	parameterized language	138	0.000016	subdomini no seleccionat
natural	1240000000	natural language	54600000	0.0440	subdomini seleccionat
japanese	882000000	japanese language	13200000	0.0149	subdomini seleccionat
ubiquitous	68300000	ubiquitous language	9900	0.000144	subdomini no seleccionat
Pàgina 3 (60 articles analitzats) / LearningRate = 47.0588 > 40 --> continua l'aplicació					
apply	1940000000	apply language	43100	0.0000222	subdomini no seleccionat
archc	18500	archc language	33	0.0017	subdomini seleccionat
plants	267000000	plants language	216	0.000000808	subdomini no seleccionat
proper	366000000	proper language	197000	0.000538	subdomini no seleccionat
transformation	256000000	transformation language	205000	0.0008	subdomini no seleccionat
test	1750000000	test language	134000	0.0000765	subdomini no seleccionat
Pàgina 4 (80 articles analitzats) / LearningRate = 39.1304 < 40 --> finalitza l'aplicació					

En aquesta prova s'ha escollit la llibreria "*IEEE. Computer Society*", ja utilitzada anteriorment per a la generació de la taxonomia a partir del terme "*cancer*". Ara però, s'ha realitzat la cerca a partir del terme "*language*".

5- CONCLUSIONS I TREBALL FUTUR

Aquest és l'últim capítol de la documentació i en ell, com el seu nom indica, es fa referència a les conclusions que es poden extreure del projecte.

El sistema dissenyat i implementat en aquest projecte, s'encarrega d'estructurar la informació o recursos d'una llibreria digital a partir d'un domini introduït per l'usuari.

L'estructura d'informació resultant és mostra en forma de taxonomia, tot i que com hem dit en la introducció, no es pot considerar una taxonomia del domini cercat.

El sistema, a partir d'una llibreria digital d'articles científics escollida per l'usuari, s'encarregarà de buscar i seleccionar candidats a subdominis que especialitzin el terme clau introduït per aquest.

Aquests possibles candidats hauran de passar una sèrie de passos o filtres abans de ser seleccionats com a bons.

Ja per últim, els subdominis que han estat acceptats pel sistema, passaran a formar part de l'estructuració final de la informació, que el sistema retornarà a l'usuari a partir de la interfície gràfica.

Al llarg d'aquesta documentació hem anat comprovant com es complien els objectius que es tenien especificats inicialment en l'apartat segon de la introducció.

S'ha vist com extreure de les planes *web* de les llibreries, els possibles candidats a subdominis, per a seleccionar més tard, els conceptes que passaran a formar part de la estructuració de la informació.

També s'ha aconseguit que l'aplicació funcioni correctament sobre més d'una llibreria digital diferent, sempre i quan aquesta compleixi certes condicions ja mencionades anteriorment en el punt 3.3.

Finalment, s'ha construït una interfície *web* que embolcalla les quatre llibreries digitals que utilitza el sistema, de forma que l'usuari pot treballar fàcilment i de forma transparent, amb cadascuna d'elles.

Com dèiem, l'objectiu principal del sistema, és l'estructuració de la informació d'una llibreria digital en forma de taxonomia. Aquesta informació pot ser utilitzada posteriorment per l'usuari per a veure ràpidament i de forma clara els subdominis relacionats directament amb el terme que ha introduït, de forma que pot ajudar-lo d'una forma significativa, en la cerca de la informació que necessita.

L'execució del sistema amb una amplada de la cerca senzilla i una precisió de la cerca elevada ens retornarà una estructuració dels recursos poc completa, però amb l'avantatge de retornar-la amb un temps relativament baix. Aquesta opció pot interessar a un usuari que pretén aconseguir una resposta ràpida i superficial de les especialitzacions existents d'un domini determinat.

Per a un usuari que pretén aconseguir una estructuració molt completa, és a dir, amb moltes especialitzacions del domini introduït, i que disposa de temps suficient, li interessarà una execució amb una amplada de la cerca molt complexa i una precisió de la cerca elevada o baixa, depenent aquesta de la qualitat dels subdominis que desitja aconseguir.

En tots els projectes i grans aplicacions, un cop es donen per finalitzats, sempre es poden retocar alguns aspectes o canviar-ne algunes característiques. Aquestes tasques fan referència a les possibles actuacions que es poden dur a terme per a millorar la present aplicació i s'identifiquen com a treball futur.

Com hem dit, l'usuari pot treballar de forma transparent, i a través de la interfície gràfica, amb quatre llibreries digitals diferents. Doncs bé, una possible millora que es podria dur a terme seria ampliar el llistat de repositoris d'articles disponibles per al sistema.

Un altre aspecte important a millorar, podria ser la implementació d'una *cache*, que s'utilitzaria a l'hora de realitzar les cerques. Aquesta, hauria d'emmagatzemar els resultats de les cerques que s'anessin duent a terme en el sistema, de forma que quan un usuari volgués realitzar una cerca determinada, es poguessin utilitzar els resultats d'una cerca anterior amb els mateixos paràmetres que l'actual. D'aquesta forma s'aconseguirien els resultats ràpidament. Un aspecte a tenir en compte sobre les cerques emmagatzemades, és que aquestes s'haurien d'eliminar automàticament al cap d'un cert temps. Així s'evitaria retornar resultats poc actualitzats.

Aquesta aplicació genera una estructuració de la informació, existent en una llibreria digital o repositori d'articles, a partir de conceptes que especialitzen directament el terme introduït per l'usuari. Si posem per cas l'exemple del terme càncer, com a terme clau, aquests conceptes faran referència generalment a tipus concrets de càncer. Es podrien però, crear altres tipus de relacions semàntiques per a estructurar la informació, de forma que a part de tipus de càncer, poguéssim obtenir associacions d'ajuda contra el càncer, fundacions, centres de recerca,...

Quan s'utilitza de forma molt continuada l'*API* de *Google* per a realitzar les cerques, el rendiment d'aquest disminueix, de forma que augmenta el temps de resposta. Per tant, un altre dels possibles aspectes que es podrien millorar, seria utilitzar, a part del *Google*, altres buscadors alternatius per a obtenir els estadístics. D'aquesta forma, multiplexant els buscadors per a realitzar les diferents cerques de conceptes, s'agilitaria el procés d'estructuració dels recursos.

M'agradaria també comentar que durant el curs de la carrera d'ETIG, he adquirit, a partir de les assignatures que he anat cursant, molta base necessària per al disseny i desenvolupament d'aquest projecte. Encara que si n'hagués de nombrar algunes en particular, destacaria: Llenguatges de Programació (LP), Sistemes Oberts (SOB) i Desenvolupament d'Aplicacions Internet i Intranet (DAII), aquesta última de la Enginyeria Informàtica.

I ja per finalitzar el capítol de les conclusions, m'agradaria comentar, que la realització d'aquesta aplicació ha estat molt satisfactòria per a mi. Es pot dir que he millorat satisfactòriament el meu nivell de programació amb Java i JSP. A part, he après a estructurar de forma correcta els apartats que componen la documentació per a un projecte d'aquestes magnituds.

Per últim, agrair l'ajuda que he rebut del professor Robert Rallo, pel que fa a la part de la construcció de la interfície *web* amb *JSP*'s. I com no, destacar sobretot la dedicació i la paciència que ha tingut amb mi el director del projecte, David Sánchez, membre del grup de recerca GRUSMA de la ETSE. Ha estat un plaer tenir-lo com a director.

REFERÈNCIES

Com a recursos utilitzats per a la realització del projecte podem destacar bàsicament dos tipus. Els recursos *software*, i les pàgines *web* referenciades o de les quals hem extret informació. A continuació farem referència a aquests tipus de recursos que hem anat enumerant durant la present documentació.

[1]: *Google*

Info.: Cercador d'informació a la web

URL: <http://www.google.es/>

[2]: *Clusty*

Info.: Cercador taxonòmic d'informació a la web

URL: <http://clusty.com/>

[3]: *Vivisimo*

Info.: Cercador taxonòmic d'informació a la web

URL: <http://vivisimo.com/>

[4]:

Títol: Sistema Multi-Agente para la construcción automática de ontologías

Autor: Laura Prieto Rebollo

Tipus: projecte fi de carrera

[5]:

Títol: Scalable knowledge acquisition from the Web

Autor: David Sánchez (director present PFC)

Tipus: projecte de tesi

[6]: *Sun Microsystems*

URL: <http://es.sun.com/>

[7]: HTMLParser

Info.: eina utilitzada per a extreure el text de la pàgina *web* obtinguda.

Llibreria: htmlparser.jar

URL: <http://htmlparser.sourceforge.net/>

[8]: Sourceforge

URL: <http://sourceforge.net/>

[9]: Snowball

Info.: eina utilitzada per a comprovar si dos paraules pertanyen a la mateixa família, és a dir, tenen la mateixa arrel.

Llibreria: snowball-1.0.jar

URL: <http://snowball.tartarus.org/>

- [10]: OpenNLP
Info.: eina utilitzada per a classificar sintàcticament les paraules, és a dir, determinar si són verbs, noms, articles, números, ...
Llibreria: opennlp.jar
URL: <http://opennlp.sourceforge.net/>
- [11]: Swing
Info.: llibreria pertanyent a la Java Foundation Classes (JFC) que suporta la construcció d'interfícies gràfiques d'usuari (GUI), i funcionalitats gràfiques per a aplicacions client que han de funcionar sobre plataformes tals com Microsoft Windows, Linux i MAC OSX.
URL: <http://java.sun.com/products/jfc/>
- [12]: *Eclipse Platform v.3.0.1*
Info.: entorn de programació JAVA utilitzat per al desenvolupament de l'aplicació dissenyada en aquest projecte.
URL: <http://www.eclipse.org/>
- [13]: *Jakarta Tomcat v.4.1.31*
Info.: servidor web utilitzat en aquest projecte per a fer córrer la nostra aplicació.
URL: <http://tomcat.apache.org/>
- [14]: *Porter stemming algorithm* (o *Porter stemmer*), creat per Martin Porter en 1980.
URL: <http://www.tartarus.org/~martin/PorterStemmer/>
- [15]: *Google API*
Info.: eina utilitzada per a realitzar cerques al *Google*, de forma que retorni el número de resultats de pàgines relacionades amb el terme buscat.
Llibreria: googleapi.jar
URL: <http://www.google.com/apis>
- [16]: PubMed
Info.: llibreria digital d'articles científics que està inclosa en el sistema.
URL: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>
- [17]: *Association for Computing Machinery. The Guide*
Info.: llibreria digital d'articles científics que està inclosa en el sistema.
URL: <http://portal.acm.org/guide.cfm?coll=portal&dl=GUIDE&CFID=76520774&CFTOKEN=23776300>
- [18]: *IEEE. Computer Society*
Info.: llibreria digital d'articles científics que està inclosa en el sistema.
URL: <http://www.computer.org/portal/site/ieeecsc/>
- [19]: *NASA Astrophysics Data System*
Info.: llibreria digital d'articles científics que està inclosa en el sistema.
URL: <http://adswww.harvard.edu/index.html>