Universitat Politècnica de Catalunya - BarcelonaTech

Projecte de Fi de Grau

Especialitat Computació

Atles: Mapa de relacions entre àrees de coneixement de la UPC

Autor:
Ferran Maria Toda i Casaban

Directora:
Marta Cuatrecases Capdevila

Ponent:
Albert Renom Vilaró

28 de Setembre de 2020

Resum

Una tasca que pren molt de temps, no sempre és acurada i no s'aprofita al màxim és l'etiquetat de les publicacions dels investigadors. En aquest projecte es desenvoluparà procés que ho automatitzi, i a més a més generi un mapa d'àrees de coneixement de la UPC. Es faran tests amb diferentes eines de NLP, *machine learning* i algorismes de *clústering* per garantir els millors resultats possibles. Els mapes generats seran molt útils per saber temes emergents i facilitar les preses de decisions sobre on invertir i encaminar futures investigacions.

Resumen

Una tarea que requiere de mucho tiempo, no es exacta y no se aprovecha al máximo es el etiqueado de las publicaciones de los investigadores. En este proyecto se desarrollará un proceso para automatizarlo, además de generar un mapa con las áreas de conocimiento de la UPC. Se harán tests con diferentes herramientas de NLP, *machine learning* y algoritmos de *clústering* para garantizar los mejores resultados posibles. Los mapas generados seran muy útiles para conocer los temas emergentes y facilitar la toma de decisiones sobre donde invertir y encaminar futuras investigaciones.

Abstract

A task that requires a lot of time, that is not always accurate and that is not exploited to its maximum is the labelling of the investigator's publications. In this project, a process that will automatize it will be developed, and moreover, it will generate a map of knowledge areas within the UPC. Different tests will be conducted using NLP tools, *machine learning* and *clustering* algorithms in order to ensure the best possible results. The maps generated will be highly useful to show cutting-edge subjects and to facilitate the decision making on where to invest and direct future investigations.

Índex

Índex de figures

Índex de taules

1 Contextualització

1.1 Introducció

Aquest és un projecte de Treball de Fi de Grau del Grau en Enginyeria Informàtica (menció de Computació) de la Facultat d'Informàtica de Barcelona, que ha estat desenvolupat en la modalitat B amb un conveni de Cooperació Educativa a l'inLab FIB[1] dirigit per la Marta Cuatrecasas i amb l'Albert Renom com a ponent.

Es vol crear un procediment per representar en un mapa del coneixement l'activitat d'investigació de la UPC de forma automàtica per a detectar àrees i disciplines emergents i fomentar la interdisciplinaritat i ajudar en la presa de decisions tàctiques i estratègiques d'inversió en la comunitat.

S'utilitzaran tècniques de processament de llenguatge natural per a trobar relacions semàntiques entre publicacions científiques de la comunitat UPC, i així generar la informació necessària per a ser mostrada d'un mode accessible, per exemple, en taules, gràfiques i finalment, en forma de graf.

1.2 Context

La Universitat Politècnica de Catalunya (UPC) abasta àrees de coneixement molt extenses i variades, al voltant de 300 grups de recerca i 3600 investigadors de la universitat realitzen investigacions i publicacions científiques a revistes i portals de divulgació. De mitjana, més de 200 publicacions al mes [2] són llançades al món i és feina dels propis portals d'etiquetar-les i classificar-les manualment segons la temàtica. Això suposa tanta feina que en la majoria dels casos és el propi investigador qui ha d'escriure les paraules clau per encabir-ho dins de categories majors.

Aquest mètode tampoc és fiable, el món canvia contínuament i nous temes es creen i s'expandeixen molt més ràpidament del que es poden actualitzar les categories dels portals on es publiquen. Les revistes generalment fan una classificació *top-down*, és a dir, es classifica de dalt a baix seguint un arbre establert d'etiquetes. Un exemple és SCOPUS [3], una base de dades bibliogràfica per a articles de revistes acadèmiques que fa aquest tipus de classificació en 3 nivells: 4 grans àrees temàtiques es divideixen en 27 disciplines generals i en un tercer nivell de 358 àrees temàtiques menors. Però qualsevol nou tema que no estigui englobat dins d'algun de les àrees anteriors és classificat de manera errònia, fet que no dona espai a la detecció de nous temes emergents.
Una altra possibilitat és el mal etiquetatge per part de l'investigador que, com és d'esperar, aprofitarà per col · locar-lo a categories que criden més l'atenció però que alhora poden estar més allunyades del tema que tracten, tot per donar més visibilitat a les seves publicacions.

És per això que neix la necessitat d'un sistema automàtic per classificar documents (*bottom-up*), relacionar àrees de coneixement i detectar noves sinergies entre aquestes i acabar traient una generació automàtica de mapes en forma de graf d'àrees de recerca de la UPC que mostri aquestes relacions i sinergies.

La visualització d'aquestes dades, si es fa de forma accessible i visual, pot ajudar molt en el procés de presa de decisions i obtenir resultats menys esbiaixats i suportats de manera empírica.

1.3 Definició de conceptes i terminologia

En aquesta secció hi ha una llista dels tòpics que es necessiten per entendre el projecte:

- **NLP**
  El Processament del llenguatge natural, o NLP per les sigles en anglès, és una branca de la informàtica que s'encarrega de tractar computacionalment les llengües naturals, o els llenguatges humans.

- **Graf**
  És una representació abstracta d'un conjunt d'objectes on alguns dels parells d'objectes estan connectats per enllaços. Els objectes interconnectats són anomenats *vèrtex*, i els enllaços que connecten alguns parells de vèrtex s'anomenen *arestes*.

- **Disciplina**
  Una disciplina és una branca del coneixement que s'investiga i s'ensenya a centres d'educació superior. Les disciplines són reconegudes com a tals a través de les publicacions acadèmiques on els investigadors exposen els seus resultats.

- **Interdisciplinaritat**
  Es diu d'aquella àrea de coneixement que pertany a una zona de confluència d'un conjunt de disciplines però no es consolida com a disciplina individual.

- **Clusterització**
  És la classificació d'objectes similars en diferents grups, o més precisament, la partició de les dades en diferents subconjunts (o clústers). Així doncs, les dades de cada subgrup idealment comparteixen un tret comú.

- **Lematització**
  La lematització és el procés d'agrupar totes les formes derivades en un sol item, l'arrel d'aquestes paraules anomenat lema.

- **Corpus**
  S'anomena *corpus* a una col · lecció d'escrits.

- **Doc2Vec**
  És un algorisme d'aprenentatge no supervisat per a generar vectors a partir de documents de text. Aquests vectors poden ser utilitzats per a diferents tasques com trobar la similitud entre parelles de documents.

- **TF-IDF**
  És un terme utilitzat en anàlisi de text, i és la freqüència d'ocurrència del terme en un document concret en relació a la presència que el terme té en el conjunt de documents analitzats.

1.4 Col · lectius implicats

En aquest projecte hi participen un conjunt de persones expertes implicades directament en el desenvolupament del projecte i actors implicats en el seus resultats.

- Persones implicades en el desenvolupament:

- Dr. Lluis Padró: Departament de Ciències de la Computació. Expert tecnològic, especialment en processament de llenguatge natural.
- Dr. Eduard Alarcon: Departament d'Enginyeria Electrònica. Impulsor de la idea inicial.
- Ana Rovira: Servei de Biblioteques. Subministrament del corpus.
- Meritxell Oncins: Servei Suport TIC. Subministrament de la maquinària per a un futur desplegament.

- Actors implicats en el resultat:
  Aquest projecte va dirigit a governança de la UPC: Les connexions que s'observen al graf aporten informació molt útil per a gestionar equips de recerca, detectar departaments que estan investigant sobre una mateixa disciplina i tema o poder identificar àrees de recerca emergents, i així la UPC serà pionera en aquestes àrees i prendrà millors decisions estratègiques tant de finançament com d'investigació.

  Altrament, aquest projecte també serà usat per tot el conjunt de investigadors, professors, docents, i, en general, personal divulgador. Tot l'etiquetatge els vindrà donat amb la col · locació de la seva publicació en el mapa de coneixement. A més a més, serà més fàcil pels equips de recerca trobar altres equips de recerca i establir relacions per col · laborar entre ells.

  L'usuari estàndard podrà, a simple cop de vista, fer-se una idea general de en quins àmbits és especialista la UPC així com el creixement i evolució de les àrees.

  La xarxa de biblioteques UPC es beneficiarà d'aquest projecte ja que s'inclouran noves funcionalitats a FUTUR[4], essent punters en la representació gràfica del coneixement en portals universitaris.

2 Justificació

Com ja s'ha explicat abans, el projecte consisteix en generar de forma automàtica mapes de relacions entre àrees de coneixement de la UPC per acabar identificant sinergies entre diferents àrees d'investigació o temes emergents. Així doncs, s'analitzaran solucions que intenten aconseguir alguna cosa similar.

Una d'aquestes solucions és integrar el projecte a FUTUR [4], el portal de producció científica dels investigadors de la UPC administrat pel mateix servei de biblioteques de la UPC. Aquest portal ofereix informació acadèmica dels investigadors actius de la UPC. A més a més de tota la informació que exposa, també ofereixen uns mapes de coneixement on es visualitzen els àmbits de coneixement UPC i la relació que hi ha entre ells basada en la producció científica dels investigadors de la UPC. Els mapes actuals mostren únicament relacions entre investigadors i la quantitat d'activitat d'investigació en uns àmbits de coneixement preestablerts, és a dir, esbiaixada.

Un altre apropament és el de SCOPUS. Com ja comentat anteriorment, l'etiquetatge que fa sobre els articles és a partir d'uns àmbits o temàtiques ja establertes, i no donen peu a realitzar aquesta identificació de la transdisciplinarietat que és una de les finalitats del projecte.

Detectar aquest conjunt d'àrees i disciplines permet convertir-se en punters en la creació de nous equips d'investigació i recerca en les transdisciplines detectades abans que ningú, així com la detecció d'àrees de recerca emergents i decreixents, bé sigui per falta de publicacions o per l'escissió d'una part d'una disciplina en una de nova, en la què es continua una nova línia d'investigació i es deixa enrere l'antiga.

3 Abast

3.1 Objectius

L'objectiu principal d'aquest projecte és generar de forma automàtica mapes de coneixement de les àrees de recerca de la UPC a partir de les publicacions proporcionades per la biblioteca a través de tècniques de processament de llenguatge natural, així com també, permetre de manera visual detectar àrees de confluència inter, trans i multidisciplinar.

A partir d'aquest objectiu es definiran una sèrie de sub-objectius per a poder assolir l'objectiu general:

- Convertir un corpus de documents en format pdf a un format que sigui fàcil de tractar i treballar com pot ser txt.
- Processament del text: lematization/ tokenization.
- Transformar els documents a vectors*vector* de números per a fer el processament posterior.
- Clusteritzar / classificar.
- Validar i analitzar els resultats
- Generar el mapa de relacions.

3.2 Obstacles i riscos

Durant el desenvolupament del projecte poden sorgir obstacles que provoquin que s'endarrereixi la planificació establerta i impedir que es compleixin parcialment els objectius proposats.

Un dels principals obstacles que es pot presentar és la obtenció del corpus d'articles de la UPC. Com s'ha comentat anteriorment, serà el servei de biblioteca qui proporcioni aquest recull d'articles i pot ser que no disposi de forma d'obtenció al iniciar el projecte. De totes maneres molts dels objectius es poden completar amb un corpus molt reduït generat de manera manual.

Un altre obstacle que pot aparèixer és que la quantitat de documents a tractar sigui massiu i la maquinària de la que es disposa no sigui capaç de fer tot el procediment, que provocarà haver de treballar també amb un corpus reduït.

Aquests obstacles no impedeixen l'assoliment total dels objectius i el projecte es podrà seguir desenvolupant en tot moment.

4 Metodologia i Rigor

4.1 Mètode *Agile*

Per desenvolupar aquest projecte s'ha optat per utilitzar una metodologia de treball àgil o *agile* [5] amb cicles curts de 15 dies (també anomenats *sprints*) basat en *scrum*. Aquesta forma de treballar permet adaptar-se a obstacles o canvis que vagin sorgint per augmentar així les probabilitats d'èxit del projecte. La metodologia *scrum* defineix uns rols per a la bona organització de l'equip. En aquest cas la *Product Owner* serà la Soraya Hidalgo de l'Àrea de Recerca i Transferència / Ctt, l'Albert Renom agafarà el rol de *Scrum Master* i l'equip de desenvolupament serà un equip de l'inLab de 3 persones, on l'autor d'aquest treball hi forma part.

Al final de cada *sprint* es realitza una reunió amb tot l'equip on es demostra la feina feta durant el període, es resoldrà qualsevol dubte que hagi sorgit i així evitar avançar per vies errònies. Aquestes reunions també serveixen per a proposar la feina a fer en el pròxim *sprint*.

A més a més, cada dia hi ha un *daily* (reunió breu diària) amb l'Albert Renom, responsable del projecte a l'InLab, per sincronitzar l'equip, parlar de *stoppers*(coses que podrien frenar el procés de desenvolupament) i com superar-los.

Aprofitant que la metodologia inclou reunions cada 15 dies, la validació del treball es realitza de manera periòdica i constant.

4.2 Eines de seguiment

Durant el desenvolupament del projecte han estat emprades les eines següents:

- **Git.** El control de versions es realitza a través de Git [6], és una eina de codi obert, gratuïta i d'abast mundial. Git manté un seguiment dels canvis que es produeixen a mesura que es desenvolupa el projecte, d'aquesta manera es pot localitzar fàcilment d'on provenen possibles errors.

- **Trello.** Trello [7] és una eina per ajudar a planificar les tasques a realitzar. Així se sap en tot moment quines tasques ja s'han finalitzat, quines estan en procés, quines en procés però bloquejades i quines estan en el *backlog* (la bossa de tasques a començar).

5 Recursos

Per a la realització d'aquest projecte es necessiten dos tipus de recursos: personals i materials.

Recursos personals

- **Líder de projecte:** persona encarregada de dirigir i estructurar el projecte.
- **Expert en processament de llenguatge natural:** persona encarregada d'aportar el seu coneixement per a desenvolupar el projecte.
- **Equip de desenvolupament:** tres estudiants encarregats del desenvolupament del projecte. L'autor d'aquest treball hi dedicarà més hores que la resta de l'equip.

Recursos materials

- **Eines de desenvolupament**
  - Un ordinador per a desenvolupar el projecte i un altre a mode de servidor.
  - Sistemes operatius GNU Linux on s'hi treballarà.
  - Llenguatges de programació: Python i bash, amb els quals es desenvoluparà el treball.
  - Els entorns de treball Visual Studio Code per a desenvolupar el codi i Overleaf per a escriure la memòria.
- **Eines de control i gestió**
  - Trello per a gestionar les tasques.
  - Git per al control de versions.
- **Eines de comunicació**
  - Slack per a la comunicació interna de l'equip de desenvolupament.
  - Correu electrònic per a la comunicació amb la resta de l'equip.

6 Descripció de les tasques

Aquest projecte té una durada de 760 hores, que es repartiran al voltant de 8 mesos i van des del 1 d'Octubre de 2019 fins al 1 de Juny de 2020.
Pel que fa al desenvolupament del TFG, es dedicarà al voltant de 4 hores diàries de treball. Cal tenir en compte que fets com els exàmens de les diverses assignatures i les classes a assistir el fan variar una mica.

T1: Gestió del Projecte
Una bona organització del projecte és la pedra angular per a la bona realització d'aquest, és indispensable planificar amb antelació punts tant bàsics com els objectius a complir i el seu abast si no es vol perdre el full de ruta. Dins d'aquesta tasca es troben els punts:

- Reunió inicial
- Reunions de planificació
- Reunions de revisió
- Memòria del desenvolupament del projecte
- Determinar l'abast del projecte
- Planificació temporal
- Realitzar el pressupost
- Informe de sostenibilitat

T2: Scripts per al Preprocessament dels Documents
Un document requereix passar per un seguit de transformacions per tal de poder ser processat posteriorment, d'altra manera, molt text innecessari i inútil entraria dins l'*input* dels algorismes, que pot portar a resultats erronis. Dins d'aquesta tasca es troben totes les tasques que s'han dut a terme per tal d'assolir un bon text "plaï facilitar la feina a les tasques següents, en formen part:

- Cerca d'informació de les tècniques i eines a utilitzar.
- Implementació d'un script que extreu el text dels articles en pdf.
- Script per identificar i eliminar informació innecessària del text extret.
- Lematització del text.

T3: De Documents a Vectors i classificació
Aquesta tasca consisteix en passar a vectors els documents de text per a una millor representació d'aquests. La tasca inclou:

- Cerca d'informació de les tècniques i algorismes a utilitzar.
- Implementació de l'algorisme que dona la representació dels documents en vectors.

- Algorisme per al Clustering o classificació.

T4: Anàlisi i validació dels resultats
En aquesta tasca s'analitza i es valida el bon funcionament de les tasques anteriors i, si s'escau, retocar-les o modificar-les per a millorar els resultats i corregir-ne els erronis. La tasca inclou:

- Cerca d'informació de les tècniques i algorismes a utilitzar.
- Construcció de corpus escollit a mà per a fer els experiments.
- Realització dels experiments i anàlisi dels resultats.
- Possible modificació dels algorismes de la tasca anterior.

T5: Generació de mapes de relacions

- Cerca d'informació de les tècniques i algorismes a utilitzar.
- Implementació de l'algorisme de posicionament en un Graf.
- Generació de l'eina de visualització.

7 Estimacions i Gantt

El pas previ a poder realitzar un diagrama de Gantt és identificar quines dependències existeixen entre les tasques.

Presenten un seguit bastant lineal i uniforme, la Tasca 1 no depèn de ninguna, i la resta de Tasques depenen de tasques anteriors.

La Taula 1 mostra un resum de les tasques amb el seu codi, el seu temps estimat i les seves dependències.

La Figura 1 mostra el diagrama de Gantt, que permet, a simple vista exposar el temps de dedicació previst per a cada tasca.

Codi	Tasca	Temps	Dependències
T1	**Gestió del Projecte**	**200 h**	
T1.1	Reunió Inicial	2 h	
T1.2	Reunions de Planificació	26 h	
T1.3	Reunions de Revisió	26 h	
T1.4	Memòria del Desenvolupament del Projecte	102 h	
T1.5	Determinar l'Abast del Projecte	10 h	
T1.6	Planificació Temporal	14 h	
T1.7	Realitzar el Pressupost	10 h	
T1.8	Informe de sostenibilitat	10 h	
T2	**Scripts per al Preprocessament dels Documents**	**170 h**	
T2.1	Cerca d'informació de les tècniques i eines a utilitzar	50 h	
T2.2	Implementació d'un script que extreu el text dels articles en pdf	35 h	T2.1
T2.3	Script per Identificar i Eliminar Informació Innecessària	46 h	T2.1, T2.2
T2.4	Lematització del Text	39 h	T2.1, T2.2
T3	**De Documents a Vectors i Classificació**	**120 h**	
T3.1	Cerca d'informació de les tècniques i algorismes a utilitzar	42 h	
T3.2	Implementació de l'Algorisme de Text a Vectors	45 h	T2.2, T3.1
T3.3	Algorisme de Clustering	33 h	T3.1, T3.2
T4	**Anàlisi i Validació de Resultats**	**188 h**	
T4.1	Cerca d'informació de les tècniques i algorismes a utilitzar	37 h	
T4.2	Construcció de Corpus per a fer Experiments	40 h	
T4.3	Realització d'Experiments i Anàlisi de Resultats	52 h	T4.1, T4.2, T3.2
T4.4	Possible Modificació dels Algorismes	59 h	T4.3
T5	**Generació de Mapes de Relacions**	**82 h**	
T5.1	Cerca d'informació de les tècniques i algorismes a utilitzar	36 h	
T5.2	Implementació de l'Algorisme de posicionament en un Graf	26 h	T5.1, T3.2
T5.3	Generació de l'Eina de Visualització	20 h	

Taula 1: Taula d'estimacions horàries. Elaboració pròpia.

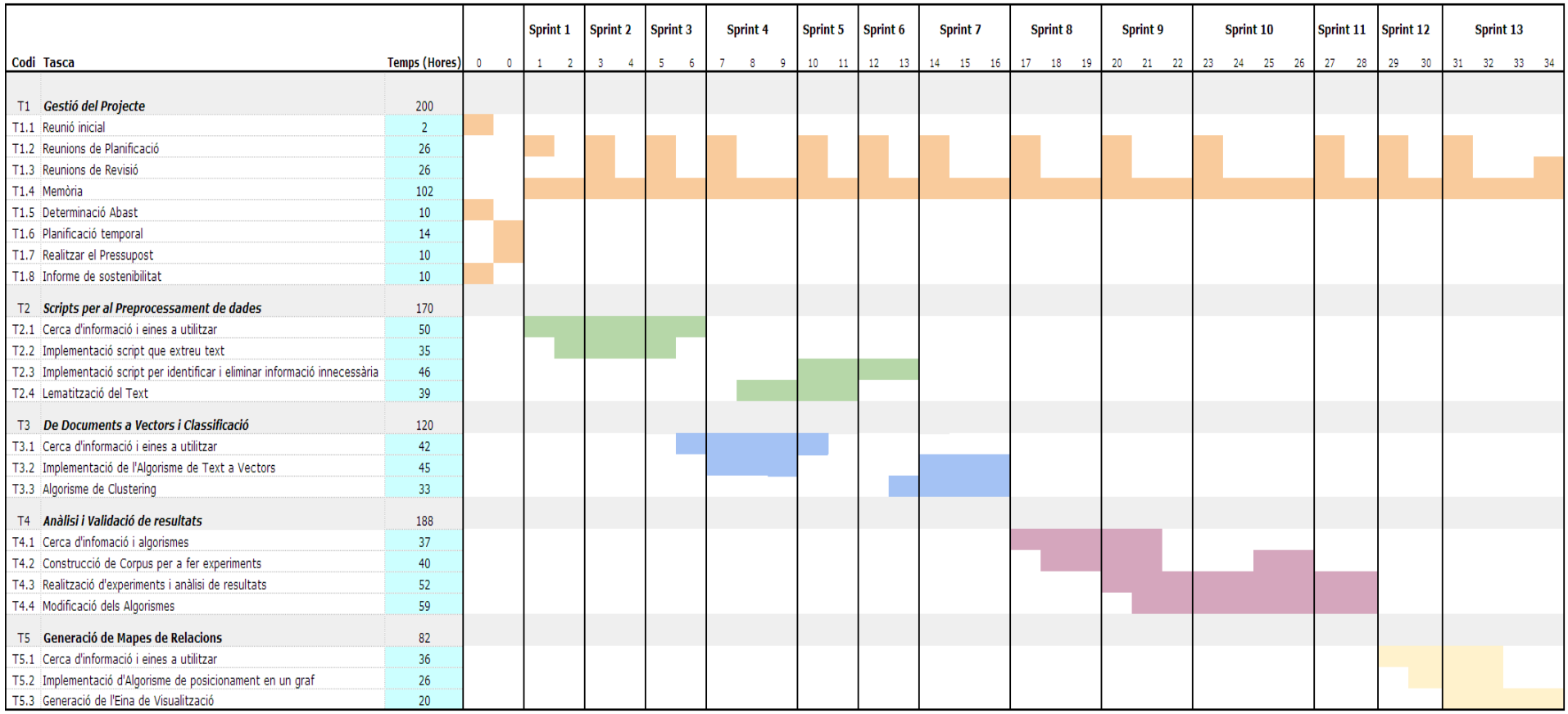

Figura 1: Diagrama de Gantt. Elaboració pròpia.

8 Implementació

En aquest apartat s'explicaran les tècniques i algorismes usats que, junts, formen aquest procediment per a generar mapes de coneixement. Primer s'explicaran les tècniques de pre-processament de les dades i després els algorismes per obtenir resultats. Finalment un petit pas a la generació dels mapes.

8.1 Mètodes, tècniques i algorismes

8.1.1 Pre-processament de les dades

D'ara en endavant s'anomenarà *corpus* al conjunt de documents amb els que es treballarà. Aquest son proporcionats pel servei de biblioteques de la UPC, i utilitzen el format PDF (Portable Document File). El problema amb els documents en PDF és que es tracta d'un format pensat per l'emmagatzemament i no la seva edició. És per a això que s'ha d'usar una forma que permeti separar entre dades i metadades dels mateixos i passar cap a un format més fàcil de tractar.

També, la majoria de documents que proporciona la biblioteca estan en anglès, és un detall a tenir en compte ja que la varietat d'idiomes dins el corpus suposa una problemàtica per als algorismes i tècniques que s'usaran, fet que s'explica més endavant.

Per al pre-processament de les dades fem servir aquestes eines:

Tika

Tika és un conjunt d'eines en Java que detecta i extrau les metadades i el text de documents en múltiples formats PPT, XLS i PDF[8]. L'avantatge d'aquest software respecte d'altres amb la mateixa finalitat és que permet mantenir l'estructura lògica de lectura del document, per exemple, un *paper* acostuma a estar maquetat en una o dues columnes, i Tika segueix el seu ordre lògic d'esquerra a dreta, quan altres eines no ho tenen en compte, és per això que s'ha triat.

Normalment, molts d'aquests documents contenen imatges o gràfiques que no es poden traspassar a caràcters llegibles i s'han d'eliminar en el traspàs a format *txt*. A més a més, cada document pot contenir una capçalera i un peu de pàgina que es repeteixen al llarg del document en funció del nombre de pàgines, s'han d'eliminar per a no tenir repeticions en el text que podria esbiaixar els resultats. Per a eliminar aquests fragments un petit algorisme que dona bons resultats és detectar els bytes que es repeteixen a l'inici de cada pàgina, i eliminar aquesta quantitat a cada pàgina. El mateix amb el final de cada pàgina. Aprofitem les metadades que ens proporciona Tika per a saber on comença i acaba cada pàgina i poder fer el procés anterior, ja que si només es tingués el document en format *txt* no podria saber-se.

La resta de metadades que extreu Tika [8] com per exemple el títol o la data de publicació, no és útil ja que ho proporcionarà la UPC de la mateixa manera que la biblioteca subministra els PDF, amb una API.

S'observa també que algunes equacions queden codificades en format *txt* de manera intel · ligible, per tant es passa un filtre per a eliminar aquestes paraules amb caràcters "estranys". Un cop arri-

bats aquest punt, els documents del corpus han passat d'estar en format PDF a ser *plain text*.

Freeling

A partir d'aquí és interessant treballar paraula per paraula per a poder categoritzar-les adequadament. Per a realitzar el procés s'ha triat *Freeling*, ja que és una eina *open-source* desenvolupada al centre de recerca TALP de la UPC [9].

Freeling és, entre altres, capaç també de detectar l'idioma d'un document, per a descartar-lo directament si no està en anglès. Això és necessari per evitar un problema amb el conjunt de documents i no del processament dels textos en si. Un cop extretes les paraules clau, suposant dos textos sobre fruita per posar un exemple, un en català i un altre en anglès, extreurà paraules com "poma" i "apple"respectivament, però l'algorisme ho detectarà com paraules diferents, tot i ser la mateixa. Al no disposar d'una eina de traducció fiable al cent per cent, es prefereix descartar directament tot document que no sigui en anglès.

Es necessita extreure l'arrel de cada paraula, per a això existeixen dos tècniques de normalització de textos: el *lemmatization* i el *stemming*.

- Stemming: trunca, a partir d'unes regles, el final de la paraula esperant obtenir un resultat correcte, tot i que funciona per la majoria dels casos, no sempre dona el resultat correcte.
- Lemmatization: és més elaborat, fa servir un anàlisi morfològic i de vocabulari per trobar el "lema"de la paraula, és a dir, l'arrel.

Alguns exemples es poden trobar a la taula 2:

	lemmatization	**stemming**
studying	study	study
studies	study	studi
was	be	wa

Taula 2: Exemples de lematització i stemming. Elaboració pròpia.

Es farà servir *lemmatization* ja que garanteix uns resultats més acurats i el *Freeling* dóna aquesta possibilitat. El text ha de passar per un procés on ho converteix tot en *tokens*, és a dir, es separa tot espais i es tracta cada paraula com un *token*. Un cop "tokenitzat" i "lematitzat"el text, separat per paraules, *Freeling* també assigna una etiqueta (substantiu, adverbi, signe de puntuació, etc). Caràcters que no interessen com poden ser números o signes de puntuació són eliminats i només deixa paraules amb significat.

8.1.2 Processament de les dades

Vectorització

Arribats a aquest punt, tenim un arxiu per cada article amb un conjunt de paraules i la seva categoria gramatical (si són substantius, adverbis, etc).

Ara el que es necessita és aconseguir una representació numèrica per a cada document, que es requereixen per a poder establir i observar relacions semàntiques entre aquests documents. Per a aconseguir aquesta representació es farà servir un algorisme d'aprenentatge no supervisat anomenat *paragraph vectors*, o **Doc2Vec**[10], que està basat en **Word2Vec**, que s'explicarà a continuació. S'utilitzarà aquest model, en comptes d'altres com el *TF-IDF*, ja que està pensat per a corpus de la mida de Google i, com ja se sap, de la UPC surten unes 200 publicacions al mes.

Word2Vec[11] és un algorisme que pretén reconstruir el context lingüístic de les paraules. Permet capturar relacions entre elles i ho fa generant una representació numèrica de cada paraula. Cada representació numèrica és un vector dins un espai i així es pot analitzar la relació que tenen diferents mots dins d'un conjunt de documents, com sinònims, antònims o analogies com la de la figura 2.

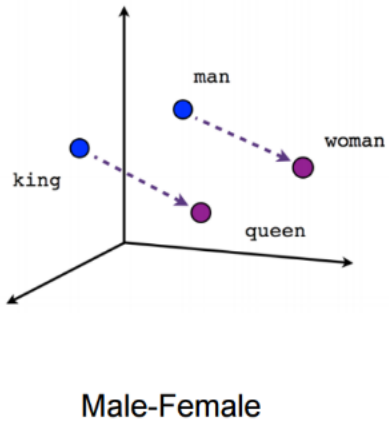

Figura 2: King is to queen like man is to woman. Imatge extreta de medium.com.

Utilitzant el concepte CBOW *(Continous Bag of Words)*, **Word2Vec** funciona com una xarxa neuronal on s'utilitza el context que envolta una paraula per poder predir-la, emmagatzemant-les com vectors de "característiques". Més endavant, quan s'ha entrenat l'algorisme, els vectors de carecterístiques passen a ser els propis vectors de paraules. Una altre forma de arribar al mateix resultat és el concepte de *Skip-gram*, contrari a l'anterior, on s'utilitza una paraula per a predir el seu context. Els dos algorismes funcionen dins del **Word2Vec**, CBOW és més ràpid i és emprat en paraules més freqüents i *Skip-gram* és molt més lent però molt més acurat en paraules menys comunes.

Doc2Vec és molt més actual i està basat en **Word2Vec** per tant, un cop explicat aquest és més fàcil d'entendre com funciona. **Doc2Vec** s'alimenta d'un corpus i genera una representació numèrica d'un document independentment de la seva longitud. El problema que s'origina al treballar amb documents i no paraules, és que els documents no segueixen una estructura lògica com l'estructura gramatical de les oracions. Per solventar-ho, s'usa el concepte de Mikilov and Le, on, seguint el model del CBOW, a més d'utilitzar el context d'una paraula per a predir-la, s'hi afegeix un vector extra que indica a quin paragraf, o en aquest cas document, pertany.

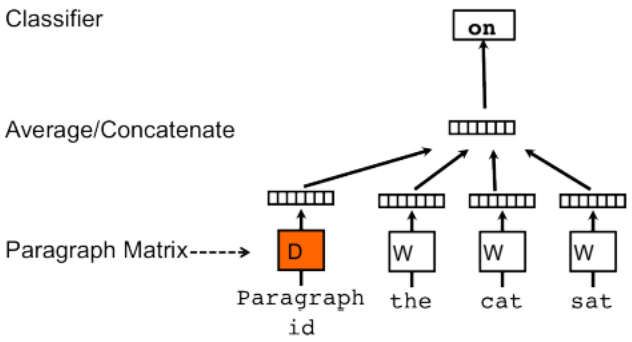

Figura 3: Architecture of PV-DM (Mikolov et al., 2014).

Aquest vector és únic per a cada document però, en canvi, el vector que representa una paraula concreta és el mateix en cada document. Al final de l'entrenament aquest vector afegit és el que conté una representació numèrica del document.

El model anterior és anomenat Memòria distribuïda d'un Vector Paràgraf(PD-DM, en anglès). Mentre que els vectors de paraules representen el concepte d'una paraula, el vector d'un document representa el concepte d'un document.

En aquest cas, utilitzarem la implementació de Doc2Vec que ha fet Gensim[12] ja que aquest *framework* es molt fàcil d'utilitzar i al estar fet en Python lliga amb la resta d'etapes.

Clústering

El *clústering* és una manera d'agrupar dades que consisteix en classificar-les segons un tret comú, per tal que comparteixin més característiques entre elles que en grups diferents.

Un cop tenim els vectors de paraules, necessitem els algorismes de *clústering* per aconseguir fer les agrupacions entre articles que tractin d'un tema similar, en el nostre cas 2 articles seran similars si els seus respectius vectors estan aprop en l'espai, obtenint així una agrupació per representar les àrees de coneixement de la UPC.

Provarem 2 algorismes de *clústering*:

- **K-means**
  La K del nom de l'algorisme fa referència al número de clústers que genera. Per fer-ho, s'inicialitza amb k punts aleatoris que pertanyin al conjunt de dades, i cada un d'ells serà un clúster. A partir d'aquí l'algorisme repeteix els següents passos fins a convergir:

  – Assignar cada dada al clúster més proper en distància euclídia.

  – Calcular el nou centre de clúster, fent la mitjana aritmètica entre els punts que hi pertanyen. El centre d'un clúster s'anomena centroide.

  Aquest algorisme presenta una sèrie d'avantatges i inconvenients.

  Avantatges

  – La seva implementació és senzilla.

  – Escala acceptablement bé amb el nombre de dades que ha de clusteritzar.

  – El seu funcionament és intuïtiu.

  Inconvenients

  – És molt limitat en quant a la possible forma dels clústers. La manera en què es calculen només permet formes entre circulars i ovalades.

  – Necessitem saber el número de clústers a priori.

  – És incapaç de sortir d'òptims locals.

  – Les assignacions són binàries. Això vol dir que un punt només pot pertànyer a un clúster.

  Alguns d'aquests inconvenients poden ser superats amb diferents tècniques.

  Pel número de clústers podem fer servir el mètode del colze *(elbow method)*, que consisteix en executar l'algorisme amb diferents valors de K entre 1 i l'arrel al nombre de dades. Per cada execució en calculem la variància, resultat de sumar les variàncies de cada clúster, i ho agrupem en un gràfic. El nombre de clústers òptim serà el que es trobi al "genoll de la corva" *(knee of the curve)*.

  Per les assignacions binàries, existeix una extensió anomenada *Fuzzy* que descobreix punts que poden pertànyer a diferents clústers. Per exemple, una samarreta pot ser blanca o negra (k-means normal), però pot ser que sigui blanca **i** negra, que és el que *Fuzzy* té en compte.

- **HDBSCAN**
  (Hierarchical Density-Based Spatial Clustering of Applications with Noise)

  És una extensió del DBSCAN que introdueix la component jeràrquica en què es basen altres algorismes de clustering.

  DBSCAN classifica els punts de tres maneres rebent 2 paràmetres, nombre mínim de punts *p* i distància ε:

  - Nucli: Un punt és considerat nucli si com a mínim p punts estan a distància $\leq \varepsilon$.
  - Accessible: Un punt *q* és considerat accessible si està a distància $\leq \varepsilon$ d'un nucli.
  - Soroll: Un punt que no és assolible des de cap altre punt és considerat soroll.

  Els punts nucli formen clústers amb altres punts nucli i punts accessibles, amb els accessibles actuant de frontera del clúster.

  Per fer l'extensió a jeràrquic, se segueixen una sèrie de passos.

  1. Canviar la mètrica de distància. Si abans es feia servir distància euclídia, ara es calcula com al màxim entre el *core* de cada un dels punts i la distància euclídia, sent el *core* d'un punt la distància entre ell i el *p* punt més proper.
  2. Calcular el *Minimum Spanning Tree* del graf complet en què cada punt és un node i les arestes tenen pes igual a la mètrica del punt anterior.
  3. Construir la jerarquia de components connexes. Per fer-ho, s'ordenen les arestes per pes i s'itera sobre elles, creant un nou clúster fusionat per cada aresta. Això es pot fer fàcilment amb una estructura de *union-find*.
  4. Condensar l'arbre resultant. Per cada nivell començant des de l'arrel, es descarten els fills que tenen nombre de nodes $< p$, mentre que els altres s'agrupen.
  5. Extreure els clústers. Es calcula l'estabilitat de cada grup de la següent manera. Sigui λ l'invers del pes, λ_{birth} correspondrà a l'aresta a partir de la qual s'ha decidit separar dos grups. L'estabilitat de cada grup serà $\sum_{p \in cluster}(\lambda_p - \lambda_{birth})$. Començant per les fulles, recorrem tots els nodes *n*. Sigui *t* el pare de *n*, si l'estabilitat de *t* és major o igual que la de la suma de les estabilitats dels fills de *t*, llavors aquest grup es considera clúster. En cas contrari, l'estabilitat de *t* passa a ser la suma d'estabilitats dels fills de *t*.

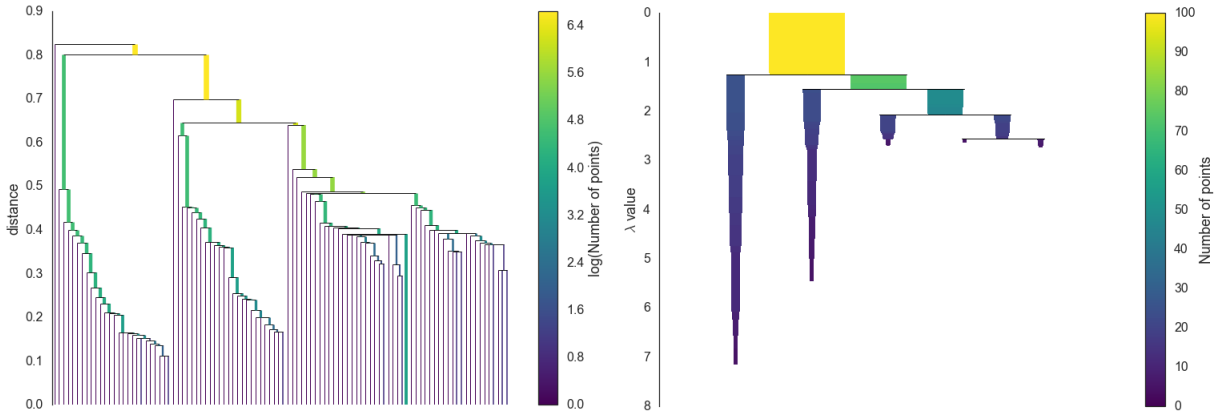

Figura 4: Exemple de resultat dels passos 3 i 4 en l'algorisme HDBSCAN. A l'esquerra dendrograma de components connexes ordenades per pes. A la dreta el dendrograma anterior condensat.

Aquest algorisme és més acurat donant resultats, no demana saber el nombre de clústers des de l'inici i pot computar clústers de mides complexes, com per exemple un clúster que envolta un altre. La part negativa és que és molt menys intuïtiu.

Ambdós algorismes estan implementats a la llibreria de Python d'aprenentatge automàtic *scikit-learn* i, per tant, s'utilitzen d'una forma semblant i se'n farà us.

8.1.3 Execució

En aquest apartat es mostraran un seguit d'experiments utilitzant les tècniques i algorismes comentats en l'apartat anterior fent proves amb diferents paràmetres per a escollir el més adient i comprovar el funcionament del procediment.

Tot això es farà sobre un corpus de documents controlat que ha estat proporcionat pel servei de biblioteques per a poder validar els resultats.

El *corpus* amb el que es treballarà està format per 1131 articles provinents de 18 departaments diferents, d'un total de 6 facultats. Els articles s'han triat de manera que n'hi hagi uns quants de cada tipus per a tenir prou dades com per a fer el *clustering*.

Seguint el procés de vectorització de cada article descrit en el punt anterior, obtenim vectors de característiques de diferents mides triades arbitràriament amb la finalitat de comparar-los entre ells. Típicament, per a projectes d'aquest estil s'utilitzen vectors de mides relativament grans, i per tant les mides dels vectors de característiques per als experiments seran 50, 100, 200, 300, 400 i 500.

K-means

Per a aquest experiment en concret el nombre de clústers ve donat pel nombre de departaments en què es basa el dataset. Això vol dir que un dels problemes de K-means, el fet de necessitar el nombre de clústers amb antelació, en aquest cas resulta ser un avantatge perquè sabem en quantes classes volem classificar els articles. Per aquest motiu, la K triada és 18.

El primer de tot és decidir la mida dels vectors que representen cada article. Per a fer això s'ha executat el mateix algorisme de *clustering* per a les diferents mides establertes.
Com que no s'utilitza cap sistema d'etiquetatge i es vol comprovar el percentatge d'encerts del procés, es farà servir l'algorisme hungarès (o algorisme de *Munkres*)[13], que resol el problema d'assignació en temps polinòmic. Així es pot assignar a cada departament un dels clústers resultants.

	50	**100**	**200**	**300**	**400**	**500**
Percentatge d'encerts	81%	80%	79%	81%	79%	79%

Taula 3: Percentatge d'encerts per a diferents mides dels vectors.

La taula de resultats anterior no mostra cap millora en afegir més característiques respecte les 50 inicials. Això probablement serà degut a que el nombre documents no es gaire elevat i no es necessita una dimensionalitat tant gran.

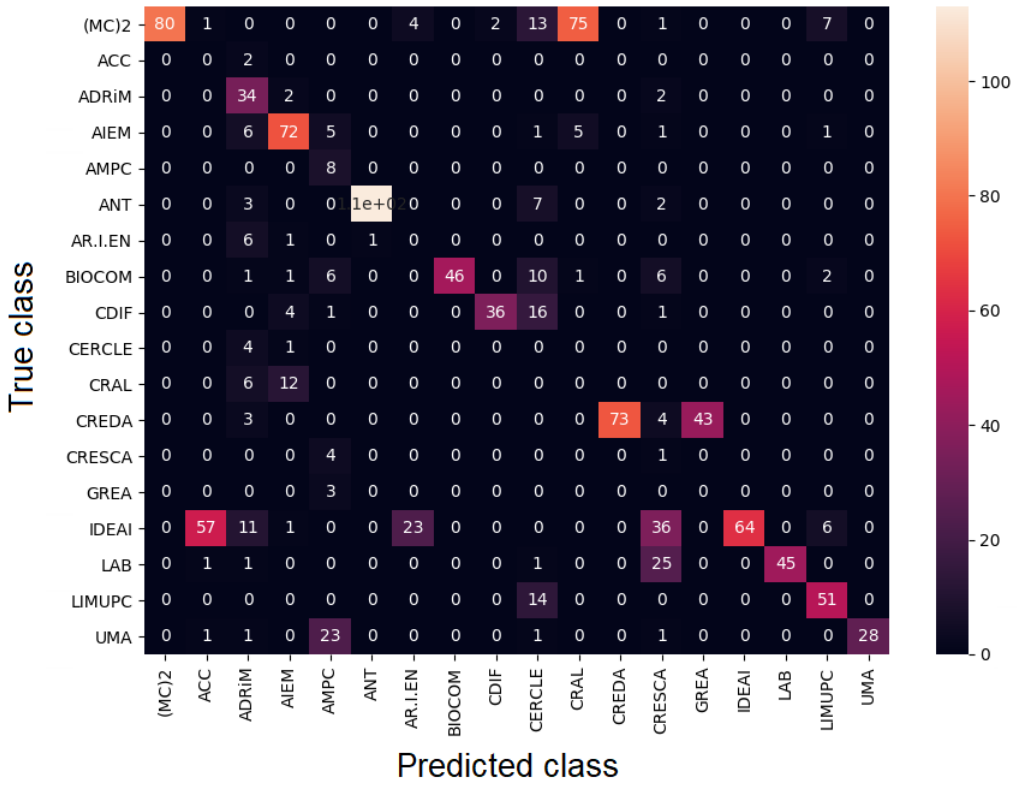

Figura 5: Matriu de confusió per a l'algorisme K-means.

Aquesta és la matriu de confusió resultant de l'execució amb K=18 fent servir el vector de 50 característiques com a entrada. A la diagonal principal apareixen els articles que han estat classificats correctament, és a dir, la predicció de la seva classe correspon a l'autèntica. Tota la resta de valors són articles que han estat mal classificats.

Hi ha casos en què el valor de la diagonal és 0. Això passa perquè d'aquella classe en concret no hi havia suficients mostres i els punts han estat classificats en altres llocs. Per altra banda, hi ha classes que han estat dividides en 2, com passa per exemple a la primera fila de la matriu. El motiu és que K-means tendeix a fer divisions més o menys homogènies, i el resultat és que classes desproporcionadament grans o petites queden dividides o absorbides, respectivament.

HDBSCAN

Com aquest algorisme no rep el nombre de clústers a l'entrada, l'objectiu serà que el nombre de clústers sigui el més proper al real possible. Fent experiments variant la mida mínima del clúster s'obtenen els resultats recollits en la taula 4.

Mida mínima d'un clúster	nº clústers	nº articles soroll
2	90	378
3	48	414
4	30	484
5	21	495
6	20	493
7	19	499
8	15	478
9	13	519

Taula 4: Nombre de clústers identificats i punts classificats com a soroll per a diversos valors de mida mínima dels clústers en l'algorisme HDBSCAN.

A la vista dels resultats, la mida mínima que s'ajusta més a l'objectiu seria el 7, i per tant la opció a triar.

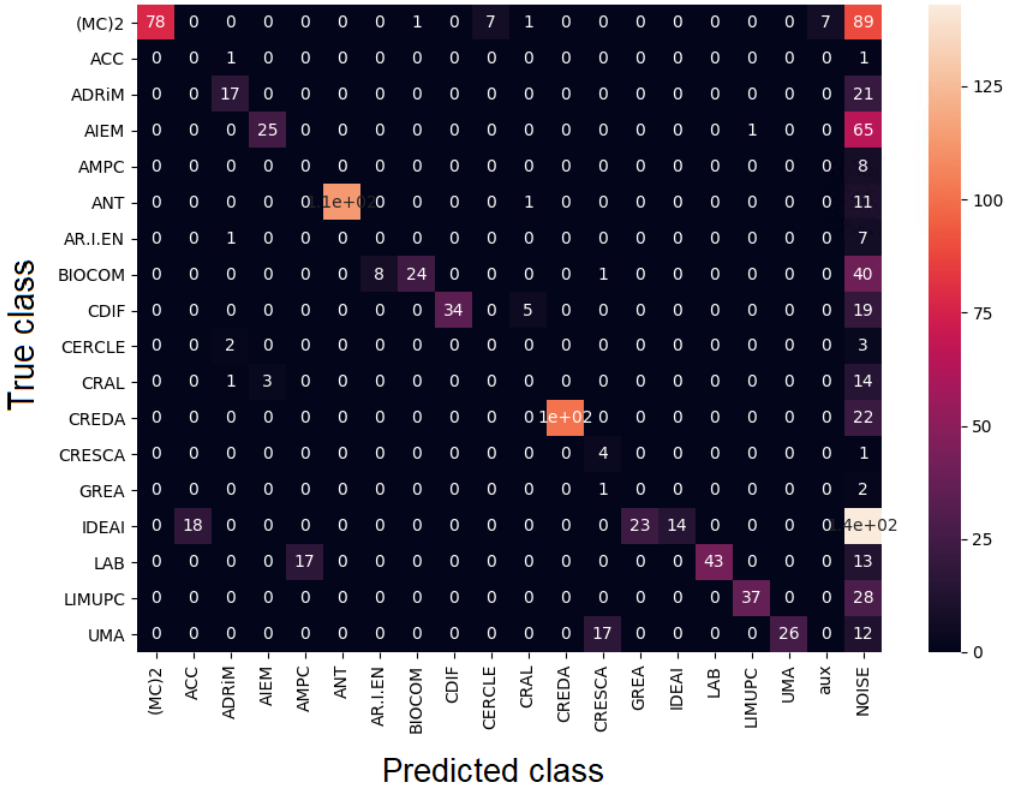

Figura 6: Matriu de confusió per a l'algorisme HDBSCAN.

De manera similar a K-means, el resultat esperat és que els articles dels departaments amb poques mostres acabin classificats com a soroll, i els que tenen suficients mostres acabin sent la seva pròpia classe. Els resultats, però, s'allunyen una mica d'aquest objectiu. Acaba havent-hi un gran nombre d'articles que queden classificats com a soroll, i repercuteix negativament en el percentatge d'encerts, que acaba sent d'un 68% per a la opció triada.

Experiment d'escalabilitat

Un altre aspecte a tenir en compte en el projecte és l'escalabilitat. Com la intenció és generar mapes amb tots els articles que es publiquen a la UPC, és important que la complexitat del procés escali acceptablement bé. Per aquest motiu es fa també un experiment per avaluar-ho. De les 5 etapes del procés, les primeres 3 tenen cost lineal ja que fan el tractament document a document. Així, només queda analitzar les etapes de l'algorisme de vectorització i de *clustering*.

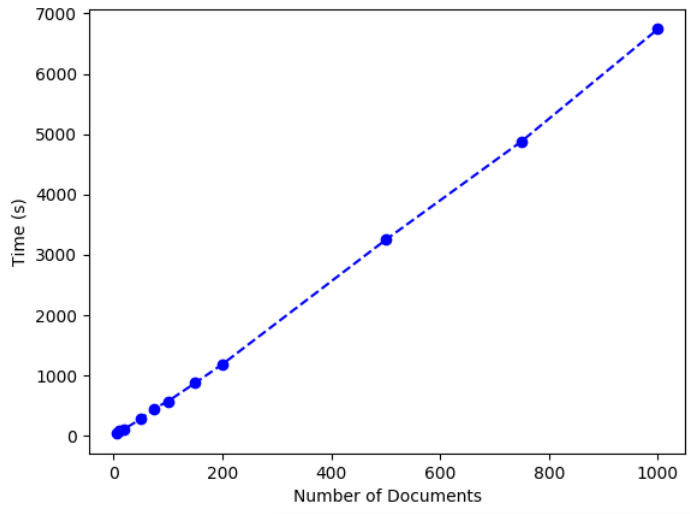

Figura 7: Temps d'execució de l'algorisme de vectorització per a diferents nombres de documents.

En la figura 6 es pot apreciar que el creixement en funció del nombre de documents és lineal. Per altra banda, els algorismes de *clústering* tenen un cost conegut que és quadràtic, i aquest és el millor cost al qual es pot aspirar.

8.2 Generació de mapes

El darrer pas del procés és generar un mapes que mostrin les relacions que es puguin establir entre articles, centres de recerca, autors, etc.

Degut a la gran dimensionalitat de les dades que s'extreuen del procés explicat no és possible mostrar aquest mapa com punts en l'espai. Una forma d'evitar aquest problema és transformar aquestes dades en un graf, utilitzant cada article com a node i les relacions entre aquests com a arestes. Les relacions entre els articles es poden definir de diverses formes: utilitzant les distàncies entre els vectors, ja sigui posant una aresta de cada node cap als nodes més propers sota un llindar o generar el graf complet i utilitzant com a pes de les arestes aquesta distància; utilitzant les paraules més rellevants de cada article, aquesta informació es pot obtenir utilitzant la mateixa xarxa neuronal que genera el Doc2Vec o amb tècniques com TF-IDF, i posar arestes entre els articles que comparteixin aquestes paraules; o utilitzant la clusterització i posar arestes de més pes entre articles que estiguin dins del mateix clúster.

Per a fer aquest graf llegible s'utilitzarà un algorisme *force-directed graph* per a establir la posició de cada node. Aquest algorisme busca minimitzar el creuament d'arestes i equilibrar la llargada de les arestes simulant forces entre els nodes per a empènyer o atreure altres nodes amb els que comparteixen arestes. Una aresta de més pes aplicarà una força major a una de menys pes.

En la figura 8 es mostra el corpus utilitzat per als experiments acolorint amb el mateix color els nodes que pertanyen al mateix clúster i utilitzant les 5 paraules més rellevants de cada article per a posar arestes entre aquells nodes que en comparteixin.

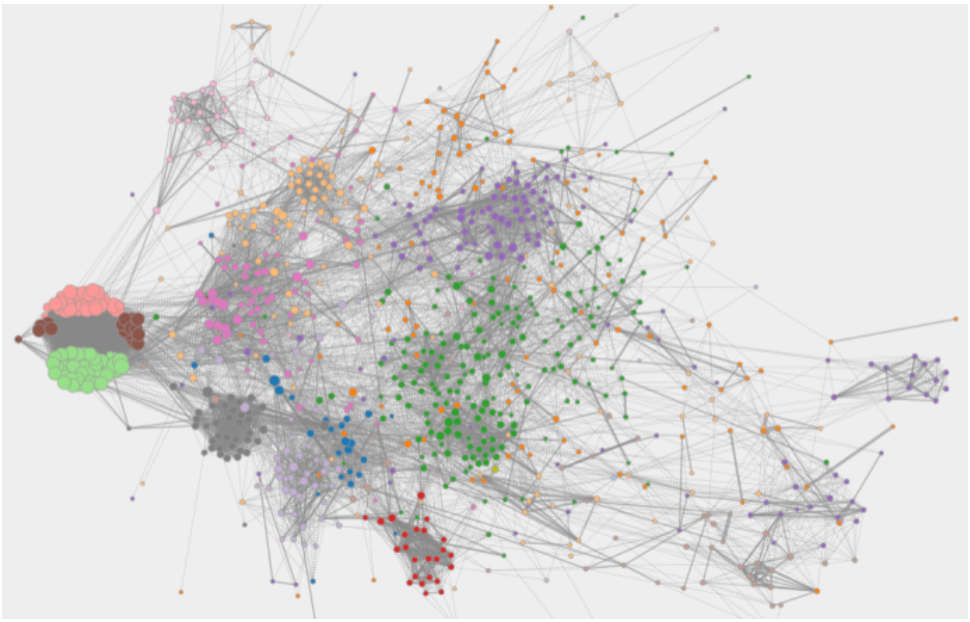

Figura 8: El mateix corpus que en els experiments com a *force-directed graph*.

Aquest mapa podria donar, a través de diferents consultes, informació rellevant per a que un expert pugui arribar a conclusions. Aquestes consultes poden ser per exemple autors freqüents en una zona, termes rellevants, consultar articles que estan a l'extrem entre 2 clústers utilitzant l'extenció *fuzzy* del k-means, entre altres.

9 Gestió del risc

En aquesta secció es tornaran a introduir els possibles obstacles que poden sorgir durant el desenvolupament del projecte i es proposaran solucions per a intentar, malgrat els impediments, acabar-lo en el temps establert.

Com ja es va mencionar en el lliurament anterior, els principals obstacles són:

Retard en el mitjà d'obtenció de documents
La tasca 4 necessita un mètode d'obtenció de corpus de documents reals per a poder fer experiments. Com ja s'ha explicat en el lliurament 1, és el servei de biblioteca qui proporcionara aquest mitjà. És un servei totalment independent als desenvolupadors del projecte així que la obtenció es pot veure afectada temporalment.

Aquest retard en l'obtenció dels documents faria impossible la realització de la Tasca 4. No es podrien fer els experiments necessaris per a validar que els algorismes i scripts anteriors tenen un funcionament correcte. Fer experiments amb un corpus fet a mà no fa creïbles els resultats que els algorismes poden donar.

Per a solucionar aquest problema, en el cas que es produeixi, s'optarà per a desenvolupar la tasca 5 abans que la tasca 4, ja que no hi ha dependències entre elles i la validació de la Tasca 4 es pot realitzar perfectament sense que afecti a la Tasca 5.

Poc potencial de les màquines
El volum de documents d'investigació que recull la UPC supera les 230.000 unitats. Processar-los tots a la vegada pot requerir una capacitat molt superior a la de les màquines de les que es disposa.

Si no resultessin ser prou potents, es tornaria a veure afectada la realització de la Tasca 4, ja que no es podrien fer experiments ni generar els mapes finals.

Una solució a aquest problema podria ser augmentar el pressupost per a obtenir màquines més potents que tinguin la capacitat per fer els càlculs que es requereixen. O, alternativament, contractar sistemes *cloud* més potents amb la mateixa finalitat. Una segona solució seria afegir una última tasca per a modificar els algorismes, i que aquests, a més a més, permetin fer els experiments de forma eficient en memòria, però això podria provocar que s'excedís el temps planificat. Com a última opció es reduiria l'abast del projecte: no es farien els experiments amb el corpus total de documents i s'hauria de dedicar més hores a la construcció manual d'un corpus que sigui considerable per a realitzar els experiments.

10 Gestió econòmica

Aquest projecte comporta diferents tipus de costos. Es classifiquen en diverses categories: costos de personal per activitat, genèrics, de contingència i d'imprevistos. Tots ells es calculen i es justifiquen a continuació.

10.1 Costos de personal per activitat

Cada tasca definida en l'apartat anterior comporta un cost econòmic. Durant aquest apartat es farà una estimació d'aquest cost respecte al cost humà per tasca.
El primer pas és determinar el sou per hora del les persones implicades en el projecte. Dins, hi trobem els perfils de treballador, que són: cap de projecte, programador, dissenyador i analista. La taula 2 mostra les retribucions dels rols que hi participen [14] tenint en compte l'aportació a la Seguretat Social (30% afegit al sou brut).

Rol	Retribució
Cap de Projecte	18€/h
Programador	12€/h
Analista	12€/h
Dissenyador	12€/h

Taula 5: Retribucions per a cada rol. Elaboració pròpia.

El rol de cap de projecte no estarà directament relacionat amb una persona física, sinó que es reconeix com a tal l'equip que participa a les reunions del final de cada sprint on l'autor del TFG en forma part. Els programadors, dissenyadors i analistes seran tot l'equip de l'inLab FIB encarregat de desenvolupar el projecte, on també s'hi inclou l'autor.

A partir de l'estimació horària de les tasques descrites a l'apartat de planificació temporal mostrades en la taula 1 i la seva planificació temporal en el Gantt a la figura 1, es pot crear una nova taula amb els costos per tasca. En aquesta, es calcula el cost tenint en compte les hores invertides a cadascuna d'elles. Quan més d'un rol intervé dins d'una tasca, es calcularà el cost a partir de la mitjana dels seus sous.

10.2 Costos genèrics

Els costos genèrics són els que no s'adjudiquen directament a una tasca, sinó que intervenen de manera indirecta a totes. Aquests són:

- Amortitzacions, de hardware i software
- Espai de treball
- Consum elèctric
- Desplaçament

Codi	Tasca	Temps	Rol	Cost
T1	**Gestió del Projecte**	**200 h**		
T1.1	Reunió Inicial	2 h	Cap de projecte	36€
T1.2	Reunions de Planificació	26 h	Cap de projecte	468€
T1.3	Reunions de Revisió	26 h	Cap de projecte	468€
T1.4	Memòria del Desenvolupament del Projecte	102 h	Cap de projecte	1836€
T1.5	Determinar l'Abast del Projecte	10 h	Cap de projecte	180€
T1.6	Planificació Temporal	14 h	Cap de projecte	252€
T1.7	Realitzar el Pressupost	10 h	Cap de projecte	180€
T1.8	Informe de sostenibilitat	10 h	Cap de projecte	180€
T2	**Scripts per al Preprocessament dels Documents**	**170 h**		
T2.1	Cerca d'informació de les tècniques i eines a utilitzar	50 h	Cap de projecte i programador	750€
T2.2	Implementació d'un script que extreu el text dels articles en pdf	35 h	Programador	420€
T2.3	Script per Identificar i Eliminar Informació Innecessària	46 h	Programador	552€
T2.4	Lematització del Text	39 h	Programador	468€
T3	**De Documents a Vectors i Classificació**	**120 h**		
T3.1	Cerca d'informació de les tècniques i algorismes a utilitzar	42 h	Cap de projecte i programador	630€
T3.2	Implementació de l'Algorisme de Text a Vectors	45 h	programador	540€
T3.3	Algorisme de Clustering	33 h	programador	396€
T4	**Anàlisi i Validació de Resultats**	**188 h**		
T4.1	Cerca d'informació de les tècniques i algorismes a utilitzar	37 h	Cap de projecte i programador	555€
T4.2	Construcció de Corpus per a fer Experiments	40 h	Analista	480€
T4.3	Realització d'Experiments i Anàlisi de Resultats	52 h	Analista i programador	624€
T4.4	Possible Modificació dels Algorismes	59 h	Programador	708€
T5	**Generació de Mapes de Relacions**	**82 h**		
T5.1	Cerca d'informació de les tècniques i algorismes a utilitzar	36 h	Cap de projecte i programador	540€
T5.2	Implementació de l'Algorisme de posicionament en un Graf	26 h	Programador	624€
T5.3	Generació de l'Eina de Visualització	20 h	Dissenyador i programador	240€
Total		760h		**11.127€**

Taula 6: Taula d'estimacions de costos de personal per activitat. Elaboració pròpia.

10.2.1 Amortitzacions

Les amortitzacions es calculen per a *hardware* i *software*. En el cas d'aquest projecte, el *software* utilitzat (git, Overleaf, Linux, Visual Studio Code, etc.) és totalment gratuït, és per això, que no hi ha amortització a calcular per al *software*.
Pel que fa al *hardware*, a la taula 4 es mostren les amortitzacions tenint en compte una durada del projecte de 8 mesos.

Hardware	Preu	Vida útil	Amortització
PC 1	1000 €	5 anys	133.33 €
PC 2	700 €	3 anys	155.55 €
Total			**288.88€**

Taula 7: Cost dels recursos *hardware*. Elaboració pròpia

10.2.2 Espai de Treball

El projecte es desenvolupa totalment a l'inLab, que es troba a la Facultat d'Informàtica de Barcelona. Dins d'aquest, només és necessari un despatx amb ordinadors i una aula de reunions. El preu per metre quadrat a Pedralbes és, de mitjana, 6.197 € [15], tenint en compte que la sala ocupa uns 60 metres quadrats s'afegeix un cost de 371,82€.

10.2.3 Consum elèctric i d'internet

El cost actual del kWh és de 0,1198€ [16]. Tenint en compte el *hardware* emprat per al projecte i les seves hores aproximades d'ús (és a dir, només quan està encès) es pot calcular el cost total a partir d'aquesta taula:

Hardware	Potència	Hores	Cost
PC 1	300 W	700 h	25€
PC 2	100 W	3672 h	44€
Total			**69€**

Taula 8: Cost del consum elèctric del *hardware*. Elaboració pròpia

Internet també consumeix recursos. Una tarifa actual d'internet costa aproximadament 40€ al mes (tarifes per a empreses i amb fibra òptica), tenint en compte la durada del projecte suposa un cost total de 320€.

10.3 Altres costos

10.3.1 Contingència

El fons de contingència neix de la necessitat de prevenir imprevistos en cas de que succeeixin. S'ha acordat fixar-lo en un 10%, per tant, la suma entre Cost Per Activitat (11.127€) més els Costos Generals és de 12.176,75 €, el fons de contingència suposa uns 1.217,677 €.

10.3.2 Imprevistos

És probable que durant el projecte s'hagin d'aplicar plans alternatius per culpa d'imprevistos, com es contempla en apartats anteriors. Ponderant el cost econòmic que suposarien i la probabilitat de la incidència, a la taula següent es calculen els costos per solucionar imprevistos:

Imprevist	Preu	Risc	Cost
Averia PC 1	1000 €	5%	50 €
Averia PC 2	700 €	5%	35 €
Poc potencial de les màquines (contractar sistema cloud) [17]	100 €	30%	30€
Total			**115€**

Taula 9: Costos generats per imprevistos. Elaboració pròpia.

10.4 Cost total del projecte

Finalment, a la taula 6 es mostra el cost total estimat del projecte, fet a partir de les estimacions parcials esmentades anteriorment.

Tipus	Cost
Personal	11.127 €
Amortitzacions	288,88 €
Espai de treball	371,82 €
Consum elèctric	389 €
Imprevistos	115€
Contingències	1.217,675 €
Total	**13.509,375€**

Taula 10: Estimació dels costos totals del Projecte. Elaboració pròpia

11 Sostenibilitat i compromís social

11.1 Autoavaluació enquesta

Un cop feta l'enquesta, n'he tret dues conclusions de les meves pròpies respostes. La primera és que mai m'havia plantejat l'impacte social i mediambiental dels meus projectes, fet que ha influït significativament a la majoria de les puntuacions dins d'aquesta. No ha estat fins aquest projecte que he hagut de requerir un informe de sostenibilitat, i, atenent a l'escenari global i el motor de canvi que suposem els enginyers, és obligatori per a nosaltres tenir en compte les conseqüències que pot tenir sobre el medi ambient la nostra feina.

La segona conclusió que n'extrec és la disparitat entre els meus coneixements sobre sostenibilitat i compromís social i la meva capacitat d'aplicar mesures per rebaixar-ne els que pugui generar. És a dir, algunes assignatures del grau tenien com a competència transversal "Sostenibilitat", on ens feien veure documentals i fer treballs sobre temes relacionats amb aquesta, però en ninguna assignatura obligatòria ens ensenyen, al menys, directament mètodes per reduir el nostre impacte tant ambiental com social. És per això que a l'enquesta he marcat amb puntuació alta les preguntes sobre coneixements sobre valoració d'impactes de les TIC, però amb puntuació baixa les de coneixements sobre aplicació de tecnologies "sostenibilístiques"aplicables.

Com a conclusió final, afirmo que hi ha moltes àrees relacionades amb la sostenibilitat a les TIC que encara desconec, però assumeixo la meva responsabilitat com a persona a punt d'entrar al mercat laboral de comprometre'm a ser més conscient amb l'impacte de tots els projectes en els que hi treballi d'ara en endavant.

11.2 Dimensió econòmica

Durant el desenvolupament d'aquest projecte s'ha procurat utilitzar programari lliure, de manera que no suposa cap cost econòmic durant tota la vida el producte generat.

A més, la maquinària necessària per a desenvolupar i donar el servei del producte generat no comporta un gran consum elèctric, el que redueix també el cost econòmic.

11.3 Dimensió ambiental

La dimensió ambiental d'aquest projecte tampoc és gaire extensa, s'utilitzen molt pocs recursos i, que siguin realment rellevants per al medi ambient, només el consum d'electricitat. Minimitzar els recursos que s'empren és bastant difícil, l'única solució és comprar dispositius que consumeixin menys, però afectaria a algunes tasques al només disposar de màquines poc potents, com ja es va explicar en apartats anteriors.
Aquesta canvi tampoc suposaria una millora del projecte, tenint en compte els recursos que s'han emprat en ell, és molt difícil reduir o optimitzar aquest aspecte..

11.4 Dimensió social

Personalment, aquest projecte només suposa el meu dia a dia en el meu lloc de treball. Tot i això, considero que m'ha aportat un munt de coneixements els quals no t'ensenyen a classe, com l'organització de projectes i altres aspectes més aviat relacionats amb el món laboral. Per a la gent a qui va dirigida aquest projecte (investigadors, docents, bibliotecaris, etc.) serà un gran avantatge i facilitarà molt la tasca tant de la classificació i l'etiquetatge dels articles, a simple vista pot semblar que no consumeix tant de temps i que no és necessari desenvolupar un projecte sencer per això, però la detecció d'àrees de coneixement emergents i comunitats entre elles és molt útil per encaminar les decisions que pot prendre la UPC sobre les seves inversions.

12 Conclusions

El processament del llenguatge natural segueix sent un procés complex però alhora extremadament útil per fer diverses tasques d'alt nivell. El fet de poder automatitzar qualsevol activitat farragosa sempre és una millora en l'ús del nostre temps, sempre i quant aquest procés sigui bo i fiable.

És destacable, per tant, el 80% d'encerts assolit en aquest projecte. Com s'ha vist, l'escalabilitat és bona i per tant es podria dur a terme amb un conjunt de dades substancialment més gran, que era, en definitiva, l'objectiu inicial. En definitiva, el procés de classificació de documents ha resultat ser automatitzable.

No s'han d'oblidar, però, els inconvenients que han sorgit. Els resultats obtinguts són bons però podrien haver estat millors. Cal destacar també que el corpus amb el que s'ha treballat és molt reduït i s'hauria d'experimentar amb una quantitat de documents de l'ordre del que es produeix a la UPC per obtenir millors resultats.

Tot i que encara queden moltes parts a polir, el projecte té un gran potencial. Les possibilitats que ofereix no s'han explorat al complet, i fins i tot es podria extrapolar a altres tasques de classificació d'estil similar. Hi falta molta investigació i experimentació per aconseguir crear un procés completament robust.

Referències

[1] Inlab fib, qui som? https://inlab.fib.upc.edu/ca/inlab-fib [Online, Accedit a 15-Febrer-2020].

[2] Dades estadístiques i de gestió. https://gpaq.upc.edu/lldades/ [Online, Accedit a 23-Febrer-2020].

[3] Elsevier. Scopus info. https://www.elsevier.com/solutions/scopus/how-scopus-works/content [Online, Accedit a 20-Febrer-2020].

[4] Futur. website for the scientific production of upc researchers. https://futur.upc.edu/ [Online, Accedit a 20-Febrer-2020].

[5] Agile Alliance. *"What is Agile Software Development?"*, 2013. https://www.agilealliance.org/agile101/ [Online, Accedit a 20-Febrer-2020].

[6] *"Git About"*. https://git-scm.com/about [Online, Accedit a 20-Febrer-2020].

[7] *"What is Trello?"*. https://help.trello.com/article/708-what-is-trello [Online, Accedit a 20-Febrer-2020].

[8] The Apache Software Foundation. Apache tika - a content analysis toolkit. https://tika.apache.org/ [Online, Accedit a 9-Març-2020].

[9] Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.

[10] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR.

[11] Tomas Mikolov, G.s Corrado, Kai Chen, and Je□rey Dean. E□cient estimation of word representations in vector space. pages 1–12, 01 2013.

[12] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[13] H. W. Kuhn and Bryn Yaw. The hungarian method for the assignment problem. *Naval Res. Logist. Quart*, pages 83–97, 1955.

[14] Indeed. Comparación de sueldos, buscar sueldos. indeed. https://www.indeed.es/salaries/ [Online, Accedit a 9-Març-2020].

[15] Ajuntament de Barcelona. Mitjana del preu d'oferta als barris 2013-2019. https://www.bcn.cat/estadistica/catala/dades/timm/ipreus/hab2mave/evo/t2mab.htm [Online, Accedit a 9-Març-2020].

[16] Papernest. Precio kwh españa: Información y tarifas 2020. https://www.companias-de-luz.com/precio-de-la-luz/kwh/espana/ [Online, Accedit a 9-Març-2020].

[17] Google. Todos los precios de compute engine. https://cloud.google.com/compute/all-pricing?hl=es [Online, Accedit a 9-Març-2020].