



Treball final de grau

**GRAU EN ENGINYERIA  
INFORMÀTICA**

Facultat de Matemàtiques  
Universitat de Barcelona

---

**EINES BASADES EN TEXT  
PER A LA VISUALITZACIÓ I  
GEOLOCALITZACIÓ DE  
NOTÍCIES**

---

**Autor: Solanes Font, Aleix**

**Director: Dr. Jordi Vitrià Marca**  
**Realitzat a: Departament de Matemàtica Aplicada  
i Anàlisi**

**Barcelona, 22 de juny de 2015**

# Abstract

This project came up after a few meetings with the professor Dr. Jordi Vitrià talking about Data Science, the possibilities inside this sector, about interesting projects that people made using Data Science, we talked about some cases that could be interesting to analyse, all with a spot in mind, the possibility of reaching a result and visualize that result in an understandable way. So, after those inspiring meetings we achieved the main idea of this project: Geolocation of news.

Regarding the journalism sector, hundreds of news are published every day, but unfortunately not all of them include a proper geographic reference. All those news with a reference to a geographic location let the door open to a new way of distribution, and also let new ways of analysis based on this new parameter, the location. As an example, imagine that a developer wants to let a user with a smartphone or a tablet, receive automatically the news related to the place where the user is. Without a reference to where the news took place it is difficult to face the problem, however with a reference to a place the problem now seems affordable.

During the realization of this project, various tools and methods are used in order to obtain automatically a set of georeferenced tags from the news and thus be able to show the results in an easy way to analyse.

Despite being this project based in the concept of Data Science, I am not explicitly using algorithms from the world of Data Science (in this report I will explain the reasons why some algorithms could not be used), I follow, however, the main principles: look for databases that can help me face the problem, clean all data in order to be able to use it, and finally I try to show the results in an understandable way to reach results or conclusions.

This document outlined in detail the design, development and implementation of every single step that was necessary to achieve the final result.

In order to see the results in a simple way, I also created a website with a summary of all the project and some visualizations of the results under the following URL: <http://alsolanes.github.io/TFG>

## Resum

Aquest projecte va néixer al cap d'algunes reunions amb el professor Dr. Jordi Vitrà en les quals parlàvem sobre Data Science i les seves possibilitats, sobre projectes interessants que la gent havia creat, sobre casos que seria interessant de poder analitzar, i tot amb una fita: la possibilitat d'arribar a algun resultat i fer que aquest es pogués representar visualment d'una forma entenedora.

En el sector del periodisme es publiquen centenars de notícies cada dia, però malauradament no totes elles inclouen la seva corresponent referència geogràfica. Que una notícia tingui aquesta referència, pot obrir nous camins en la distribució d'aquestes, així com permetre noves opcions d'anàlisi. Per exemple, considerem un desenvolupador que vol permetre a un usuari amb un smartphone o una tablet, que en funció d'on estigui, rebi les notícies automàticament d'aquesta zona. Sense cap referència a aquestes notícies seria difícil encara el problema, no obstant, si disposem d'aquesta informació el problema es simplifica.

En el marc d'aquest projecte, s'utilitzaran diferents eines i mètodes per a poder obtenir d'una forma automàtica les referències geogràfiques d'un conjunt de notícies i així poder-ne representar els resultats d'una forma que sigui senzilla d'analitzar posteriorment.

Aquest projecte neix de la curiositat que em despertava el món de la Data Science, o més concretament, el poder acabar extraient conclusions d'una sèrie de dades. Si bé no s'utilitzen explícitament algorismes pròpiament del món de Data Science (durant aquesta memòria s'explicaran els motius pels quals alguns algorismes no s'han pogut utilitzar), si que es segueixen les idees fonamentals d'agafar una gran quantitat de dades, netejar-les i finalment mirar de visualitzar les dades per a facilitar la tasca d'obtenció de resultats o conclusions.

Per tal de facilitar l'accés als resultats també s'ha habilitat una pàgina web amb el procés resumit així com la visualització dels diferents resultats obtinguts. La direcció en qüestió és:

<http://alsolanes.github.io/TFG>

# Agraïments

Vull agrair a ...

# Índex

<b>1</b>	<b>Introducció</b>	<b>1</b>
1.1	Big Data . . . . .	1
1.2	Data Science . . . . .	2
1.3	El projecte . . . . .	4
1.4	Antecedents . . . . .	4
1.5	Estructura de la Memòria . . . . .	7
<b>2</b>	<b>Definició d'objectius i motivació del problema</b>	<b>8</b>
2.1	Objectius . . . . .	8
2.2	Problema . . . . .	8
2.2.1	Cicle de vida d'una notícia . . . . .	8
<b>3</b>	<b>Desenvolupament</b>	<b>10</b>
3.1	Eines utilitzades . . . . .	10
3.2	Fonts de les dades . . . . .	12
3.3	Una visió alternativa, altres eines . . . . .	12
3.4	Preparació de l'entorn MongoDB i conceptes bàsics . . . . .	12
3.5	Obtenció i neteja de les dades . . . . .	14
3.6	Obtenció de localitats en un text . . . . .	20
3.7	Obtenció de les coordenades d'una localitat . . . . .	20
3.8	Formatació de les dades per a CartoDB . . . . .	20
3.9	Aplicació de l'algorisme DBScan . . . . .	20
3.10	Visualització de les dades . . . . .	20
3.11	Divulgació dels resultats . . . . .	20
<b>4</b>	<b>Metodologia i resultats</b>	<b>21</b>
4.1	Metodologia . . . . .	21
4.2	Resultats . . . . .	21
<b>5</b>	<b>Concordança de resultats i objectius</b>	<b>22</b>
<b>6</b>	<b>Conclusions</b>	<b>23</b>

# 1 Introducció

Uns dies abans d'escollir el projecte, vaig tenir l'oportunitat de reunir-me amb el Dr. Jordi Vitrià, per tal de parlar sobre les possibilitats i temes en general relacionats amb Data Science. Durant aquestes converses, van sorgir diferents conceptes, diferents projectes que s'havien realitzat al voltant del món de la Data Science, així com idees que alimentaven la idea de que aprofitant l'oportunitat d'escollir un treball final de grau, seria una bona idea experimentar lleugerament amb algun projecte que hi estigués relacionat.

Abans d'introduir el projecte, però, introduiré els conceptes de Big Data i Data Science, ja que en són les arrels.

## 1.1 Big Data

Amb l'evolució de les tecnologies de la informació, s'ha incrementat també la quantitat de dades que es produeixen a internet. En plantejar-se com tractar tota aquesta nova quantitat d'informació, es va veure que per exemple no era viable carregar totes aquestes dades en una base de dades relacional per al seu anàlisi. D'aquesta manera, va aparèixer el concepte de Big Data, per a fer referència a tota aquella informació que no pot ésser processada o analitzada utilitzant processos o eines que hi havia fins aleshores.

### Quins tipus de dades es poden generar?

Aquesta gran quantitat d'informació pot venir de diferents fonts, i fins i tot s'ha de tenir en compte que no només aquesta informació és generada pels humans, sinó que existeixen dades creades per màquines, com pot ser el cas de les M2M (Machine-to-Machine). Tipus de dades:

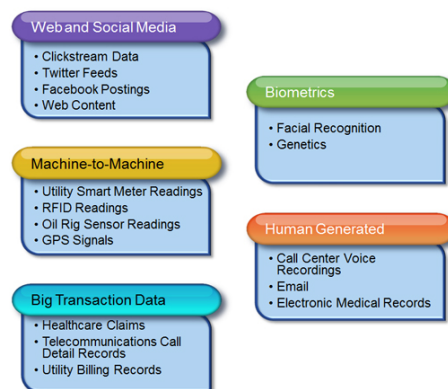


Figura 1: Tipus de dades a analitzar dins el sector Big Data.

- **Web i xarxes socials:** Es tracta de la informació generada en pàgines web, així com les dades obtingudes a través de les Xarxes Socials.

- **Machine-to-Machine(M2M):** Són les tecnologies que permeten la interconnexió entre dispositius. S'utilitzen dispositius com ara sensors o mesuradors per a generar dades, i fer que un altre màquina pugui interpretar aquestes dades i donar-los-hi sentit.
- **Big Transaction Data:** Són dades transaccionals, disponibles tant de forma semiestructurada com no estructurada. Entre els tipus de dades que inclou, s'hi troben per exemple registres de facturació o registres detallats de trucades (CDR).
- **Biomètriques:** Són dades biomètriques que inclouen informació sobre empremtes dactilars, escanejos de retina, reconeixement facial, genètica, etc. Són dades especialment importants en els sectors de seguretat.
- **Dades generades per humans:** És la informació que generem al enviar correus, escriure documents electrònics, en fer-nos estudis mèdics, deixar missatges de veu,...

I dins tota aquesta quantitat d'informació hi podríem incloure també la que pot generar una empresa de premsa, la qual genera centenars de notícies a diari amb les seves respectives meta-dades, les imatges, vídeos, i tota la informació que correspongui a cada notícia.

Una vegada tenim totes aquestes dades, és necessari un tractament per a poder mirar d'entendre i extreure'n conclusions, i és aquí on entra el món de la Data Science.

## 1.2 Data Science

Data Science, és un nou camp que està estretament lligat amb l'anàlisi de Big Data, però no es centra exclusivament en projectes de Big Data, ja que l'objectiu principal és l'extracció de coneixement d'una font de dades.

*"A Data scientist is somebody who is inquisitive, who can stare at data and spot trends. It's almost like a Renaissance individual who really wants to learn and bring change to an organization."*

- Anjul Bhambhri, vicepresident dels productes Big Data d'IBM.

Les persones que treballen en Data Science, s'anomenen Data Scientists, i la seva formació es basa en tres grans eixos: coneixement del negoci, capacitats en programació, i formació en matemàtiques i estadística.

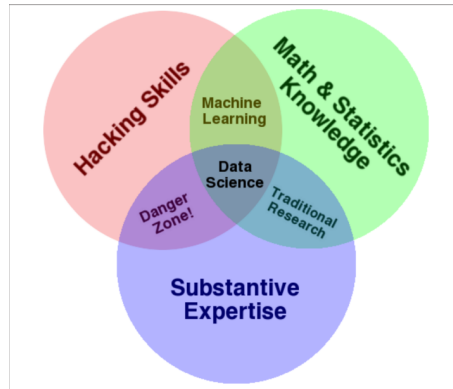


Figura 2: Diagrama de Venn de Data Science. De Drew Conway.

Segons Nathan Yau, doctor en estadística per la universitat de UCLA i redactor cap del portal divulgatiu [www.flowingdata.com](http://www.flowingdata.com), un expert en Data Science ha de tenir les següents capacitats:

- Estadística
- Data munging: capacitat per a manipular i adaptar les dades segons les necessitats (parsing, scraping i formatting data).
- Visualització de les dades (graphs, mapes,...).

L'aparició del terme Data Science, té els seus orígens en els requeriments que demanava per algunes ofertes de feina Google, i que demanaven les següents habilitats per als futurs candidats:

*"...working with a team on problems that require a hybrid skill set of stats and computer science paired with personal characteristics including curiosity and persistence."*

- Fragment d'una oferta de feina de l'any 2008 abans de l'acceptació del terme Data Science.

Veient que existia la creixent demanda d'aquest perfil, el Dr. Dhanurjay "DJ" Patil juntament amb Jeff Hammerbacher (Facebook i LinkedIn respectivament) van decidir encunyar el terme "Data Science" per a descriure aquest lloc de treball.



### 1.3 El projecte

El següent projecte es centra en l'anàlisi d'un conjunt de notícies, i obtenir de cada notícia un o varis punts geogràfics que representaran l'àmbit geogràfic d'on està parlant el text, per a posteriorment mostrar en un mapa els resultats d'aquest anàlisi.

Per a poder arribar aquests resultats el procés es divideix en quatre grans etapes:

1. Recopilació de notícies
2. Cerca i neteja d'una base de dades fiable que contingui noms de localitats i les seves coordenades.
3. Anàlisi del text de cada notícia recopilada per a extreure'n noms de ciutats.
4. Representar les localitats trobades en un mapa.

Totes aquestes etapes s'aniran detallant en el desenvolupament del següent document.

Una vegada es tinguin els resultats, els mapes resultants així com una explicació general del procés per arribar-hi, estarà disponible mitjançant una pàgina web pública, des de la qual es posarà el codi a disposició del públic.

### 1.4 Antecedents

#### Yahoo BOSS Geo Services

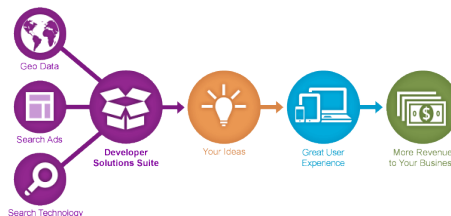


Figura 3: Esquema de negoci de Yahoo BOSS

Yahoo BOSS Geo Services, és un servei web que es divideix en dos projectes:

- PlaceFinder: Permet als seus usuaris trobar les coordenades concretes d'una direcció o localitat.
- PlaceSpotter: Donat un text que pot pertànyer a un twit de twitter, una notícia, una pàgina web, entre altres, retorna les localitats que ha pogut localitzar dins d'aquest text.

És un servei de pagament, i el seu cost varia en funció del nombre de queries diàries que es desitgin fer.

# CLAVIN



Figura 4: Logotip del projecte CLAVIN.

CLAVIN (Cartographic Location And Vinicity INdexer) és un projecte open source que té com a objectiu assignar "tags" a un document d'entrada especificant la geolocalització d'aquest text. Per a analitzar el text, combina diferents eines open source per a aplicar un processament de llenguatge natural i finalment, amb l'ajut d'un Gazetteer <sup>1</sup>, poder trobar les paraules referents a punts geogràfics correctament. Està especialitzat amb treballar sobre grans quantitats d'informació, Big Data, i és per això que per a processar aquestes grans quantitats de dades utilitza el framework Hadoop<sup>2</sup>.

Com a exemple de funcionament, aprofitarem que la seva pàgina web permet testejar la seva aplicació (<http://clavin.berico.us/clavin-web/>), agafarem un text qualsevol en anglès (en qualsevol altre idioma no funciona).

El text parla sobre un edifici de Portsmouth que el volen pintar dels colors de l'equip rival de futbol de la ciutat, i d'on a simple vista es poden localitzar com a ciutats d'on parla la notícia les ciutats de Portsmouth, Hampshire i Southampton. Podem comprovar com els resultats que ens indica són les ciutats de Portsmouth i



Figura 5: Pàgina web que permet testejar el funcionament del parsejador de punts geogràfics.

Hampshire, mentre que no ha posat la ciutat de Southampton, donat que la notícia en general no parla de Southampton i només ho menciona pel seu equip de futbol. També posa com a resultat els Estats Units en general.

<sup>1</sup>Diccionari geogràfic que conté correspondències entre informació geoespacial i noms.

<sup>2</sup>Framework de software que permet treballar amb milers de nodes i amb petabytes de dades.



Figura 6: Pàgina web que mostra a la part esquerra un mapa amb les ciutats trobades, i a la dreta la llista de ciutats junt amb les seves coordenades i el país.

## TextGrounder

TextGrounder és un projecte que pretén trobar mencions en un text sobre llocs geogràfics o referències temporals. Per a poder retornar la informació correctament, utilitza mètodes de processat de llenguatge natural (NLP) juntament amb algorismes de machine learning per a analitzar el context de la paraula i fer més fiable el resultat donat. Aquest projecte va néixer al setembre de l'any 2010 de les mans de Jason Baldrige (Universitat d'Austin) amb la intenció d'analitzar les referències temporals i geogràfiques en textos acadèmics del segle XIX.

Actualment està compost de tres subprojectes:

- Geolocalització de documents: Identifica la localització d'un document fent servir bases de dades d'entrenament, els quals contenen distribucions d'unigrames o bigrames que serveixen per a caracteritzar un text. Els textos utilitzats per a entrenar la base de dades principalment són textos de la Wikipedia.
- Geolocalització de topònims: Desambigua cada topònim d'un document, crea un model estadístic per a caracteritzar el text, igual que en l'anterior subprojecte, però en aquest cas s'utilitza un gazeteer per a verificar aquests topònims.
- Generador KML: Genera una sèrie de fitxers KML que contindran informació estadística sobre una sèrie de paraules, identificant per exemple les paraules més utilitzades al parlar d'una regió concreta de la terra.

## Google Maps API

La informació basant-se amb la base de dades dels seus mapes, està disponible a través d'aquesta API. Google ofereix dues versions de la API, una lliure i la versió "for work". La primera és una versió limitada, que ens permet realitzar un màxim de 2500 queries per dia i el projecte on s'utilitzi ha de ser gratuït per a tots els públics. La seva utilització es basa en fer crides a un webservice passant-li una sèrie de paràmetres. Així per exemple si desitgem les coordenades d'una direcció concreta, fariem:

```
https://maps.googleapis.com/maps/api/geocode/json?  
address=585+Gran+Via+de+les+Corts+Catalanes,+Barcelona+,+ES&key=API_KEY
```

Figura 7: Crida al webservice de la API de Google

Aquesta crida, ens retornaria una resposta JSON la qual contindria informació sobre aquesta direcció, informant-nos del tipus d'edifici que es tracta, la latitud i la longitud.

## 1.5 Estructura de la Memòria

La següent memòria estarà estructurada de forma que primer s'explicaran els *objectius i la definició del problema* per tal de tenir clar quina és la finalitat del treball.

A continuació es detallaran les diferents etapes de disseny i desenvolupament fins a arribar al resultat final. També es mencionaran aproximacions alternatives a la realització del problema que s'han intentat durant la realització del projecte. Aquesta informació estarà a la secció de *Desenvolupament*.

A continuació, a la secció *Metodologia i resultats*, es parlarà sobre les metodologies utilitzades durant la realització d'aquest projecte, així com les correspondències entre els resultats esperats, i els resultats reals.

Finalment, en l'apartat de *Conclusions i vies de continuació*, s'analitzarà el resultat final del treball, i es detallaran les possibles futures ampliacions que es podrien fer al treball, així com possibles solucions per a reduir les limitacions que s'han trobat durant la realització del projecte.

## 2 Definició d'objectius i motivació del problema

### 2.1 Objectius

Aquest projecte té com a principal finalitat permetre identificar els punts geogràfics on una notícia té lloc. Però per a poder arribar a aquest punt, s'han d'assolir una sèrie d'objectius on cada un d'ells té un pes igualment important per a poder obtenir uns resultats fiables. Els principals objectius marcats durant el projecte són:

1. Recopilar una sèrie de notícies per a poder ser analitzades
2. Crear una base de dades fiable amb les localitats que es tindran en compte per a la geolocalització de les notícies.
3. Permetre identificar, donat un text, les paraules que són susceptibles de ser una localitat.
4. Fent servir les bases de dades anteriors, mostrar en un mapa els diferents punts trobats, per a mostrar tant la evolució temporal, com el nombre total d'aparicions d'una localitat en les notícies.
5. Crear una pàgina web per a mostrar-ne els resultats, així com a permetre que altres usuaris puguin continuar o analitzar el desenvolupament.

### 2.2 Problema

En el sector del periodisme, es publiquen centenars de notícies cada dia. Per exemple, la font de referència utilitzada, el diari Ara en la seva versió digital, crea al voltant de 120 notícies diàries.

Cada una d'aquestes notícies conté una sèrie de "tags", no obstant, aquestes paraules no sempre segueixen un ordre lògic, o senzillament no permeten categoritzar les notícies segons la seva ubicació geogràfica.

#### 2.2.1 Cicle de vida d'una notícia

El cicle de vida d'una notícia conté tres grans etapes:

- Producció
- Distribució
- Consum

## **Producció d'una notícia**

Durant la producció, és el procés en què el periodista crea la notícia, l'escriu i li assigna una sèrie de paraules que permetran identificar aquesta notícia. Aquestes paraules, poden ser noms de persona que hi apareixen, noms d'empresa, alguna paraula clau,... però no necessàriament ha de contindre la localitat d'on s'està parlant, donat que sovint resta a disposició del periodista assignar les paraules que ell cregui rellevants per al text que ha escrit.

## **Distribució d'una notícia**

En aquesta etapa, la notícia ja ha estat creada, revisada, se li han assignat unes paraules clau, s'ha introduït a la base de dades del diari, i ja està en el punt de quedar a la disposició dels clients o usuaris que desitgin accedir a aquesta notícia. La distribució es pot fer en físic o en digital, tot i que en el nostre cas ens centrarem en la part digital. Aquesta distribució es sol realitzar a través de mètodes de distribució com ara RSS (Really Simple Syndication), a través de la seva pàgina web, o d'alguna aplicació mòbil.

## **Consum d'una notícia**

Finalment, l'usuari té la notícia a la seva disposició i la pot estar consultant. En aquesta etapa, no hi intervé res més que l'usuari, la notícia en sí i el terminal des d'on l'estigui consumint, ja que qualsevol etapa extra com podria ser el fet que s'ordeni o categoritzi les notícies es consideraria dins l'etapa de distribució d'aquesta.

Aquest projecte pretén ser una ajuda per a les etapes de producció (permetent per exemple ajudar a l'escriptor de la notícia a localitzar paraules susceptibles de ser localitats mencionades en la notícia), i per la distribució (permetent tenir les notícies categoritzades segons la zona geogràfica on té lloc la notícia).

## 3 Desenvolupament

Per al procés de desenvolupament del projecte he fet servir diferents eines que m'han facilitat les diferents etapes del procés, cada tecnologia utilitzada facilita una aproximació al problema, és per això, que també és interessant mencionar altres tecnologies que he pogut experimentar durant aquest projecte, tot i que finalment pot ser que no els hagi utilitzat per l'aproximació final. Tot seguit en presento primer les tecnologies utilitzades, i a continuació altres tecnologies que facilitarien una diferent visió del problema.

### 3.1 Eines utilitzades

#### Github



Figura 8: Logotip de Github.

És una eina que bla bla...  
flkj

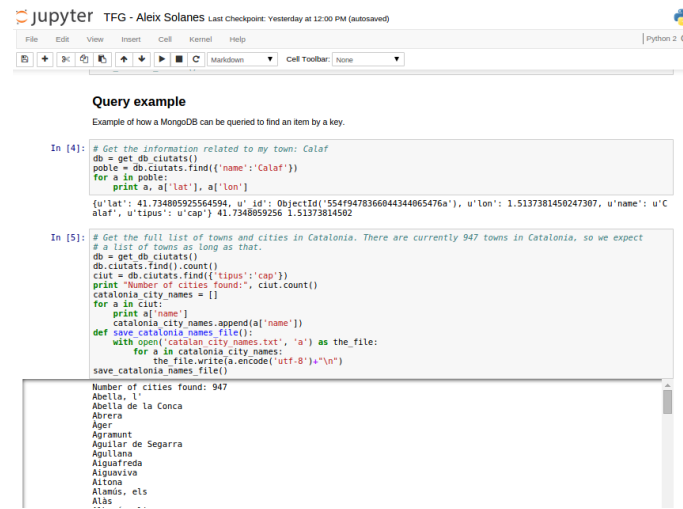
#### Python



Figura 9: Logotip de python.

És una eina que bla bla...

# IPython



**Query example**

Example of how a MongoDB can be queried to find an item by a key.

```
In [4]: # Get the information related to my town: Calaf
db = get_db.ciutats()
poble = db.ciutats.find({'name': 'Calaf'})
for a in poble:
    print a, a['lat'], a['lon']

{'u'lat': 41.734805925564594, 'u' id': ObjectId('554f9478366044344065476a'), 'u'lon': 1.5137381450247307, 'u'name': u'C
alaf', 'u'tipus': u'cap'} 41.7348059256 1.51373814502
```

```
In [5]: # Get the full list of towns and cities in Catalonia. There are currently 947 towns in Catalonia, so we expect
# a list of towns as long as that.
db = get_db.ciutats()
db.ciutats.find().count()
ciut = db.ciutats.find({'tipus': 'cap'})
print "Number of cities found:", ciut.count()
catalonia_city_names = []
for a in ciut:
    print a['name']
    catalonia_city_names.append(a['name'])
def save_catalonia_names_file():
    with open('catalan_city_names.txt', 'a') as the_file:
        for a in catalonia_city_names:
            the_file.write(a.encode('utf-8')+"\n")
save_catalonia_names_file()

Number of cities found: 947
Abella, l'
Abella de la Conca
Abrera
Ager
Agramunt
Aiguilar de Segarra
Agullana
Aiguafreda
Aiguaviva
Altona
Alamús, els
Alàs . .
```

Figura 10: Exemple d'una notebook d'IPython.



**RSS**

**MongoDB**

**JSON**

**CartoDB**

### **3.2 Fonts de les dades**

**RSS - Diaria Ara, Vilaweb, Regió7**

**Geonames**

**ICC - Institut Cartogràfic de Catalunya**

**Github.io**

### **3.3 Una visió alternativa, altres eines**

**NLP (Natural Language Processing)**

**NER (Named-Entity Recognition)**

**SPARQL**

### **3.4 Preparació de l'entorn MongoDB i conceptes bàsics**

Per a la gestió de les bases de dades que s'han utilitzat en el projecte, s'utilitzarà una aproximació NoSQL, MongoDB. Utilitzar una base de dades NoSQL no és una elecció que s'hagi pres sense motiu, ja que es buscava una escalabilitat i una agilitat que una base de dades relacional SQL clàssica no facilitava uns resultats tan bons. En definitiva el que es necessita és, partint d'un model clau-valor, fer cerques ràpides que retornin un valor al especificar una clau, i en això, els models NoSQL són idonis. Més concretament, MongoDB emmagatzema les seves bases de dades en format BSON (Binary JSON). BSON és una serialització amb codificació binària d'una sèrie de documents amb format JSON. Cada document o objecte, és el que es podria considerar com una espècie d'equivalent del que es coneix com a taula amb termes SQL.

Així doncs, una vegada definit l'entorn que s'utilitzarà per a la gestió de les bases de dades, es procedeix a la seva inicialització. La qual en un entorn python ens resultarà especialment simple fent servir la llibreria *PyMongo*. Una vegada s'ha engegat el servidor de la base de dades, que per executar-lo en un entorn Linux s'ha d'executar la comanda "*sudo mongod*", la inicialització bàsica és tan simple que

amb dues línies de codi Python permet deixar apunt per a ser utilitzada la nostra base de dades.:

```
client = MongoClient('localhost:27017')
db = client.ciutats
```

Figura 11: Inicialització d'una base de dades MongoDB amb la llibreria PyMongo.

Ara que ja es té inicialitzada la base de dades, afegir una ciutat amb les seves corresponents dades, serà fer un insert com el següent:

```
db.ciutats.insert({"name":name, "tipus":tipus,"lat":lat, "lon":lon})
```

Figura 12: Insert a una base de dades utilitzant PyMongo.

On "*name*"serà el nom del camp, i *name* el seu valor.

Per a poder fer consultes sobre una base de dades MongoDB, es pot utilitzar la comanda find de la llibreria PyMongo.

```
db.ciutats.find({'name':'Calaf'})
```

Figura 13: Exemple de consulta bàsica utilitzant PyMongo.

Com podem veure en la figura anterior, per a fer una consulta bàsica utilitzant l'instrucció find, només cal que especifiquem quin camp estem buscant i el valor concret d'aquest. En aquest cas la consulta ens retornaria una estructura JSON amb la informació disponible per al municipi de Calaf com la següent:

```
{u'lat': 41.734805925564594, u'_id': ObjectId('554f947836604436a'),
u'lon': 1.5137381450247307, u'name': u'Calaf', u'tipus': u'cap'}
```

Figura 14: Exemple de resposta d'una consulta bàsica utilitzant PyMongo.

### 3.5 Obtenció i neteja de les dades

La primera etapa del procés de desenvolupament del projecte, ha estat la cerca de les bases de dades corresponents que d'una forma fiable pugui permetre arribar a uns resultats acceptables. És una de les etapes del projecte més importants, donat que qualsevol resultat que se'n desprengui depèn estrictament d'aquestes dades. Per a poder trobar la base de dades final se n'han hagut de provar unes quantes, i fer la corresponent recerca. Fins i tot s'han intentat desenvolupaments alternatius per a construir manualment les bases de dades, com s'explica en la secció de *desenvolupaments alternatius*.

En aquesta secció, s'explicarà els processos seguits per tal d'*obtenir*, *analitzar*, i *processar* totes les bases de dades utilitzades en el projecte, per tal de en la següent etapa poder-les utilitzar sense problemes, considerant que hi haurem deixat només la informació important, i en el cas que sigui necessari haurem adaptat el format d'algunes dades per a facilitar la seva utilització.

En la introducció, hem parlat sobre el terme Data Science. Si recordem les tres habilitats que el Dr. Nathan Yau mencionava com a pilars en qualsevol Data Scientist (Estadística, Data munging i Visualització de les dades), aquesta etapa faria referència al terme "*Data munging*". Aquest terme, ve del verb en anglès "*Munge*", el qual significa "transformar dades d'una forma indefinida".

#### D'on extreure les bases de dades?

En aquest projecte, s'han utilitzat dues aproximacions per a l'obtenció de les bases de dades. Una és senzillament buscar una font fiable i descarregar-la, com en el cas de les ciutats i els seus punts geogràfics; l'altra, donada la impossibilitat de trobar de forma senzilla una base de dades que s'ajustés a les necessitats, ha estat crear-la, buscant un lloc d'on extreure les dades i emmagatzemant-la en una base de dades pròpia, com en el cas de l'obtenció de les notícies a analitzar.

#### Perquè modificar les bases de dades obtingudes?

S'ha de tenir present, que quan es facilita una base de dades, se'n sol especificar el format per tal que qualsevol persona que la vulgui utilitzar sàpiga com tractar-la. Per desgràcia, cada llenguatge pot tenir les seves particularitats alhora de tractar amb dades, i per això pot ser necessari donar un format concret a les dades prèviament obtingudes.

També s'ha de tenir en compte, que una base de dades pot incloure molta informació que de bon principi ja se sap que no és necessària per a la implementació d'un projecte, i en aquests casos, per temes d'agilitat i d'espai, també pot ser interessant destriar-ne aquella que resulti innecessària.

## Base de dades de pobles i ciutats de Catalunya

Per tal d'obtenir uns resultats el més bons possible, és molt important que les bases de dades que s'utilitzin continguin informació el màxim de fiable. És per això, que després de que s'hagin investigat i provat diferents alternatives, s'ha decidit optar per una base de dades facilitada per l'Institut Cartogràfic de Catalunya, la qual inclou més de 52.000 referències geogràfiques del territori català. Aquesta base de dades està disponible al públic sota el següent enllaç:

<http://www.icc.cat/index.php/Home-ICC/Publicacions/Nomenclator>

Aquesta informació, que s'ha definit com a base de dades, en un principi és considerat un Nomenclàtor. Aquest terme, fa referència a un document amb informació sobre un nucli poblacional, que pot ésser inferior a una unitat poblacional com pot ser el municipi (nucli de població, llogarets, parròquies,...), i en detalla informació sobre en quina forma s'hi assenta la població. A més, en facilita informació geogràfica, com ara les seves coordenades, cosa que és el que més interessa per a la realització del treball.

El primer pas abans de prosseguir és veure quins diferents tipus de dades inclou aquesta base de dades.

Abreviació	Significat
'cap'	Cap de municipi
'barri'	Barri, sector urbà (superior a 50.000 habitants)
'nucli'	Nucli de població(poble, llogaret...)
'diss.'	Veïnat disseminat
'e.m.d.'	Entitat municipal descentralitzada
mun.	Nom del municipi quan aquest no coincideix amb la capital
edif.	Edificació aïllada
edif. hist.	Edifici històric (ermita, església, castell,...)

Taula 1: Tipus de dades inclosos en el nomenclàtor del ICC.

Es pot veure doncs, que el que interessa més per aquest projecte és el terme 'cap', que fa referència a Cap de municipi. Per a comprovar que és el que s'entén com a municipi, senzillament es comprova que hi ha 947 referències sota el terme 'cap.', el mateix nombre exacte de municipis que es troben registrats a Catalunya.

### Estructura de les dades

El document que inclou tota la informació del nomenclàtor, és un document en format .xlsx (document específic del programa Microsoft Excel, a partir de la seva versió 2007). Per tant, el primer pas en obtenir unes dades, és veure'n el format, per a saber com encarar la seva refactorització per a poder utilitzar-la en el projecte.

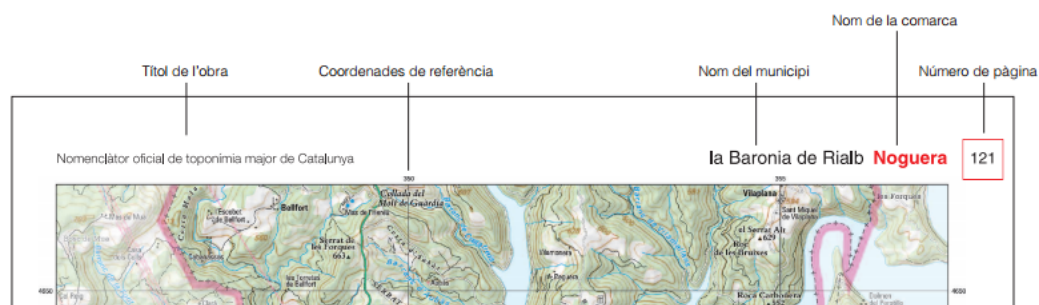


Figura 15: Nomenclàtor oficial de toponímia major de Catalunya.

Aquest fitxer d'Excel, fa referència al nomenclàtor oficial de toponímia major de Catalunya, una obra que inclou informació concreta per a cada municipi de Catalunya. És per això que en els termes que analitzarem a continuació hi podem trobar camps de referència a aquesta obra, com ara els camps Volum o el camp pàgina, que indiquen la localització dins la obra física del municipi especificat. Anem a veure les diferents columnes del fitxer d'Excel:

- **Topònim:** Nom del punt geogràfic.
- **Concepte:** Tipus d'assentament poblacional ('cap', nucli, barri, diss.,...)
- **Municipi 1:** Indica el municipi al qual pertany el topònim en qüestió, inclou 4 camps (Municipi 1, Municipi 2, Municipi 3, Municipi 4) a la taula, que indiquen el cas en que pugui pertànyer a múltiples municipis.
- **Comarca 1:** Indica a la comarca que pertany el topònim. Pot pertànyer a fins a 5 comarques, per tant inclou 5 camps (Comarca 1, Comarca 2, ..., Comarca 5).
- **UTM X i UTM Y:** Coordenades X i Y en el Sistema de Coordenades Universal Transversal de Mercator.
- **Volum:** Indicador que fa referència al volum físic de l'obra (Nomenclàtor Oficial de toponímia major de Catalunya).
- **Pàgina:** Indicador que fa referència a la pàgina dins el volum físic de l'obra.
- **UTMX i UTM Y:** Coordenades precises X i Y en el Sistema de Coordenades Universal Transversal de Mercator.

Topònim	Concepte	Municipi 1	Municipi 2	Municipi 3	Municipi 4	Comarca 1	Comarca 2	Comarca 3	Comarca 4	Comarca 5	UTM X	UTM Y	Volum	Pàgina	UTMX	UTMY
Abadal, l'	<u>edif.</u>	Avinyó				BAG					4140	46321	1	100	414000	4632100
Abadal, Torre	<u>edif.</u>	Sant Feliu de Llobregat				BLL					4204	45842	2	832	420400	4584200
Abadals, els	<u>edif.</u>	Castellbell i el Vilar				BAG					4039	46102	1	247	403900	4610200
Abadals, pla dels	<u>orogr.</u>	Sallent				BAG					4060	46275	2	772	406000	4627500

Figura 16: Nomenclàtor oficial de toponímia major de Catalunya.

## Neteja

Ara que ja sabem el format que tenen les dades, ja podem iniciar el procés per a extreure la informació que ens interessa i deixar-la apunt per a ésser manipulada. Per a fer aquest procés utilitzarem la llibreria *xlrd* per a python, la qual ens permet la manipulació de fitxers amb dades procedents del Microsoft Excel, més concretament ens permet l'extracció de dades de fitxers d'Excel .xls i .xlsx a partir de la versió 2.0.

De totes les dades que disposa la base de dades, es conservaran les dades referents al nom del municipi (columna "Topònim"), al tipus de municipi (columna "Concepte") i les seves coordenades geogràfiques corresponents a les columnes UTMX i UTMY. Per a les dades corresponents a les coordenades, es convertiran aquestes dades al sistema de coordenades esfèriques, on mantindrem els paràmetres de latitud i longitud en la nostra base de dades.

## Conversió de coordenades UTM a coordenades esfèriques (lat, lon)

La conversió entre coordenades UTM i coordenades esfèriques (lat, lon), no és una feina senzilla, ja que aquesta conversió radica en diferents sistemes de representar el món. Per a poder realitzar la conversió es realitzen diferents operacions algebraiques que involucren operacions hiperbòliques i múltiples passos en funció de la zona UTM on vulguem fer aquest canvi. És per això que el primer que es necessita saber per a fer el canvi de coordenades és en quina zona es troba el punt que es desitja convertir. Les dades que es desitgen convertir són coordenades totes situades dins de Catalunya, i per tant mirarem en el mapa de zones UTM quina és la que pertoca.

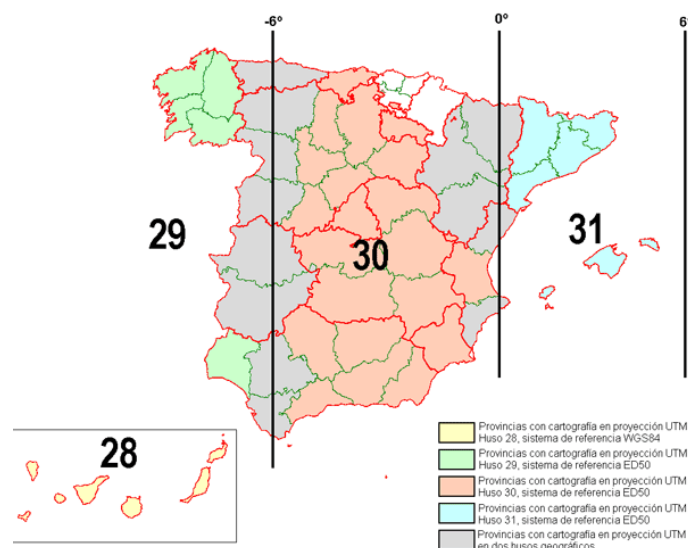


Figura 17: Mapa amb les zones UTM.

Com es pot comprovar, la zona corresponent a Catalunya és la 31. Ara es necessita saber la lletra corresponent a la zona, que per a Catalunya és la 'T'. Partint que per a la conversió, qualsevol coordenada dins de Catalunya tindrà com a zona la 31T, ja es pot fer servir una llibreria que ens permetrà la conversió de forma automàtica sense majors complicacions. S'utilitzarà la llibreria *UTM*, i més concretament la seva funció `to_latlon()`.

```
coordenades = utm.to_latlon(item['utmX'],item['utmY'],31,'T')
item['lat'] = coordenades[0]
item['lon'] = coordenades[1]
```

Figura 18: Conversió entre coordenades UTM i latitud/longitud.

Aquesta funció, com es pot veure a la figura anterior, retorna un vector amb primer el valor de la latitud, i després el corresponent a la longitud.

Una vegada ja es tenen totes les dades que es desitgen pel projecte, es recorre tot el fitxer `.xlsx` que conté la informació de l'ICC i per cada fila es van agafant els valors especificats anteriorment. Es fa la conversió de les coordenades, i finalment cada valor es guarda a la base de dades amb l'*insert* especificat anteriorment.

## Base de dades de ciutats de tot el món

El treball es centra en una base de dades de notícies escrites en català i d'àmbit nacional, és per això que s'ha fet especial èmfasi en cobrir el territori català amb detall, però a nivell mundial s'ha optat per considerar només les ciutats més poblades.

Trobat una base de dades fiable i que reunís els requisits demanats consistia en tenir informació sobre les coordenades de cada ciutat, i a més informació sobre la seva població, ja que sinó no se'n podria determinar quina ciutat és suficientment important per a ésser considerada.

La base de dades que s'ha escollit, ha estat la que manté [geonames.org](http://www.geonames.org). Conté informació que es va actualitzant dia a dia, ja que tot i que el nucli d'aquesta informació prové de fonts oficials, pot contenir errors i una comunitat activa de tot el món la va actualitzant i millorant. Aquesta base de dades és de lliure accés i ús sota una llicència Creative Commons.

La pàgina per accedir a les descàrregues de la base de dades, així com altres fitxers d'informació geogràfica diversa, es troba en el següent enllaç:

<http://download.geonames.org/export/dump/>

Si observem els diferents fitxers que se'ns facilita a l'enllaç, podem comprovar com n'hi ha de diferents tipus. Els fitxers que es poden trobar són:

- **XX.zip**: Conté informació per cada país, on XX correspon al seu codi ISO.
- **allCountries.zip**: Tots els països en un sol fitxer.
- **cities1000.zip**: Totes les ciutats amb una població major a 1000 habitants.
- **cities5000.zip**: Totes les ciutats amb una població major a 5000 habitants.
- **cities15000.zip**: Totes les ciutats amb una població major a 15000 habitants o capitals.
- **alternateNames.zip**: Conté dos fitxers: el primer que conté els noms alternatius d'una ciutat (i.e.: el nom de la ciutat en diferents idiomes, noms alternatius d'una ciutat,...); el segon, és un fitxer que conté les referències entre els codis ISO i l'idioma que representen (i.e.: el català està representat amb el codi CA o CAT)
- **admin1CodesASCII.txt**: Codis que fan referència a regions administratives en codificació ASCII.
- **admin2Codes.txt**: Codis que fan referència a regions administratives en codificació UTF-8.
- **iso-languagecodes.txt**: Codis ISO dels diferents idiomes. És el mateix fitxer inclòs dins del fitxer zip **alternateNames.zip**.



- **featureCodes.txt**: Nom i descripció d'una característica. (i.e.: per a una entitat política es representa de la següent forma: "*A.PCL political entity*").
- **timeZones.txt**: Informació sobre les hores de cada capital d'estat, concretament el desplaçament horari respecte DST i respecte el meridià de Greenwich.
- **countryInfo.txt**: Informació diversa sobre cada país, com per exemple els codis postals o els idiomes oficials.
- **...-;date;.txt**: Conté informació sobre les modificacions i eliminacions respecte l'anterior versió en cada un dels fitxers.
- **Altres**: Hi ha altres fitxers que contenen informació com per exemple els autors de les modificacions.

Com s'ha mencionat anteriorment, es desitjarà mantenir una base de dades amb les ciutats més importants, i tal com podem veure hi ha fitxers ordenats per la població de les ciutats. En aquest cas, s'utilitzarà la base de dades amb ciutats que tenen una població superior als 15.000 habitants, per tant el primer fitxer a tenir en compte serà `cities15000.zip` i el fitxer `alternateNames.zip` per a poder agafar de cada ciutat el seu nom en català.

## Obtenció

### Neteja

## Base de dades de notícies

### Obtenció

### Neteja

### 3.6 Obtenció de localitats en un text

### 3.7 Obtenció de les coordenades d'una localitat

### 3.8 Formatació de les dades per a CartoDB

### 3.9 Aplicació de l'algorisme DBScan

### 3.10 Visualització de les dades

### 3.11 Divulgació dels resultats

## 4 Metodologia i resultats

Agile?? desenvolupament iteratiu, avaluació, limitacions (producció i feedback usu-  
aris)

Metodologia es refereix a com has organitzat el desenvolupament: has fet servir  
GitHub, aproximacions successives a la solució, altres tècniques de soft engineering  
que hagi fet servir, com avalués les solucions, etc.

### 4.1 Metodologia

### 4.2 Resultats

## 5 Concordança de resultats i objectius

## 6 Conclusions

Bla bla

## Referències

- [1] Cohen, M.: TextGrounder: state of the art  
<https://mcohenlab.wordpress.com/textgrounder/>
- [2] The Google Geocoding API  
<https://developers.google.com/maps/documentation/geocoding/>
- [3] IBM Developer Networks: ¿Qué es Big Data?  
<https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/>
- [4] Yahoo BOSS Geo Services: Overview  
<https://developer.yahoo.com/boss/geo/>
- [5] Repositori GITHUB del projecte CLAVIN  
<https://github.com/Berico-Technologies/CLAVIN>
- [6] Repositori BitBucket del projecte TextGrounder  
<https://bitbucket.org/utcompling/textgrounder/>
- [7] PyPI - the Python Package Index  
<https://pypi.python.org/>
- [8] MongoDB  
<https://www.mongodb.com/>
- [9] Batut, C.; Belabas, K.; Bernardi, D.; Cohen, H.; Olivier, M.: User's guide to *PARI-GP*,  
[pari.math.u-bordeaux.fr/pub/pari/manuals/2.3.3/users.pdf](http://pari.math.u-bordeaux.fr/pub/pari/manuals/2.3.3/users.pdf), 2000.
- [10] Chen, J. R.; Wang, T. Z.: On the Goldbach problem, *Acta Math. Sinica*, 32(5):702-718, 1989.
- [11] Deshouillers, J. M.: Sur la constante de Šnirel'man, *Séminaire Delange-Pisot-Poitou, 17e année: (1975/76), Théorie des nombres: Fac. 2, Exp. No. G16*, pàg. 6, Secrèariat Math., Paris, 1977.