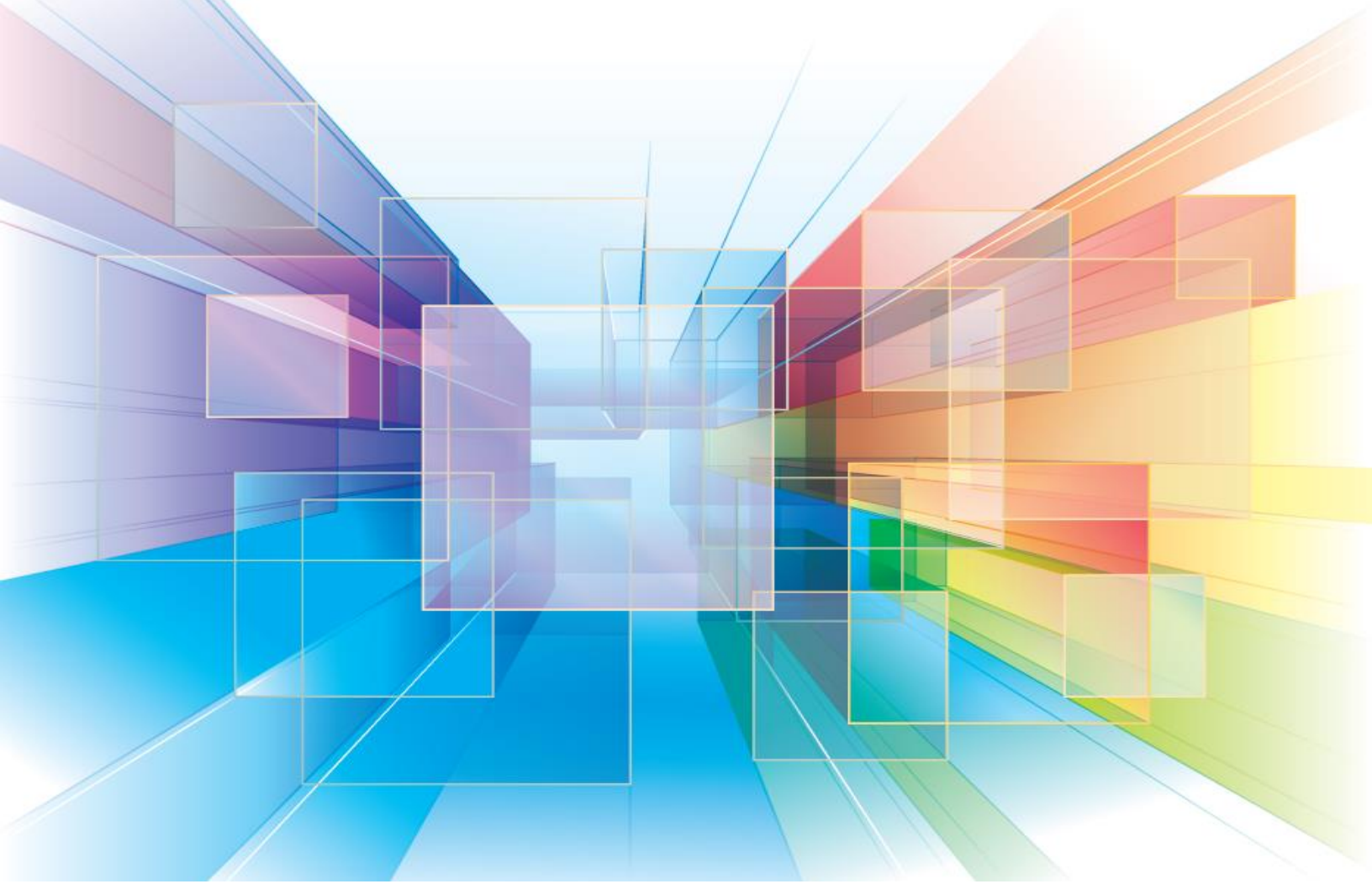
P.A.C.M.A.N. PROJECT

Group B: Madrid in Black

Crawling with Python

Developing a python code to scrape from TikTok

Sergio Castillo
Michal Dvořák
Michael Kazda
Johan Larm
Knut Lundgren
Mateo Sagaseta



INDEX

1-INTRODUCTION

What is P.A.C.M.A.N.? It is the acronym of Promoting Acquisition of Competences in a Multicultural and Associative Network. It is a project whose purpose is to connect computer science students from different countries of the European Union under the same web development work.

Our project arises from the need to create links and connections in a European area that grows every day and that requires understanding and communication among all its members. With the initiative of different institutes in some countries of the European Union and together with the collaboration of several companies aware of this reality, we have managed to create a link between programming students from these institutes so that they work together and harmonize both technical and cultural knowledge.

Grouping the students with other students of different nationalities, we propose to develop a web development project guided by companies specialized in information technology, being the tutors of said project.

In this way, we promote collaboration with people from different cultures, thus normalizing respect and consideration among students, empowering them, in the same way, in an international context and, in many cases, new to them.

"Men build too many walls and not enough bridges…"

Isaac Newton. -

1.1 Institutes and Members

There are four institutes that they could make real this project. Each one from a different country of the European Union:

- Carlow Institute of Further Education and Training from Carlow, Ireland.[1]
- IES El Lago from Madrid, Spain[2]
- SPS Na Proseku from Prague, Czech Republic[3]
- Hälsinglands Utbildningsförbund from Sörderhamn, Sweden[4]

During the meeting of all the institutes in Madrid, October 21th – 25th, our group was consolidated. We called it "Madrid in Black" and we are:

- Sergio Castillo from IES El Lago.
- Michal Dvořák from SPS Na Proseku.
- Michael Kazda from SPS Na Proseku.
- Johan Larm from Hälsinglands Utbildningsförbund.
- Knut Lundgren from Hälsinglands Utbildningsförbund.
- Mateo Sagaseta from IES El Lago.

[1] https://www.carlowifet.ie/
[2] https://www.educa2.madrid.org/web/centro.ies.ellago.madrid
[3] https://www.sps-prosek.cz/
[4] http://www.hufb.se/gymnasium/staffangymnasiet

1.2 Companies

During the development of the project, we have had the support and supervision of two companies, **BInfluencer**[5] and **Grupo OneTec**[6], both from Madrid. Our group was managing for BInfluencer, specifically, for Eduardo García Cantero, a web developer.

BInfluencer is a global Spanish *startup* that has set out to professionalize influence marketing, through a platform based on innovative technology that allows the best influencer to be identified for each campaign and defines the applicable rates according to objective parameters.

The main value that Binfluencer brings to the market is its digital platform, which combines artificial intelligence tools such as natural language processing and computer vision, as well as machine learning applications and its own algorithms. Through this platform, Binfluencer offers all types of companies, whatever their size and the sector they belong to, a comprehensive service for the management of Influence Marketing campaigns. In it you can select from more than 1 million Instagram profiles of more than 5.000 followers spread around the world.

On the other, Onetec is a great company that offers several services such as strategic consulting, digital solutions about business goals, E-payments, digital analytics, digital marketing and IT formation.

[5] https://binfluencer.io/
[6] https://www.grupoonetec.com/

2-SOFT SKILLS

To make this project possible, we have had to organize us and develop different social and technical skills like leadership, communication, understanding, empathy, etc.

The next points are the development of all this growth:

2.1 Organization

The project has been divided in several stages since the beginning. The **first one** was the week when we celebrated the meeting of all of the members of the project, students, teachers, and companies. During that week, we met each other and made the groups, participated in group activities.

The **second stage**, once the group had been consolidated, we started to learn about python through tutorials that our tutor of BInfluencer, Eduardo, gave us. In this part we practice with basic python coding.

In the **third stage**, Eduardo send us tutorials about scrape. We learned how to obtain information from webs with python's library **scrapy**.

The next step, the **fourth stage**, we began to scrape information from TikTok through Json code and stored it in MongoDB.

From this moment, Eduardo send us tutorials about how to use Jupyter Notebook, and two others python's libraries, Panda and Matplotlib, to make graphics with de scraped information. In this **fifth stage**, we learned about that and made the first graphics.

In **the final stage**, we analysed the graphs and obtained de conclusions of the whole project and write the documentation.

2.2 Responsibility

All group's members started to learn the basic information about Python, Scrape, MongoDB, and the others technologies, and make simple code about that.

When we had the knowledge, one part of the team made the crawling code to scrape Tiktok. Others was able to store the results of the scrape on MongoDB. Then, other part elaborated the graphs to analyse the results. And finally, each member made one part of the documentation that they had more control on it.

2.3 Communication

Since the beginning, our workgroup have been communicating via **WhatsApp**. This way has been fast and direct.

On the other hand, we chose a main agent in order to communicate with tutor in the institute and company, via email. He was also in charge for reporting the workgroup advances in **Teams** platform.

Once or twice a month, we had a **videocall** meeting with the company's tutor to resolve doubts and ask for the next steps to take.

2.4 Initiative

There has been initiative in each part of the project to be able to carry it out, from the beginning. Furthermore, In the same way, all members have had the predisposition to help the other with the problems that happen during the project.

2.5 Problem solving

One of the most important problems we had was that we only got basic information from TikTok users. This information was not sufficient to carry out the relevant analyses. We managed to solve it by means of a signature generator for each scrape that which let us obtain more information about user's media content.

2.6 Conflicts resolutions and Critical Thinking

In general, we have been successful in almost all group decisions. Each one could share his opinion and point of view and all of them have been considered.

Furthermore, continuous feedback was always offered among co-workers about their contributions.

2.7 Decision Making

All decisions were made and discussed under the consensus of all members, in addition, each one was allowed to give their opinion freely and frankly.

2.8 Willingness to Learn and Enthusiasm

Proactive attitude of most of the members of the group truly made possible to finish this project.

During the process, we could learn a lot about new technologies, also waking our curiosity up about deep knowledges in this area. For instance, our project has been considered a Big Data model in a small scale.

2.9 Teamwork & Cooperation

Since the beginning, our company's tutors are trying to learn us about **Scrum Methodology**. Scrum is a framework within which people can address complex adaptative problems, while productively and creatively delivering products of the highest possible value, in other words, Scrhum itself is a agile methodology for effective team collaboration on complex products.

We was able to complete this project thanks this method, because Scrum let us a quick system of analyse and sharing information step by step between all of group's members and our tutors.

2.10 European Culture

Each country owns tradition, language and culture. In spite of difficulties, we could perfectly share experiences, work and moments. Obviously, language was the hugest frontier; solved through our continuous efforts by speaking English.

3-OBJECTIVES

The main goal of the project consists of developing a **crawler** by python. A web crawler[7] (also known as a web spider or web robot) is a program or automated script which browses the internet in a methodical, automated manner. This process is called web crawling or spidering.

Specifically, we have to obtain a scraping code to make crawling from **TikTok**[8]. TikTok is a video-sharing social networking service owned by ByteDance, a Chinese company founded in 2012 by Zhang Yiming. It is used to create short dance, lip-sync, comedy and talent videos.

With this web crawler we try to scrape information about the number of followers that have the most popular users of this social networking, the different hashtags that they use and making comparation between countries. That information is stored in MongoDB, no relational data base. Furthermore, we analyse these data through graphics, using pandas and matplotlib, two libraries of python, whit the objective to find a relationship between the concept "*European culture and heritage*" and those countries.

Another objective of this project is building bridges between European countries, such us Spain, Ireland, Sweden and Czech Republic, through the collaboration of students of each country working together.

[7] https://en.wikipedia.org/wiki/Web_crawler
[8] https://www.tiktok.com

4-GANTT DIAGRAM

DIAGRAMA DE GANTT																													
TASK	START	END	02/12/2019	09/12/2019	16/12/2019	23/12/2019	30/12/2019	06/01/2020	13/01/2020	20/01/2020	27/01/2020	03/02/2020	10/02/2020	17/02/2020	24/02/2020	02/03/2020	09/03/2020	16/03/2020	23/03/2020	30/03/2020	06/04/2020	13/04/2020	20/04/2020	27/04/2020	04/05/2020	11/05/2020	18/05/2020	25/05/2020	29/05/2020
Start learning basics python	02/12/2019	22/12/2019																											
Make GitHub Repository and Learn Scrapy	23/12/2019	10/01/2020																											
Practice with Scrapy on TikTok	06/01/202	02/02/2020																											
Learn about MongoDB	28/01/2020	16/02/2020																											
Practice with MongoDB to store scraped information	03/02/2020	01/03/2020																											
Learn about Pandas, Matplotlib and Jupyter	02/03/2020	22/03/2020																											
Elaborate graphs with the scraped information	16/03/2020	26/04/2020																											
Analize graphs and write it on Jupyter	20/04/2020	10/05/2020																											
Start the documentation	11/05/2020	24/05/2020																											
Prepare the final presentation	25/05/2020	29/05/2020																											
FINAL PRESENTATION	29/05/2020																												★

5-TECHNOLOGIES

We have used several kinds of technology in this project.

- **Python**[9]: It is an interpreted, high-level, general-purpose programming language. That has been the core of whole project because it is the language that we used to make the crawler.

- **MongoDB**[10]: It is a cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with optional schemas. All information scraped by our crawling was stored on MongoDB.

- **Pandas**[11]: It is a software library written for the Python programming language for data manipulation and analysis. It offers data structures and operations for manipulating numerical tables and time series.

- **Matplotlib**[12]: It is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, etc. That library let us to show and represent the scraped data through graphs.

- **Jupyter Notebook**[13]**:** It is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numeral simulation, statistical modelling, data visualization, machine learning, etc.

[9] https://www.python.org
[10] https://www.mongodb.com
[11] https://pandas.pydata.org/
[12] https://matplotlib.org/
[13] https://jupyter.org/

- **GitHub**: It provides social networking-like functions such as feeds, followers, wikis and social network graph to display how developers work on their versions ("forks") or a repository and what fork (and branch within that fork) is newest.
- **Microsoft Teams:** It is a unified communication and collaboration platform that combines persistent workplace chat, video meetings, file storage (including collaboration on files), and application integration. We used that to be communicated with the rest of members of the project.


6- CODING & DEPLOY: PROBLEMS AND SOLUTIONS

In reference with coding, we haven't had any significant programming problems with the exception of the small typical mistakes that were solved together by helping each other.

About deploy, the main problem was that we only got the information of the user's profile and we couldn't get information about videos, hashtags, music, etc.

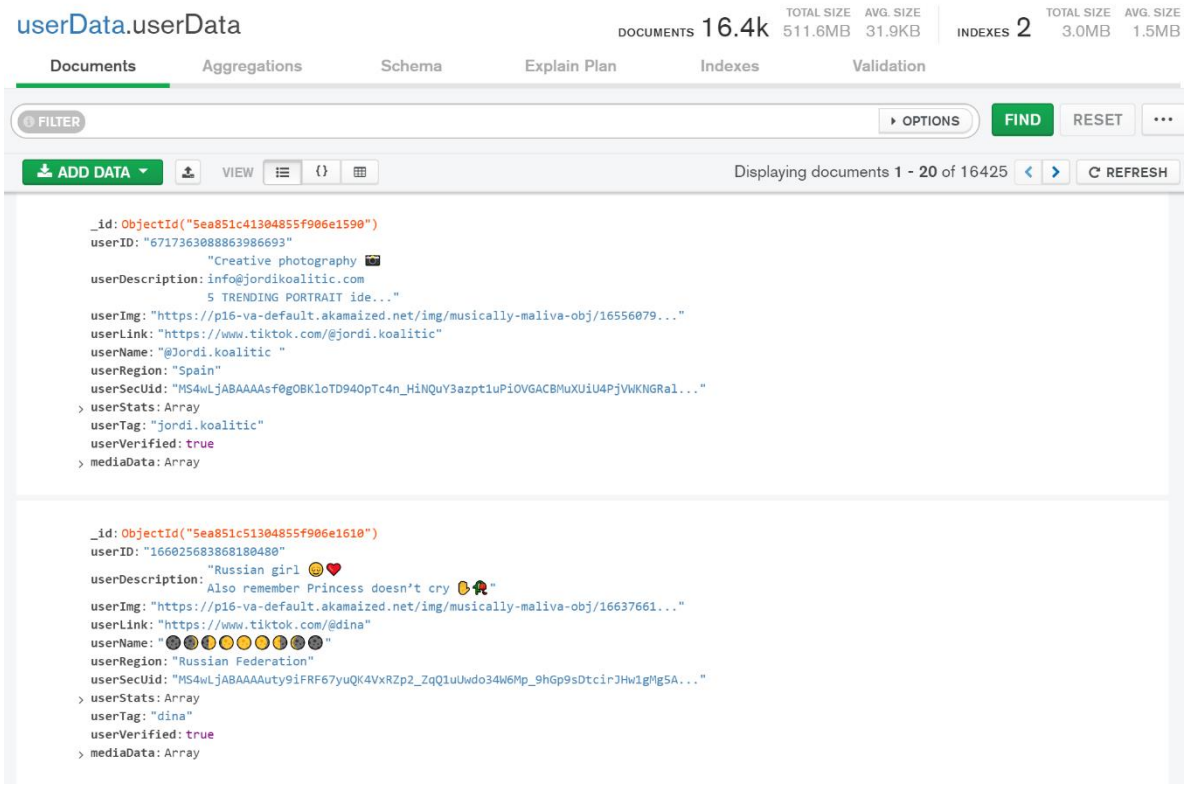
To be able to get the media data we needed a signature generator and we found a Github repository[14] which was going to give us the necessary signature to be able to continue with the work.

Once we could get the media data we kept working and started making graphs with Matplotlib and data views with Pandas.

[14] https://github.com/carcabot/tiktok-signature

7-SYSTEM TESTING

We have used **MongoDB Compass** which is program that allows any user within our team to visualize and explore our data scraped and be assured if they had been inserted into the database.



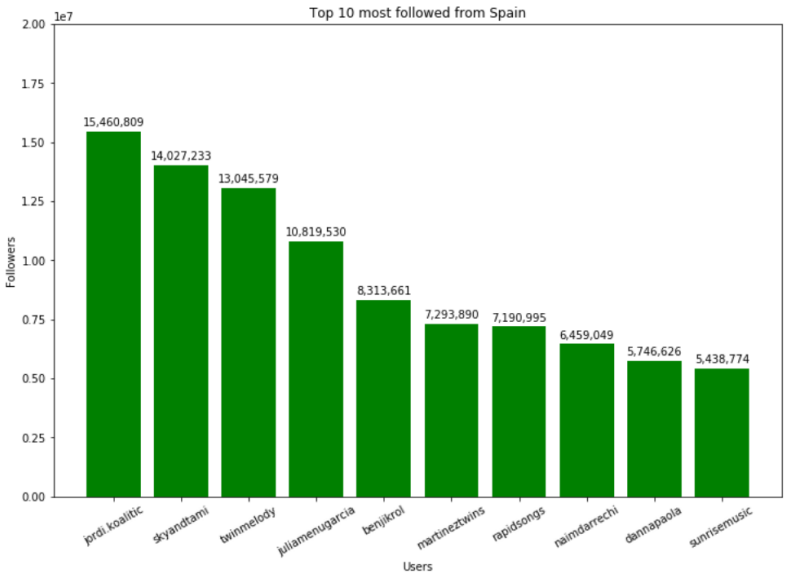
This is an example of the stored data bout two users.

We have used **Jupyter Notebook** to made some tests about the graphs and data views because it is a fast and ordered way to show them. In addition, when an alert is thrown is displayed more clearly than by console and is easier to detect the mistake.

In [2]:
```
import pymongo
import pandas as pd
from pymongo import MongoClient
import matplotlib.pyplot as plt
plt.rcParams['figure.figsize'] = [12, 8] #Set the plot size
uri = 'mongodb+srv://packman:MIB123456@packman-mib-wil2x.azure.mongodb.net/test?retryWrites=true&w=majority'
client = MongoClient(uri)
db = client.userData
collection = db.userData

europeanUnionCountries= ['Spain','Germany','France']
numUserStats = "1"
def autolabel(rects):
    #Attach a text label above each bar in *rects*, displaying its height.
    for rect in rects:
        height = rect.get_height()
        plt.annotate('{:,}'.format(height),
                    xy=(rect.get_x() + rect.get_width() / 2, height),
                    xytext=(0, 3),  # 3 points vertical offset
                    textcoords="offset points",
                    ha='center', va='bottom')

for country in europeanUnionCountries:
    usersTags = []
    followers = []
    #Top 10 most followed TikTok users by country in the database.
    for user in list(collection.find({'userRegion': country}, {'_id': False,'userTag':1,'userStats':1}).sort("userStats.
        usersTags.append(user['userTag'])
        followers.append(user['userStats'][int(numUserStats)]['userFollowers'])
    #bar graph for each country
    bar = plt.bar(usersTags,followers,color='green')
    autolabel(bar)
    plt.title("Top 10 most followed from "+country)
    plt.xlabel("Users")
    plt.ylabel("Followers")
    plt.xticks(rotation=30)
    plt.ylim(0, 20000000)
    plt.show()
```



This is an example of a page where a code with matplotlib produce a graphs with the Top 10 most followed from Spain.

8-BIG DATA

Big data[15] is a field that treats ways to analyse, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data- processing application software.

Our project can be considered a small-scale big data process because we have scraped lot of data to analyse and obtain general conclusions. Obviously, the size of the information handled in big data is much larger than ours, but it follows the same process, the four stages of the information management cycle:

1º Information acquisition

The first step is finding a good source of information and the way to exploit it. One of these methods is **Web Scraping** (a technique that uses software programs to extract information from websites). This is the technology that we have used in our project.

2º Storage

One the information is obtained, it is necessary to store it. Depending on the type of information and its future use, you can choose between different forms of storage, from spreadsheets, such as excel, or NoSQL systems that allow you to store unstructured information in a flexible and fast way. We chose MongoDB, a NoSQL data base system.

3º Management

To process the stored data, there are different types of predictive systems. You can look for repetitive patterns in the data through statistics and machine learning.

[15] https://en.wikipedia.org/wiki/Big_data

In our project, we use a simple system of data representation through tables and graphs from which we draw conclusions.

4º Analysis and conclusion:

Data by themselves do not guarantee knowledge, it is necessary to interpret them and understand the relationships between them. These relationships allow us to extract patterns that build knowledge about multiple domains and fields. In our case, we try to find patterns and relationships in different aspects of European culture in different countries that are part of it.

9-GENERAL CONCLUSIONS

Finally, once the whole project is finished, we can draw two main conclusions:

The first one is about the analysis of the scraped data. We believe that, in general, the predominant theme in Tiktok, in all the countries studied, is very similar, dance and humour, with small exceptions. This may mean that there is a common factor between different cultures about the way we have fun, in other words, each country has a common European culture.

On the other hand, a second conclusion is that we have been able to develop an international project with colleagues from other countries and cultures. This means that we have been able to overcome language barriers, through English, distance and cultural differences, precisely because we can consider the existence of a common European culture.

In addition, and to conclude, we must mention that, during an important part of the project, we have had to suffer the consequences of a pandemic, due to the coronavirus, COVID-19. This has prevented us from meeting physically to finish and present the project, but even so, we have been able to finish it by working together.

10-BIBLIOGRAPHY

- Wikipedia:
  - https://es.wikipedia.org
- Faculty of Statistical Studies of Universidad Complutense de Madrid:
  - https://www.masterbigdataucm.com/que-es-big-data/
- Real Python: Jupyter Notebook: An Introduction:
  - https://realpython.com/jupyter-notebook-introduction/
- Quick tips for using pandas with MongoDB:
  - https://www.moschetti.org/rants/mongopandas.html
- PyMongo tutorial:
  - http://zetcode.com/python/pymongo/
- Medium: Matplotlib Tutorial: Learn basics for Python's powerful Plotting library:
  - https://towardsdatascience.com/matplotlib-tutorial-learn-basics-of-pythons-powerful-plotting-library-b5d1b8f67596
- Stack Overflow: How do I sue matplotlib autopct? :
  - https://stackoverflow.com/questions/6170246/how-do-i-use-matplotlib-autopct
- Matplotlib: Grouped bar chart with labels:
  - https://matplotlib.org/3.2.1/gallery/lines_bars_and_markers/barchart.html
- Mongo DB:
  - https://docs.mongodb.com/
- Youtube: Python 3 Basics Tutorial Series:
  - https://www.youtube.com/playlist?list=PLQVvvaa0QuDe8XSftW-RAxdo6OmaeL85M

- Youtube: Learn Scrapy:
    - https://www.youtube.com/playlist?list=PLZyvi_9gamL-EE3zQJbU5N3nzJcfNeFHU