Departement 2AI

Project : Data collection and storage

SUBJECT : Identify different commercial segments and characterize them from twitter Webscrap.

Github repository

Authors:

Amine MEKKI Ahmed SIDKI Ayoub KASSI

Yunus AYDAR

Date: 10/03/2023

Table of Contents

List of Figures

Abstract

Apple had hired us, a team of data analysts to explore the economic potential of social media marketing through the social network "Twitter", one of the trendiest social networks of our time. We used a design of experiment methodology to study potential commercial segments on Twitter. The results were used to develop targeted marketing campaigns tailored to the interests and preferences of potential customers, with the goal of building brand loyalty and increasing sales.

In order to fullfil this objective, we created our own webcsraping algorithm using Python, Selenium, a MySQL database and a MongoDB database, with which we have collected and stored data of tweets and users' accounts respectively in this order, in order to do a profiling of the average twitter user's interest rate in the brand.

Then, we have proceeded with a short data analysis using basic plots in order to visualise the data distribution according to the selected factors and to identify patterns in user behavior, demographics, and interests.

Key words : Twitter, Web scrapping, commercial segments, profiling

1 Introduction

Apple is a globally recognized brand with a strong reputation for its innovative technology and sleek design. Its success has been built on a strong brand image and a commitment to excellence in all aspects of its business. In order to continue its success, the company has been exploring new marketing opportunities on social media platforms. Social media provides a vast and diverse audience of potential customers, and Apple is interested in understanding the economic potential of social media marketing.

To accomplish this goal, Apple hired four experienced data analysts to conduct a comprehensive study of potential customer segments on social media platforms. The analysts were tasked with exploring Twitter as the first platform for study. Twitter is one of the most popular social media platforms in the world and provides a unique opportunity for Apple to engage with its customers and potential customers.

The team of analysts approached the problem with a design of experiment (D.O.E.) methodology. This methodology is commonly used in research to highlight all the variables, parameters, and constraints of a problem and to methodically organize the collection of necessary data. By using D.O.E., the analysts are able to develop a clear and structured approach to the study.

The first step in the study is to collect data from Twitter. The analysts decided to use web scraping to collect the necessary data. Web scraping is a technique used to extract data from websites, which can then be analyzed and stored in a database management system. The data collected includes information on user profiles, user activity, and user engagement with Apple and its products on Twitter. Once the data is collected, the analysts are going to use statistical analysis techniques to identify potential customer segments. This involved identifying patterns in user behavior, demographics, and interests. By analyzing this data, the analysts are able to create a detailed profile of potential customers who were likely to be interested in Apple's products.

The results of the study will be used to develop targeted marketing campaigns on Twitter. Apple will use the insights gained from the study to create content that is specifically tailored to the interests and preferences of the identified customer segments. This may allow Apple to engage with its customers and potential customers on a more personal level, which may in return help to build brand loyalty and increase sales.

1.1 Problematic

To what extent can webscraping help the Apple company identify and target adequate commercial segments suiting their brand image best ?

2 Design of Experiment

In this part, we will expose the framework of our experiment. This step consists concretely in setting up an experimental design, with the aim of answering constraints of time allocated to the experiment, of availability of the data necessary to answer the problem, of economic constraints, and finally in order to minimize as well as possible the biases which can result from the correlation of the variables.

2.1 DOE approach

Our approach is based on five main milestones:

-The definition of a question, which will become our study hypothesis H0.

-The definition of a study population.

-The choice of a minimum sample size to be scraped.

-Definition of the experimental design architecture.

-Setting up a data collection strategy.

2.2 Constraints to take into account before starting the experiment:

1- Conjunctural constraints -Twitter being a private company managing the data of millions of users, it is subject to various global regulations aimed at guaranteeing the security of user data, and these measures notably make it more difficult to scrape data off the platform, a difficulty that has become more pronounced since the acquisition of the platform by the controversial billionaire Elon Musk. In addition, Twitter did not comply with our request to access their API, making it more difficult (but not impossible) to collect tweets.

2- Availability constraints - It is impossible to access the gender or age of an individual (unless they specify it in their profile, but this is rare) on the platform, just as it remains very difficult to assess the economic status and socio-professional category of the platform's members. The main data provided by the scrapping are data on the geographical location of users and data on their opinion of the Iphone product, based on the detection of certain keywords. There can also be people who put an irrelevant location (ex : Mars, the Earth the World, my place, Home, etc.). These unavailable datas are expected to generate uncertainty that is going to be quantified during the data analysis part of the experiment.

3- Problems of biases - Since Iphone is an expensive product, there will certainly be some hidden socio-economic bias that can alter the "volumes" of the segments identified, since there can be customers who want to buy the product but cannot in practice afford to buy one. Confirmation bias needs also to be taken into account

when selecting the different factors in the study.

2.3 DOE steps

2.3.1 Setting up the study hypothesis.

Let's set up our study hypothesis and its contradiction:

H0: "Twitter users in certain European countries are more interested in purchasing an Iphone".
H1: "There are no particular countries where the demand in Iphone is higher in purchasing an Iphone among Twitter users".

2.3.2 Choice of the study population.

The study population chosen is all users of the Twitter platform worldwide.

2.3.3 Choice of minimum sample.

Taking into account the constraints mentioned above, and in particular the impossibility of accessing the Twitter API, the number of samples taken will be of the order of ten thousand tweets.

2.3.4 Choice of architecture for the experimental design.

The modelling of our experiment will be done using a full factorial design composed of the following factors :
-Number of interested users per region (we define 4 different thresholds)
-different regions of study (all 6 continents)
-both genders (male and female).

We also take into account a "hidden" factor than can be a potential source of heterogenity, which is why we use two blocks in the D.O.E. :
- A block for the "Global North" (Northern part of the world, where consumers' purchase power is higher than the world average).
- A block for the "Global South" (Southern part of the world, where consumers' purchase power is lower than the world average).
Hence, we obtain a 2 blocks - 4X6X2 full factorial design that we chose to design the way it is pictured below
(note that it can be designed in the shape of a 3D- cube with colored regions, but we did not have the tools to do it).

	Male						
	Europe	Africa	Asia	Oceania	North America	Latin America	
Number x of interested users							Block 1 : "Global North"
x <= 100							
100 < x <= 500							
500 < x <= 1000							
x > 1000							
x <= 100							Block 2 : "Global South"
100 < x <= 500							
500 < x <= 1000							
x > 1000							

	Female						
	Europe	Africa	Asia	Oceania	North America	Latin America	
Number x of interested users							Block 1 : "Global North"
x <= 100							
100 < x <= 500							
500 < x <= 1000							
x > 1000							
x <= 100							Block 2 : "Global South"
100 < x <= 500							
500 < x <= 1000							
x > 1000							

2D Architecture of the chosen D.O.E

3 Algorithm explanation

Although Twitter restricts access to its data through its Application Programming Interface (API), and since we couldn't get it, we were able to develop our own web scraping algorithm for this project. While there are libraries available that can help us extract data from Twitter, we wanted to develop our own algorithm to gain a better understanding of the web scraping process. By developing our own algorithm, we were able to customize our data extraction process to meet our specific needs and requirements. This experience gave us a valuable skill that we can apply to other web scraping projects in the future.

Our algorithm scraps the twitter website and identifies different commercial segments. This algorithm enables the user to scrape and collect Tweets and users information.

The algorithm can be used to extract data about a certain topic such as tweets, the username of the person who posted the tweet, the number of likes, retweets, views, and replies, among other things. The data is collected through a search query that the user can customize according to their preference.

The program makes use of the following libraries: BeautifulSoup, Selenium, time, os, csv, dotenv, json, and pandas. It can be executed by creating an object of the TwitterAdvancedSearch class and calling the various methods.

3.1 Algorithm architecture

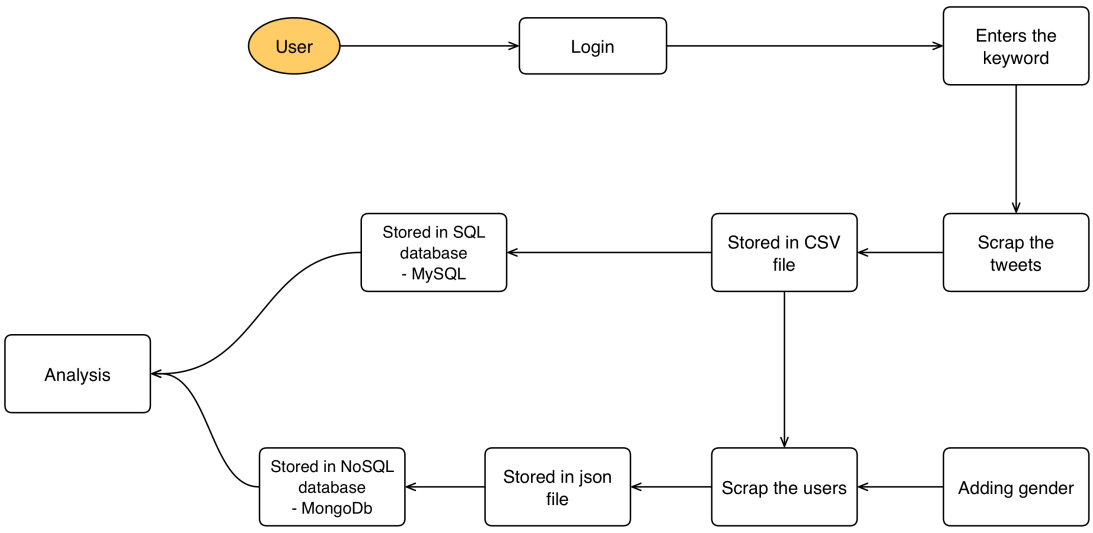

Figure 2: Algorithm general explanation flow chart

The first step is to initialize the class by providing the required search parameters that include words, the exact phrase, none words, hashtags, minimum of replies, minimum of likes, minimum of retweets, from date, and to date.

The second step is to log (login_to_twitter function) into the Twitter account of the user using his credentials as input parameters and using selenium for browser automatization. Once the login is successful, the program calls the enter_search_data() method to enter the search query on Twitter. The search results are then extracted using the collect_all_tweets_articles() method and stores them in a list. Then, the required data is extracted using the get_data_article(), such us : name, username, number of replies, among other parameters that characterize the tweets that are going to be saved as CSV file and then i MySQL database.

Furthermore, using the usernames stored in the CSV file, we are going to scrape each user profile and characterize them using scrap_users() function based on several parameters such as : username, given name of the user, user location, user id, profile description, the number of his follows, friends and tweets. These informations are then stored in json file and a MongoDB database.

As Twitter does not provide the possibility to identify the gender of its users, we have developed an algorithm to infer the gender of the users based on their name. To accomplish this, we mapped a database of names and their associated gender to the name of each user that we scrapped. The algorithm works by identifying the user's name and comparing it to the names in the database. If the name is found in the database, the gender associated with that name is attributed to the user. If the name is not found in the database, the algorithm cannot determine the gender of the user.

It's important to note that this algorithm is not 100% accurate as names can be unisex, and some cultures have different gender associations with specific names. Additionally, some users may not provide their real name on their Twitter profile or may use a pseudonym.

Despite these limitations, the gender identification algorithm provides a valuable insight into the gender distribution of our dataset and allows us to perform gender-based characterization.

To conclude the TwitterAdvancedSearch algorithm that we created is a useful tool for anyone who needs to extract data from Twitter. The program allows the user to customize the search query according to their preference, and extract the required data from the search results page. The program is easy to use and can be executed by anyone who has basic knowledge of Python. However, the program may be limited by Twitter's rate limit, which restricts the number of requests that can be made per hour. Therefore, the program may not be suitable for extracting large amounts of data from Twitter.

3.2 Sequence diagram of our webscrapping algorithm

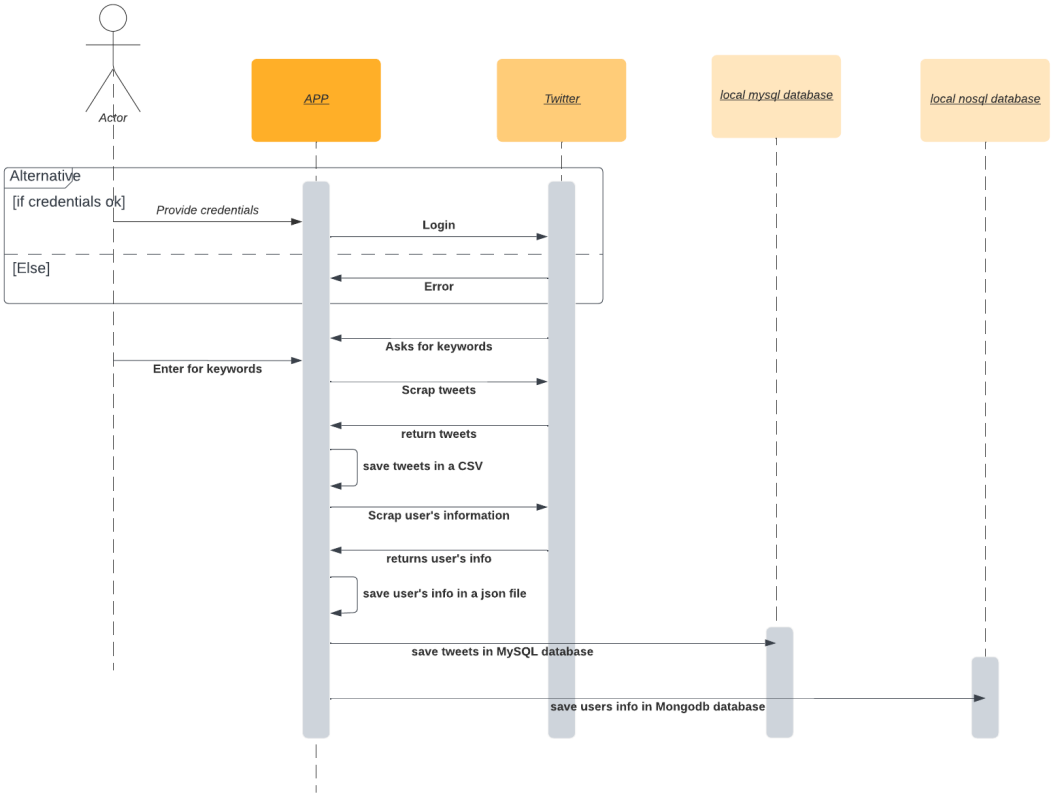

Figure 3: Algorithm explanation flow chart

4 Data Analysis

4.1 Word cloud

For this project, we collected tweets from people who mentioned "I want an iPhone" in their tweets and used a word cloud to visualize the most frequently used words. The purpose of this analysis was to gain insight into the desires and motivations of people who want to purchase an iPhone.

The resulting word cloud is shown below. The size of each word represents its frequency in the dataset.

Figure 4: Tweet's Word cloud.

The word cloud shows that the most frequently used words are "phone", "IPhone", "apple", and "want". This suggests that these people want an iPhone. The prominence of the word "apple" suggests that brand loyalty may also be a factor in the desire to own an iPhone.

Comparison with previous data or surveys could further shed light on how these desires and motivations have changed over time, and how they vary across different demographics.

4.2 Gender estimation

The purpose of implementing the gender identification algorithm was to see which gender can be more interested in buying an iPhone. While the algorithm was able to infer the gender of some users, the significant number of users with an unknown gender highlights the limitations of relying solely on name to identify gender. Further research could explore alternative methods to improve the accuracy of gender identification. Using natural language processing algorithm on each user's tweets maybe give some interesting results.

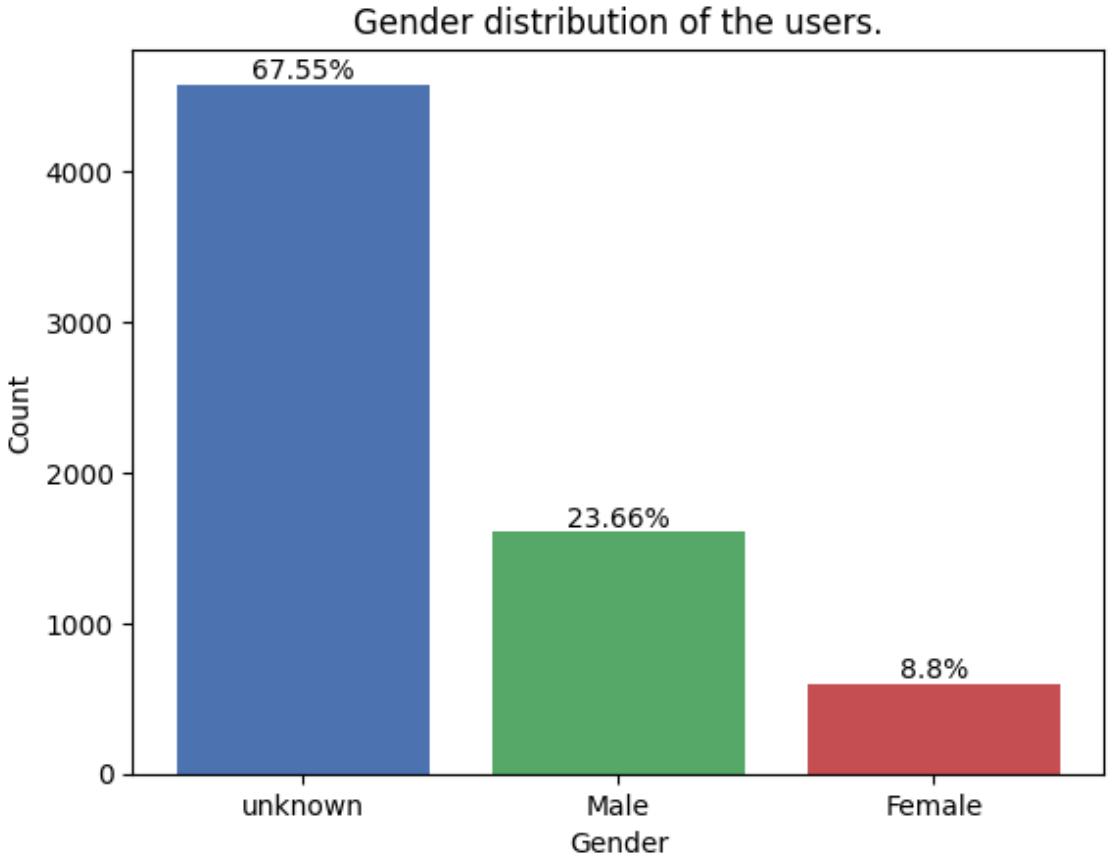

Figure 5: Gender distribution of users.

4.3 Continent extraction

In summary, identifying the location of Twitter users can be challenging due to incomplete or inaccurate information provided by the users. Some users only provide the city or state, while others provide ambiguous information such as "The world". Our algorithm may try to identify the continent based on the location provided, but the accuracy of the result may be uncertain. To improve location identification, natural language processing (NLP) techniques can be used to extract relevant information from the user's profile or tweets and infer their location more accurately. In addition, other sources of information such as language or topics discussed can be considered to refine the estimated location. Overall, a combination of techniques and sources of information is likely to result in the most accurate location estimation on Twitter.

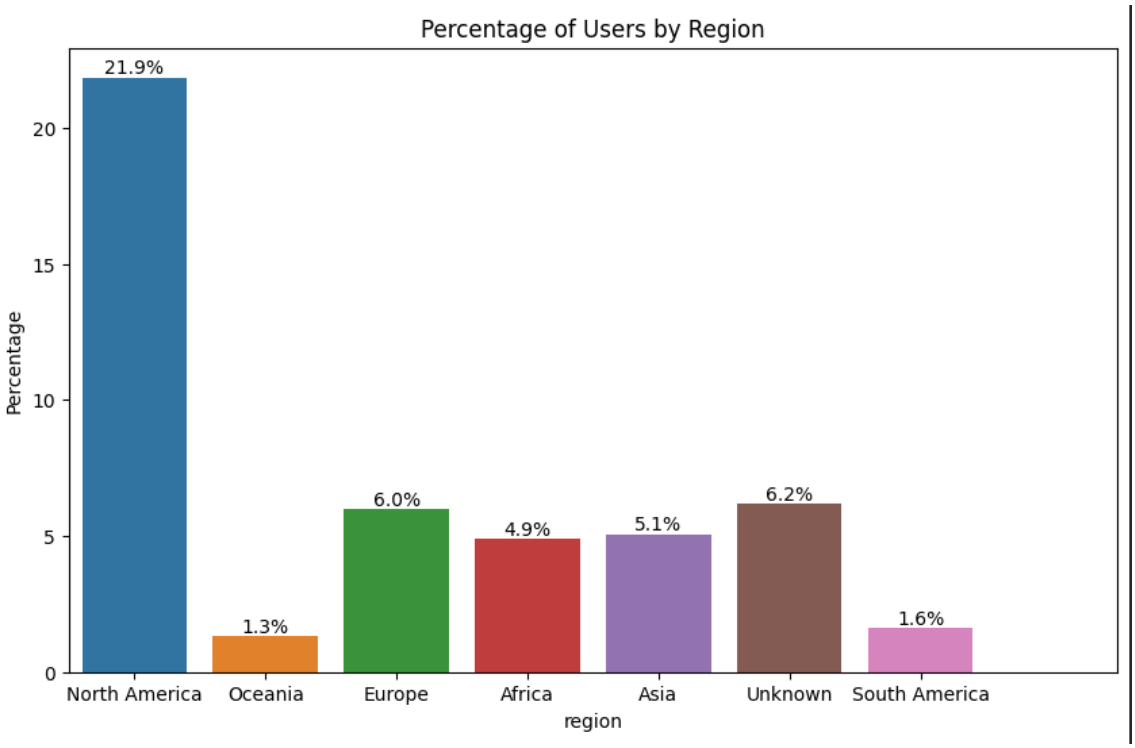

Figure 6: Gender distribution of users.

5 Conclusion

The experiment was successfully conducted and preliminary results could have been extracted from the data collected.

To sum up the strategy used, two types of data have been distinguished : tweets and users' accounts associated to these tweets.
These data have been collected using a Python algorithm with the library BeautifulSoup for the scraping, and Selenium for the automation of the process and the scrolling-down. Keywords such as "buy Iphone" have been mostly detected in tweets that seemed to show a certain interest in the product, which is why we used it as detection parameters when scraping the tweets.
Tweets have been stored in a MySQL database whereas the users' accounts' data has been stored in a MongoDB database (since the latter do not have a repeated schema, but designed according to the user's will, a scalable environment over time is more suitable in this case).
After plotting a few graphs from the data collected, we notice immediately that the uncertainty rate is very high, and is due to the many constraints of unavailable data we had to face during this project, or "random" data that has been put at the place of the real location or the username by the user : for example, we have encountered multiple times location names such as "Mars", or "The Moon", "The World", "Paradise" and so on, as well as anime or superhero-inspired usernames.
Even if we cannot give a sharp conclusion to which of both genders seems more interested in the brand's product because of strong uncertainty, men seem to be more interested in the matter (it needs to be confirmed by more reliable experiments though). Finally, the users' geographic distribution all around the world seems to show non-uniform patterns, with North America coming first by large advance compared to other countries. However, only latin alphabet has been taken into account during the experiment, which may result in a strong bias for western countries. Let us also not forget that Twitter, among numerous social networks, is periodically restrained or canceled in many countries in Asia (Russia, China, India) and in Africa, which may be an explanation of the obtained result.

Hence, the analysis needs to be improved, perhaps by implementing NLP-based deep-learning algorithm, that would be able to notice more accurately the signs of interest in tweets for the Iphone product, and even identify the gender of the user related to that tweet.