A blue and white logo

Description automatically generated with low confidence

Ecole Supérieure Privée d'Ingénierie

Et de Technologies

**Engineer Summer internship**

**Subject :**

**Social Network Sentiment Analysis : Esprit**

Developed by **Samar Jberi**

Supervised by **Mohammed Tlili**

August 6th to September 30th, 2021

Academic Year: 2021/2022

**Summary**

General Introduction……………………………………………………………………………………………………………………………….5

I General project context………………………………………..………………………………………………………………………6

1. Presentation of the host organization……………………………..………………………………………………………….6

2. Project context……………………………………………………………………………..……………………………………..…….6

2.1. Study of the existing…………………..…………………………………………………….……………….………….…….....…6

2.2. Problematic………………………………………………………………………………………………….……………...…………....7

2.3. Proposed solutions………………...…………………………………………………….……………………………………...……7

2.4. Adopted methodology - CRISP-DM……..…………………………………………….…………………………….………….7

3 Implementation……………………………………………………………………………………….……………………………...…9

3.1. Used tools …………………………………………………………………………………………………….…………………………...9

3.1.1. Microsoft Power BI…………………………………………………………………………………….……………………………….9

3.1.2. Python…………………………………………………………………………………………………………………………………….....9

4. Data Warehouse ……………………………………………………………………….……………………………………………….9

4.1. Data Warehouse modeling approach..............................................................................................9

II. Data Scraping...............................................................................................................................10

1. Data integration using Power BI………………………………………………………………………………………………....11

1.2. Data warehouse model.................................................................................................................16

III. Text Mining……………………………………………………………………………………………………………………………..18

IV. Data visualization …..………………………………………………………………………………………………………………..23

V. Conclusion……………….……………………………………………………………………………………………………………....26

**Table of figures**

Figure 1: Esprit Logo……………………………………………………………………………………….6

Figure 2: CRISP-DM Methodology………………………………………………………………………...8

Figure 3: Bottom-Up approach by Ralph Kimball………………………………………………………..10

Figure 4: Connecting to the web Page…………………………………………………………………….11

Figure 5: Data picking and storing in csv file……………………………………………………………..11

Figure 6: Calendar Table ………………………………………………………………………………....12

Figure 7: Posts\_Fact\_table..………………………………………………………………………….……13

Figure 8: Revieweres LookUp…………………………………………………………………………….13

Figure 9: Skills LookUp\_Table……………………………………………………………………………14

Figure 10: What They\_Do LookUp table…………………………………………………………………14

Figure 11: Where\_They\_live lookUp Table……………………………………………………………....15

Figure 12: Where\_They\_Studied LookUp Table………………………………………………………….15

Figure 13: LinkedIn Warehouse…………………………………………………………………………..16

Figure 14: Facebook Warehouse ………………………………………………………………………….17

Figure15: Cleaning data …………………………………………………………………………………..18

Figure 16: Study of frequency…….………………………………………………………………………19

Figure 17: Frequency plot of single terms…………………………………………………………..…....19

Figure 19: Frequency Plot of coupled term……………………………………………………………….20

Figure 18: Content filtering………………………………………………………………………………..20

Figure 20: Word Stemming…………………………………………………….……………………….....21

Figure 21: Words\_Clouds ………………………………………………………………………………...21

Figure 22: Bar\_Plot………………………………………………………………………………………..22

Figure 23: Data\_training…………………………………………………………………………………..22

Figure 24: Facebook\_dashboard\_1………………………………………………………………………..23

Figure25: facebook\_Dashboard\_2………………………………………………………………………...23

Figure 26:Skills………………………………...………………………………………………………….24

Figure 27:Where\_They\_Live……………………………………………………………………………...24

Figure 28:Where\_They\_Studied…………………………………………………………………………..25

Figure 29:What\_They\_Do………………………………………………………………………………...25

**General Introduction**

In order to apply the different skills we have acquired during the past couple of years, students conduct a summer internship after their second year of engineering studies.

Supported by the Private Higher School of Engineering and Technologies Esprit, I was offered the opportunity to work on the subject entitled "Social Network analysis" that took place from 06/08/2020 to 30/09/2021, under the supervision and help of Mr. Mohammed Tlili.

During this internship I had the chance to discover new technologies and acquire new skills.

Dived into the data and created a neat deliverable out of it.

In this report, I will list the different tasks performed as well as the steps I followed to develop my project.

In the first chapter, I will start by a presentation of the subject, study of the existing, and proposed solutions.

The second chapter will deal with the text mining part of my project, while the third chapter will deal with data visualization.

Finally, a conclusion will summarize the work done and the results obtained.

1. **General project context:**

In this chapter I will start by presenting the host organization, followed by a deep analysis of the project context including a study of the existing, the problematic at hand and the proposed solution. Lastly, I will specify the adopted methodology.

1. **Presentation of the host organization:**

A blue and white logo

Description automatically generated with low confidence

Figure 1: Esprit Logo

Founded in 2003, Esprit has managed to become an internationally recognized establishment.

In addition to being accredited by the Tunisian Ministry of Higher Education and Scientific Research, the quality of Esprit education has become well renown over the years.

Following the lead of bigger higher education facilities such as MIT, Esprit builds the pedagogy of engineering training around professional scenarios. Esprit aims towards excellence.

And for this exact reason Esprit cares deeply about its public image, collecting and taking into consideration all the feedback in order to continuously improve its programs and keep up with the expectations to offer the best college experience possible for its students and employees.

1. **Project context:**

**2.1 Study of the existing:**

Nowadays, people tend to share so much of themselves and their feelings on internet and social networks.

This usually leads to communicate large amount of data.

these data can be very useful if it was correctly analyzed.

And since Esprit is one of the biggest private schools, there are lots of unexploited data that can make a huge difference in the University growth.

**2.2 Problematic:**

The educational system has developed to include198 public higher education institution versus **63** private institutions according to a report conducted by the British council called "Education in North Africa" in 2014.

the numbers of students enrolled in higher education has experienced a significant boost from over **2.6%** in 1974 to **35.2%** by 2015.

With such tight competition within the private sector, each private institute aims to step up in order to maintain its success and notoriety

The question asked, how can these data help in the growth of Esprit against other competitors in the private sector?

**2.3 Proposed solutions:**

Based on identified needs and my knowledge about the subject since I’m an Esprit student myself, I was assigned the project “Social Network analysis Esprit” in pursuance of finding solutions for the issues mentioned in the problematic.

Since every private facility relies heavily on its public image and, I realized that it is an essential starting point to consult the public reviews online about Esprit, study its foundation and followers and evaluate the interactions that take place on the social networks in order to identify and solve the most frequent problems.

To serve this purpose, I had to go through several steps:

• Scrapping the most relevant information (followers, employees, comments, interactions…) about Esprit on social networks (Facebook, LinkedIn)

• Shaping and cleaning the data, establishing relationships between the different data.

• Text Mining to detect the nature of the content and the most frequent searched terms.

• Using Data Visualization tools to classify posts based on different criteria to make the data more legible and have a general grasp on the image of Esprit.

**2. 4. Adopted methodology - CRISP-DM**

on this project I used CRISP-DM methodology which stands for cross-industry process for data mining.

it provides a structured approach to planning a data mining project.

it's composed essentially of 6 stages:

1. Business understanding

2. Data understanding

3. Data preparation

4. Modeling

5. Evaluation

6. Deployment

Diagram

Description automatically generated

Figure 2: CRISP-DM Methodology

**Conclusion**

In this chapter, I got to explain the project’s context starting from the study of the existing to the proposed solution. Then proceeded to present my methodology. In the next chapter, I will be demonstrating the implementation of my data.

**3. Implementation:**

In this chapter I’m going to give a detailed description of the practical work I did. I’m going to start by listing the different tools I used, followed by the data model, screenshots of the integration work on power BI, screenshots of the Text Mining and Data Visualization work.

**3.1. Used tools**

**3.1.1.Microsoft Power BI**

Power BI is a business analytics service by Microsoft.

it comes with interactive visuals and business intelligence capabilities.

it is known for being super user friendly. Power BI provides cloud-based business intelligence

services, known as "Power BI Services", along with a desktop-based interface, called "Power BI

Desktop".

**3.1.2.Python**

Python is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

**4.Data Warehouse**

**4.1.Data Warehouse modeling approach**

Designing a [Data Warehouse](https://www.geeksforgeeks.org/data-warehouse-architecture/) is an essential part of business development. In this segment, I’m going to take an in depth look at the data warehouse modeling approach that I chose:

**Bottom-Up approach by Ralph Kimball:**

I used Ralph Kimball approach because it matched my needs. Since it is a bottom-up approach. i started by creating data marts going to my final data warehouse which made the ideal choice.

Shape

Description automatically generated

Figure 3: Bottom-Up approach by Ralph Kimball

**II. Data Scraping:**

In order to gather maximum of information about Esprit online, I did some web scraping using python, relying essentially on two important libraries: selenium and Beautifulsoup.

Beautiful Soup is a Python HTML and XML document parsing library created by Leonard Richardson. It produces a syntax tree which can be used to search for elements or modify them.

Selenium is a powerful tool for controlling web browsers through programs and performing browser automation. It is functional for all browsers, works on all major OS and its scripts are written in various languages i.e. Python, Java, C#.

l will be working with Python.

In this first part I started by importing the web driver and connected to the web page then fetched the XML content using beautifulSoup.

Text

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

Figure 5: Data picking and storing in csv file

Figure 4: Connecting to the web Page

1. **Data integration using power BI**

Data integration is the process of combining data from different sources into a single, unified view. Integration begins with the ingestion process and includes steps such as cleansing and transformation. Data integration ultimately enables analytics tools to produce effective, actionable business intelligence.

The scraped data of the chapter before lacks structure and meaning , that’s when Power Bi comes in handy. Applying the different necessary tools lead to the creation of the tables figuring in the in the following screenshots.

Table

Description automatically generated with medium confidenceThis figure demonstrates a detailed calendar of the posts updates which make it very flexible and useful with the data visualization.

Figure 6: Calender Table

This figure indicates posts information, starting with the comments, number of likes, number of reacting comments, type of comment (whether positive or negative), along with the foreign keys of the lookup tables in which our Data table is associated.

Table

Description automatically generated

Figure 7: Posts\_Fact\_table

Table

Description automatically generatedThis figure shows the information about the Reviewers members, starting by their full names, recommendation (whether they recommend esprit or not), their gender and their nationality.

Figure 8: Revieweres LookUp

This figure demonstrates the number of esprit employees matched to their skill according to LinkedIn

Table

Description automatically generated

Figure 9: skills LookUp\_Table

This figure demonstrates the number of Esprit employees matched to their occupation according to LinkedIn.

Table

Description automatically generated

Figure 10: What They\_Do LookUp table

This figure demonstrates the number of Esprit employees matched to where they live according to LinkedIn.

Table

Description automatically generated

Figure 11: Where\_They\_live lookUp Table

This figure demonstrates the number of Esprit employees matched to where they studied according to LinkedIn.

Graphical user interface, table

Description automatically generated

Figure 12: Where\_They\_Studied LookUp Table

* 1. **Data warehouse model:**

On this part, I had to figure out how link the tables together in way it can serve my purpose.

Since Data was gathered from two completely different sources (LinkedIn and Facebook) and represents two different ideas. I had to create two different Data warehouses following the star model.

Diagram

Description automatically generated

Figure 13: LinkedIn Warehouse

Diagram

Description automatically generated

Figure 14: Facebook Warehouse

**III.Text Mining**

Text mining, also referred to as text data mining, like text analysis, is the process of deriving high-quality information from text.

Within the following, I’ll be demonstrating screenshots about my work of text Mining of esprit reviewers Facebook page.

At this first stage, I started by fetching the Data by scrapping them out of the reviewers’ page on Facebook.

The data fetched might contain irrelevant data, and this may parasitize with the data we want to work on. That’s why I cleaned the Data out of stop words and punctuations and applied the different methods such as stemming and Lemmatization.

Graphical user interface, text, application, email

Description automatically generatedAlso, I studied words frequency and highlighted them to have a clear vision about the most significant terms related to Esprit.

Figure 15: Cleaning data

Graphical user interface, application

Description automatically generated

Figure 16: Studying of frequency

Chart, bar chart

Description automatically generated

Figure 17: frequency plot of single terms

Chart

Description automatically generated

Figure 18: Frequency Plot of coupled term

Graphical user interface, text, application

Description automatically generated

Figure19: Content filtering

Graphical user interface, text, chat or text message

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

Figure 21: Words\_Clouds

Figure 20: Word Stemming

Chart

Description automatically generated with medium confidence

Figure 22: Bar\_Plot

Graphical user interface, text, application, email

Description automatically generatedAs a final step of this chapter, I applied a supervised learning approach of classification called Naive Bayes to figure out the accuracy of my data since it is popular and used for sentiment classification.

Figure 23: Data\_training

**IV.Data Visualization with Power BI:**

Graphical user interface

Description automatically generatedGraphical user interface, chart, application

Description automatically generatedOn this part, I will be demonstrating my final dashboard including the different reports for a neat and clear readability of my data and the relation between its different elements.

Figure24: Facebook\_dashboard\_1

Figure25: facebook\_Dashboard\_2

Graphical user interface

Description automatically generated with low confidenceChart, treemap chart

Description automatically generated

Figure 27:Where\_They\_Live

Figure 26:Skills

Graphical user interface, application

Description automatically generated

Figure28:Where\_They\_Studied

Graphical user interface

Description automatically generated with low confidence

Figure 29:What\_They\_Do

**V.conclusion :**

Since data is massively growing over the years, learning the adequate tools, applying good analysis and visualizing data in a the best readable way is a shortcut towards successful entrepreneurial planning.

During this project, I covered the data scraping, text Mining going to data visualization. Power Bi and Python are two powerful tools that led into creating a dashboard so user friendly, for the optimum results and the best decision making that shall be affected by Esprit.

Using this project, Esprit can have a better idea about its public image, note its weaknesses and apply the necessary changes in order to remain the best destination for lots of students in Tunisia and otherwise.