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 Sentimental Analyses : 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……..…………………………………………….…………………………….………….8

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...............................................................................................................................11

1. Data integration using Power BI………………………………………………………………………………………………....12

1.2. Data warehouse model.................................................................................................................17

III. Text Mining……………………………………………………………………………………………………………………………..20

IV. Data visualisation …..………………………………………………………………………………………………………………..27

V. Conclusion……………….……………………………………………………………………………………………………………....30

**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……………………………………………………………..12

Figure 6: Calendar Table ………………………………………………………………………………....13

Figure 7: Posts\_Fact\_table..………………………………………………………………………….……14

Figure 8: Revieweres LookUp…………………………………………………………………………….15

Figure 9: Skills LookUp\_Table……………………………………………………………………………16

Figure 10: What They\_Do LookUp table…………………………………………………………………16

Figure 11: Where\_They\_live lookUp Table……………………………………………………………....17

Figure 12: Where\_They\_Studied LookUp Table………………………………………………………….17

Figure 13: LinkedIn Warehouse…………………………………………………………………………..18

Figure 14: Facebook Warehouse ………………………………………………………………………….19

Figure15: Cleaning data …………………………………………………………………………………..20

Figure 16: Study of frequency…….………………………………………………………………………21

Figure 17: Frequency plot of single terms……………………………………………………………....21

Figure 18: Content filtering………………………………………………………………………………..22

Figure 19: Frequency Plot of coupled term……………………………………………………………….22

Figure 20: Word Lemmatisation ……………………………………………………………………….....23

Figure 21: Words\_Clouds ………………………………………………………………………………...23

Figure 22: Bar\_Plot………………………………………………………………………………………..24

Figure 23: Number of Likes basing on gender……………………………………………………………25

Figure 24: General aspect………………………………………………………………………………....25

Figure 25: Data\_training…………………………………………………………………………………..26

Figure 26: Facebook\_dashboard\_1………………………………………………………………………..27

Figure27: facebook\_Dashboard\_2………………………………………………………………………...27

Figure 28:Skills………………………………...………………………………………………………….28

Figure 29:Where\_They\_Live……………………………………………………………………………...28

Figure 30:Where\_They\_Studied…………………………………………………………………………..29

Figure 31:What\_They\_Do………………………………………………………………………………...29

**General Introduction**

*In order to apply the different skills we have aquired 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.*

**I. 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 : 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*.

**2 Project context:**

1. **1 .Study of the existing:**

*Esprit is an engineering school, ~~that's all~~, and it trains students according to the most demanding international ~~quality~~ standards. Esprit brings together a ~~very~~ broad number of skills in the sector: more than 60 experts, including the main aspects in public engineering studies reforms, around 100 permanent teachers. All in the service of a unique but very essential objective: to train excellent engineers ready to take over the IT market.*

*From around 30 in 2003 to over 2,400 in 2010-2011 (day and evening schooling), the number of students has not ceased to increase, with an average annual growth rate of over* ***85%****.*

**2. 2.Problematic:**

*According to a 2014 British Council report titled “Education in North Africa”, since then the system has developed to include 198 public higher education institutions, 63 private institutions, 24 higher institutes of technical studies and six higher institutes for teacher training. The expansion of the higher education system led a boom from 17,000 students in 1975 to half a million students in 2015(equivalent to* ***8%*** *of the country’s student population), which represents an increase in the enrolment rate at the tertiary level from* ***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.

**I don’t see any problem here: you should formulate a question that your project answers**

**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, we 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**

CRISP-DM stands for cross-industry process for data mining. The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology. CRISP-DM is the most widely used chain for the development of Data Mining projects. It is composed essentially of 6 Stages:

1. [Business understanding](https://www.sv-europe.com/crisp-dm-methodology/#one)

2. [Data understanding](https://www.sv-europe.com/crisp-dm-methodology/#two)

3. [Data preparation](https://www.sv-europe.com/crisp-dm-methodology/#three)

4. [Modeling](https://www.sv-europe.com/crisp-dm-methodology/#four)

5. [Evaluation](https://www.sv-europe.com/crisp-dm-methodology/#five)

6. [Deployment](https://www.sv-europe.com/crisp-dm-methodology/#six)

Diagram

Description automatically generated

Figure : 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.

1. **Implemantation:**

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 aims to provide interactive [visualizations](https://en.wikipedia.org/wiki/Data_visualization) and [business intelligence](https://en.wikipedia.org/wiki/Business_intelligence) capabilities with an interface simple enough for end users to create their own reports and dashboards. It is part of the [Microsoft Power Platform](https://en.wikipedia.org/wiki/Microsoft_Power_Platform). Power BI provides [cloud](https://en.wikipedia.org/wiki/Cloud_computing)-based business intelligence services, known as "Power BI Services", along with a desktop-based interface, called "Power BI Desktop". It offers [data warehouse](https://en.wikipedia.org/wiki/Data_warehouse) capabilities including [data preparation](https://en.wikipedia.org/wiki/Data_preparation), [data discovery](https://en.wikipedia.org/wiki/Data_discovery) and interactive dashboards.

**3.1.2.Python**

Python is an [interpreted](https://en.wikipedia.org/wiki/Interpreted_language) [high-level](https://en.wikipedia.org/wiki/High-level_programming_language) [general-purpose programming language](https://en.wikipedia.org/wiki/General-purpose_programming_language). Python's design philosophy emphasizes [code readability](https://en.wikipedia.org/wiki/Code_readability) with its notable use of [significant indentation](https://en.wikipedia.org/wiki/Off-side_rule). Its [language constructs](https://en.wikipedia.org/wiki/Language_construct) as well as its [object-oriented](https://en.wikipedia.org/wiki/Object-oriented_programming) approach aim to help [programmers](https://en.wikipedia.org/wiki/Programmers) write clear, logical code for small and large-scale projects. Python is [dynamically-typed](https://en.wikipedia.org/wiki/Dynamic_programming_language) and [garbage-collected](https://en.wikipedia.org/wiki/Garbage_collection_(computer_science)). It supports multiple programming paradigms, including [structured](https://en.wikipedia.org/wiki/Structured_programming) (particularly, [procedural](https://en.wikipedia.org/wiki/Procedural_programming)), [object-oriented](https://en.wikipedia.org/wiki/Object-oriented_programming) and [functional programming](https://en.wikipedia.org/wiki/Functional_programming). Python is often described as a "batteries included" language due to its comprehensive [standard library](https://en.wikipedia.org/wiki/Standard_library).

**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:**

Contrast to Bill Inmon approach, Ralph Kimball recommends building the data warehouse that follows the bottom-up approach. In Kimball’s philosophy, it first starts with mission-critical data marts that serve analytic needs of departments. Then it is integrating these data marts for data consistency through a so-called information bus. Kimball makes uses of the dimensional model to address the needs of departments in various areas within the enterprise.

Shape

Description automatically generated

Figure : Bottom-Up approach by Ralph Kimball

For my data model approach, I chose the Bottom-Up approach by **Ralph Kimball**. My project’s specific objectives required me to define the data Marts and then merge them together. Moreover, our data model takes the shape of the Star schema, which is used for the Bottom-Up approach.

**II. Data Scrapping:**

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](https://www.geeksforgeeks.org/python-programming-language/), [Java](https://www.geeksforgeeks.org/java/), [C#](https://www.geeksforgeeks.org/csharp-programming-language/).

l will be working with Python.

Graphical user interface, text, application, email

Description automatically generated

Figure : Connecting to the web Page

Figure 5: Data picking and storing in csv file

Text

Description automatically generated

**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.

Data integration architects develop data integration software programs and data integration platforms that facilitate an automated data integration process for connecting and routing data from source systems to target systems. An example of those software is Power BI.

In the following screenshots, I will demonstrate my work in how to clean the data (the data scrapped earlier) and loading it into my data warehouse.

Table

Description automatically generated

Figure : Calender Table

This figure demonstrates a detailed calendar of the posts updates which make it very flexible and useful with the data visualization.

Table

Description automatically generated

Figure : Posts\_Fact\_table

This figure indicates posts informations, starting with the comments, number of likes, number of reacting comments, type of comment (wether postive or negative), along with the foreign keys of the lookUp tables in which our Data table is associated.

Table

Description automatically generated

Figure : Revieweres LookUp

This 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.

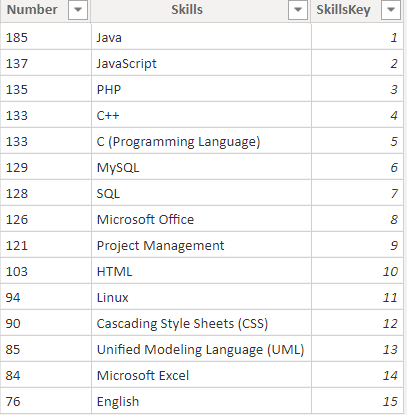


Figure : skills LookUp\_Table

Table

Description automatically generatedThis figure demonstrates the number of esprit employees matched to their skill according to LinkedIn.

Figure : What They\_Do LookUp table

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

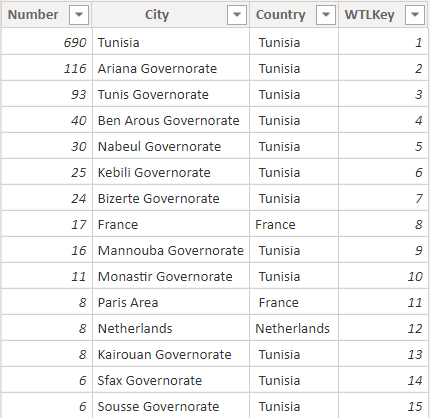


Figure : Where\_They\_live lookUp Table

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

Graphical user interface, table

Description automatically generated

Figure : Where\_They\_Studied LookUp Table

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

**2. Data warehouse model:**

Since Data was gathered from two completely different sources (LinkedIn and Facebook) I had to create two different Data warehouses*.*

Diagram

Description automatically generated

Figure : LinkedIn Warehouse

Diagram

Description automatically generated

Figure : Facebook Warehouse

**III.Text Mining**

Text mining, also referred to as text data mining, similar to text analysis, is the process of deriving high-quality [information](https://en.wikipedia.org/wiki/Information) from [text](https://en.wikipedia.org/wiki/Plain_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 strated by fetching the Data by scrapping them out of the reviewers’ page on Facebook.

After fetching my data, I got down to the text Mining phase. Starting by implementing the necessary liberaries, cleaning the Data out of stop words and punctuations, applying the different methods such as stemming and Lemmatization. And studiying words frequency and highlighting them in order to have a clear vision about the most significant terms related to Esprit.

Figure : Cleaning data



Figure 16: Studying of frequency

Chart, bar chart

Description automatically generated

Figure : frequency plot of single terms

Text

Description automatically generated

Figure : Content filtering

Chart

Description automatically generated

Figure : Frequency Plot of coupled term

Graphical user interface, text, application, email

Description automatically generated

Figure : Word Lemmatisation

Graphical user interface, text, chat or text message

Description automatically generated

Figure : Words\_Clouds

Chart

Description automatically generated with medium confidence

Figure : Bar\_Plot

In this phase, I’m going to take a close look on the mass of data I gathered.

Graphical user interface, application

Description automatically generatedA picture containing graphical user interface

Description automatically generated

Figure : Number of Likes basing on gender

Figure : General aspect

Here I will be closuring my sentimental analysis by 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.

Graphical user interface, text, application, email

Description automatically generated

Figure : Data\_training

**IV.Data Visualization with Power BI**

Graphical user interface

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.

Graphical user interface, chart, application

Description automatically generated

Figure26: Facebook\_dashboard\_1

Figure27: facebook\_Dashboard\_2

Graphical user interface

Description automatically generated with low confidenceChart, treemap chart

Description automatically generated

Figure :Where\_They\_Live

Figure :Skills

Graphical user interface, application

Description automatically generated

Figure :Where\_They\_Studied

Graphical user interface

Description automatically generated with low confidence

Figure :What\_They\_Do

**V.conclusion :**

Since data is massively growing over the years, learning the adequate tools to handle and manage it in the best ways is a shortcut towards successful entrepreneurial planning.

During this project, I have been through lots of phases from data scraping to text Mining going to data visualization. Power Bi and Python are two powerful tools that led into creating a dashboard so user friendly, ~~so~~ readable for the optimum results and the best decision making.