Brain Drain reasons analysis

Text Mining in Sociology

Lobna FEZAI & Mouna BELAID

TEXT MINING COURSE, NATIONAL SCHOOL OF ENGINEERING TUNIS

This report was done under the supervision of Dr.Chiraz Ben Abdelkader in purpose to analyze the reasons behind the willing of the Tunisian students to immigrate abroad.

November 2018

Contents

1. Introduction

1.1 Motivation

The motivation behind this work is simply the answer we are nowadays hearing whenever the future plans question is mentioned to us the students or the graduated ones. A good future plan term is henceforth tightly linked to working or living abroad. Therefore, the phenomenon of *Brain Drain* starts to be one of the main issues we are facing. An emergent analysis is needed to highlight the reasons behind this trend which can be really dangerous to our country specially in this critical growing period in order to treat it.

> It's creating a brain drain. We would end up with a society without knowledge. How can such a society make progress?

1.2 Objective

The idea was to create a data base from the answers of a prepared survey. The overall goal of this project is to understand and summarize the questionnaire answers; and eventually to extract major trends.

Our project is about Brain Drain in Tunisia. In this regard, a survey was conducted among Tunisian people in order to get answers of several questions about this topic. As a team, we are mainly interested in making analysis about two basic questions among all the questions being answered :

- What are the reasons that would push you to leave Tunisia ?
- Socially, what's the difference between Tunisia and abroad in social life ?

1.3 A bit of context

1.3.1 Definition

Brain drain is the sociological phenomenon resulting a lost of educated and talented students and workers moving to developed countries searching for different or better conditions.

Definition Brain drain can also be defined as the loss of the academic and technological labor force through the moving of human capital to more favorable geographic, economic, or professional environments. More often than not, the movement occurs from developing countries to developed countries or areas. [**1**]

1.3.2 Causes of Brain Drain

Multiple issues might be the causes of the process of brain drain relative to the country experiencing it. The main causes include unemployment or seeking higher paying jobs, poor work or studies conditions, lack of freedom of creativity and innovation, difficult economic state of the country, political instability, and to search for a better quality of life, modern lifestyle, stable economic and political environment, better education system.

1.3.3 Effects of Brain Drain on the Home Country

Brain drain can have real negative impact on a developing country including the loss of talented, qualified students and workers which will lead to worse situation in all the areas where we need this potential people such as economics, politics, science... This phenomenon is somehow auto-causing effect since more it's occurring, more the situation is getting worse, more people will desire to leave rather than stay.

Brain drain is usually described as a problem that needs to be solved. However, there are benefits that can be derived from the phenomena. When people move from LDC countries to developed countries, they learn new skills and expertise, which they can utilize to the advantage of the home economy once they return. Another benefit is remittances; the migrants send the money they earn back to the home country, which can help to stimulate the home country's economy. [**1**]

2. The approach of processing the survey

2.1 Text mining approach introduction

Data Mining is the process of extracting/discovering interesting high-quality knowledge from large volumes of data, with the help of computers. The Data Mining process is used by several kinds of people in different contexts like scientific researches, traditional business, web and even government. In our project context, we are mainly interested in the process of text mining which is the process of data mining with text data. That means the automated processing of text data we have in order to extract patterns from text written/expressed in any human language, such as Arabic, French, Chinese, etc.

Our text mining project pipeline would be as follow :

1. Read questionnaire data from a csv file
2. Initialize raw corpus (for example, select the question we hope to analyze)
3. Text cleaning
4. Text representation
5. Topic modeling
6. Clustering
7. Predictive modeling
8. Visualization and interpretation of final results

For text mining there are a bunch of software available, Along our project, we resort technically to the language programming Python which is widely used in general, and for text mining in particular. In fact, we use Jupyter which is one of the popular IDEs (Integrated Development Environment) for Python.

Through this part of report, we will expose our approach used.

For running the code used and displaying the results of every step, this link can be useful.
`https://drive.google.com/drive/folders/1qzpbUtbpl_OvabMlAorcwElBzLi2lSyW?usp=`

`sharing`

2.2 Understanding our dataset

In this step, we first read questionnaire data from a csv file. Data are collected through a survey conducted in Tunisia. There are variables that express social information about the interviewees like Age, Gender, Civil Status, Level of Studies, Establishment of Studies, Professional Situation, Region and Field of Studies. Beside these information, we find:
Timestamp
Q1 : Qu'est-ce qui vous ferait rester en Tunisie ?
Q2 : Quel salaire vous fera rester en Tunisie ?
Q3 : Qu'est-ce qui encouragerait les étrangers à venir en Tunisie ?
Q4 : Qu'est-ce qui vous ferait partir à l'étranger ?
Q5 : Quelle est la différence au point de vue social entre la Tunisie et l'étranger ?
Q6 : Quelle est la différence au point de vue professionnel entre la Tunisie et l'étranger ?

We got finally a dataset composed as below :

	Etat civil	Niveau d'étude	Etablissement d'études	Situation professionnelle	Région	Domaine d'études	de la personne qui vous a envoyé ce formulaire ? (Cela doit etre un nombre entre 1 et 52. Ecrire "autre" si vous ne l'avez pas)	Q1 : Qu'est-ce qui vous ferait rester en Tunisie ?	Q2 : Quel salaire vous fera rester en Tunisie ?	Q3 : Qu'est-ce qui encouragerait les étrangers à venir en Tunisie ?	Q4 : Qu'est-ce qui vous ferait partir à l'étranger ?	Q5 : Quelle est la différence au point de vue social entre la Tunisie et l'étranger ?	Q6 : Quelle est la différence au point de vue professionnel entre la Tunisie et l'étranger ?
	Célibataire	Ingénieur	INSAT	Etudiant	Grand Tunis	Sciences de l'Ingénieur	student's order in the official list of TICV s...	mes études, ma famille et mon entourage	2000	le tourisme	la volonté de faire de l'expérience et de fair...	A l'étranger on trouve plus de respect.	A l'étranger ils payent plus et ils sont plus ...
	Célibataire	Ingénieur	ENIT	Etudiant	Grand Tunis	Sciences de l'Ingénieur	Autre	Rien	2000	Réputation	Niveau de conscience populaire	La valeur d homme	Recherche
	Célibataire	Mastère	ESAD	Fonctionnaire	Grand Tunis	Artistique	22	La mentalité	2000 dinars	la sécurité	l'argent	la tolérance	la motivation
	Marié	Licence	FST	Fonctionnaire	Grand Tunis	Sciences de l'Ingénieur	22	Qualité de vie	3500	Pour un investisseur, la rémunération des empl...	L'expérience + le salaire	La mentalité sur le niveau professionnel et so...	À l'étranger, les expériences sont plus intere

Figure 2.1: Dataset of our study

The dataset is composed of 258 rows and 16 variables. We were mainly interested in only two questions which the analysis would be about; the 4th and 5th questions.

- What are the reasons that would push you to leave Tunisia ?
- Socially, what's the difference between Tunisia and abroad in social life ?

2.3 Data processing

2.3.1 Corpus preparation

First, let's prepare corpus for analyzing both questions. We select the *raw* data we are going to process. The important things are to learn how to input the data correctly, establish the methods to

treat it and find the best way to visualize the results and interpret them correctly. First of all, we prepared the corpus of every question. A corpus is a list of documents. In this context, a document refers to a row in our dataset. Therefore, we got 258 documents in every corpus built. A lot of operations can be done, such as visualizing the distribution of characters in every corpus.

Samples represent characters in the corpus related to the question. A plot shows the distribution of characters according to their cumulative frequencies. We note that the total number of characters is 12298 in the first question corpus and 14270 in the corpus related to the second question. We also note that the top 20 characters cover more than 95% of all character occurrences in the corpus.
As we can notice for example that there are characters like this symbol " ' " that should be removed. We will do such a task later in the step of cleaning.

2.3.2 Data Cleaning

This section is where you prepare your data to be processed, you have to make sure that you have the less possible quantity of outliers. Data cleaning and transformation are called also data wrangling because getting your data in a form that's suitable to work often feels like a fight! So let's start wrangling over data. This task consists of a sequence of several text processing sub-tasks like below :

1. Language identification
2. Remove useless (non-word) characters, Lowercase and Tokenize
3. Remove stop words
4. Word normalization or stemming

Language Identification

First of all, we needed to determine in which languages, the answers are. The purpose of this step is basically to separate the corpus into multiple uni-lingual corpora. Ideally, all subsequent steps in the entire text mining pipeline should be done based on a unilingual corpus. The questions were asked in French language. However, we noticed that some answers of both questions were not in French. Some language identification methods are actually based on text mining. Python libraries NLTK and Scikit-Learn both contain methods for language identification that are based on text mining. We found that most of answers on the 4th and 5th questions were in French language. As a result, we removed non-French documents from corpus and we obtained 251 documents for the first question and 252 documents for the second question.

Many methods can be used for that purpose:

- Translate the the non-French documents in French by using the goslate library that helps to translate documents.
- Manually translate the non French answers.
- Delete the non French words.

Remove non-word characters

The notion of "useless" really depends on the application and the data. As a general rule, we should remove character that do not contribute to meaning. Major types of unwanted characters :

- Punctuation marks
- Math symbols

- Numbers
- Emoticons
- Hyperlinks (URL)
- Email addresses

After executing this step, we converted the documents into lowercase ones. Then it came the step of tokenization which consists of segmenting the character stream into words and that's based on precise linguistic rules for how words are generally separated in the French language.

Remove stop words

When we deal with text problem in Natural Language Processing, stop words removal is one of the important step to have a better input for any models later. Stop words means that it is a very common words in a language. In this step, we looked into using a library pre-defined stop words, we added the words that we think that they are missing in the library and we removed them for both questions.

Later on, we visualize the distribution of word lengths in every corpus.

After this step, the total number of words in the corpus of the fourth question was reduced from 12298 to only 1061 words and for the fifth question from 14270 to only 1227 words.

Word normalization or stemming

Normalization is a process that converts a list of words to a more uniform sequence. This is useful in preparing text for later processing. By transforming the words to a standard format, other operations are able to work with the data and will not have to deal with issues that might compromise the process. Normalization operations include stemming and lemmatization. We resorted to two techniques. The first one is the Snowball method from NLTK library. The second method is suggested by our supervisor Dr.Chiraz which consists of removing 's' at the end of the word if the word contains at least 6 characters and keeping only the first 7 characters of the word.

Document representation

Pattern extraction usually requires data to be in structured form, i.e. as numeric or symbolic attributes, called features as this kind of representation facilitates processing and mathematical modeling of data. In the case of text data, the simplest way is to represent documents based on words: i.e. where each feature corresponds to a unique word. There are two major word-based approaches :

- Bag-of-Words
- Word2vec & Glove

In the context of our project, we were interested in the Bag-of-Words approach. The reason behind the name of this approach is that the values in the feature vector do not depend on the order of words in the document. [**2**]

In the context of our project, we tried different configuration parameters to build the BOW model. First , we used the CountVectorizer function in order to construct it. Then we used the TfidfVectorizer function. We we ended up with finding the same output. That means we got the same features. Therefore we got the document term matrix.

3. Data analysis results and conclusions

3.1 Question 4: What are the reasons that would push you to leave Tunisia ?

This question might be one of the most important questions of the survey since knowing the reasons behind the phenomena would help us to find solutions to reduce it. The reasons as mentioned in the first chapter are related to different areas such as economics, politics, social... The reasons depends on too many factors but also it's different from one country to an other. This survey can help us to highlight the reasons which are related to the *Brain Drain* in Tunisia.

3.1.1 Word frequency

The first step is to highlight the frequent or the words with the highest occurrence in the answers.

	IDF	Word
90	2.435085	vie
66	2.578185	salaire
85	2.679281	travail
51	3.162133	opportu
39	3.351375	meilleu
55	3.438387	plus
46	3.484907	niveau
12	3.484907	conditi
63	3.484907	respect
27	3.533697	expérie

Figure 3.1: Words frequency

We can see that *vie* is the word the most frequent in our dataset which is logic. In fact, all the reasons of this phenomena are related to seeking for a better level of life. Thus, we found the words *"meilleur"* and *"niveau"*. The words *"salaire", "travail", "opportunité", "conditions", "respect" and*

"experience" highlight the needs pushing the qualified citizens to leave the country. Indeed, we were always searching for a better salary, better work conditions, better opportunities, respect of our abilities and efforts and better and new experience.

R The words are missing some letters due to the stemming we have done while the data processing.

We can see almost the same results (words) in this plot.

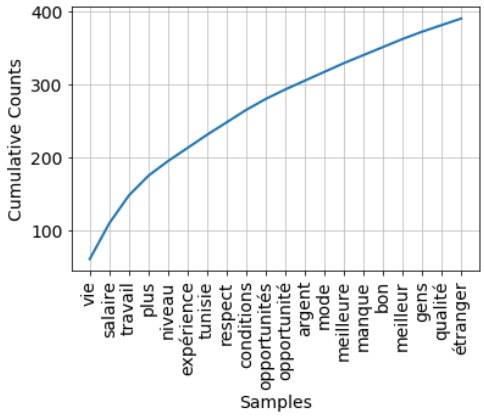

Figure 3.2: Cumulative counts of words

3.1.2 Topic Modeling

In this step, we were trying to search for the main topics mentioned in the answers. We suggested first two topic and later three. The results with three topics for this question seemed more interesting. These are the topics resulting:

```
Top 10 words for each topic:

Topic 0: respect recherc pays surtout faire gens dévelop salaire compéte manque
Topic 1: salaire opportu meilleu expérie conditi tunisie argent vie plus étude
Topic 2: travail vie niveau mode liberté salaire qualité mentali nouvell bon
```

Figure 3.3: Main three topics

Based on the coefficients, we could recognize three topics which are:

- Topic 1: Searching for respect of their competences

Topic 0

	term	topic_weight
63	respect	6.345199
60	recherc	4.134422
53	pays	3.883954
78	surtout	3.719391
28	faire	3.527976
31	gens	3.521488
22	dévelop	3.125341
66	salaire	3.093476
11	compéte	2.987671
37	manque	2.830552

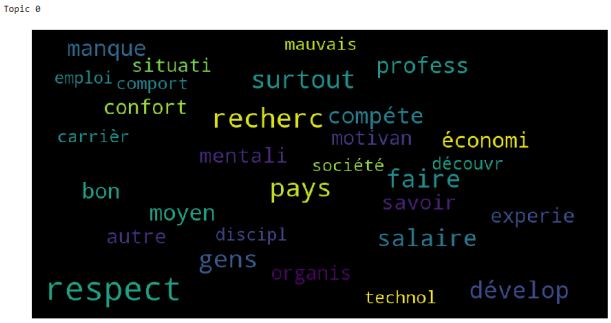

Figure 3.4: Topic one wordcloud

We can see in the wordcloud graph some other words that might be important for our study.

- Topic 2: Is devised into two sub-topics searching for a better salary and better opportunities.

Topic 1

	term	topic_weight
66	salaire	15.057613
51	opportu	13.865216
39	meilleu	12.721165
27	expérie	9.347029
12	conditi	9.072138
88	tunisie	8.435536
2	argent	8.021492
90	vie	7.612372
55	plus	6.860451
99	étude	4.723505

Figure 3.5: Topic two wordcloud

- Topic 3: Searching for better life style.

Topic 2

	term	topic_weight
85	travail	22.543945
90	vie	16.532591
46	niveau	9.883653
42	mode	7.359227
36	liberté	4.857764
66	salaire	4.466154
59	qualité	4.324984
40	mentali	3.791928
49	nouvell	3.617524
5	bon	3.466518

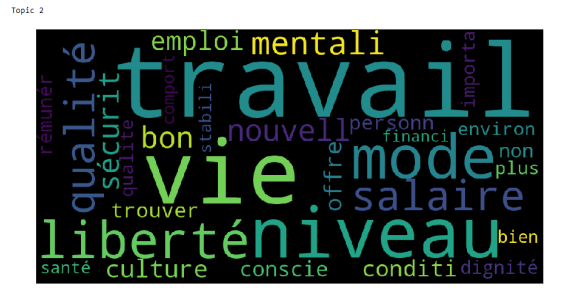

Figure 3.6: Topic three wordcloud

We can also see the distribution of the documents relatively to the topics we have put. We can see that a lot documents are related to one topic O, 1 or 2. But also, a lot of them are related to both topics 1 and 0 or 2 and 0.

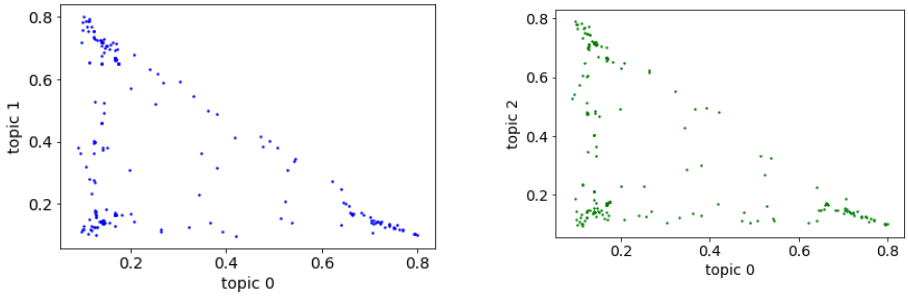

Figure 3.7: The documents distribution relatively to the topics

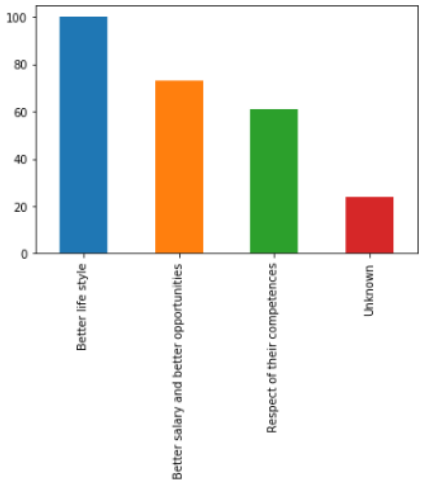

Figure 3.8: Plot showing the trend of the people toward the topics

3.1.3 Classification for prediction

We would like to extract a predictive model of the categories according to the other non-text attributes which will help us understand how the answers vary based on the characteristics of the respondent. Obviously, this is a classification problem.

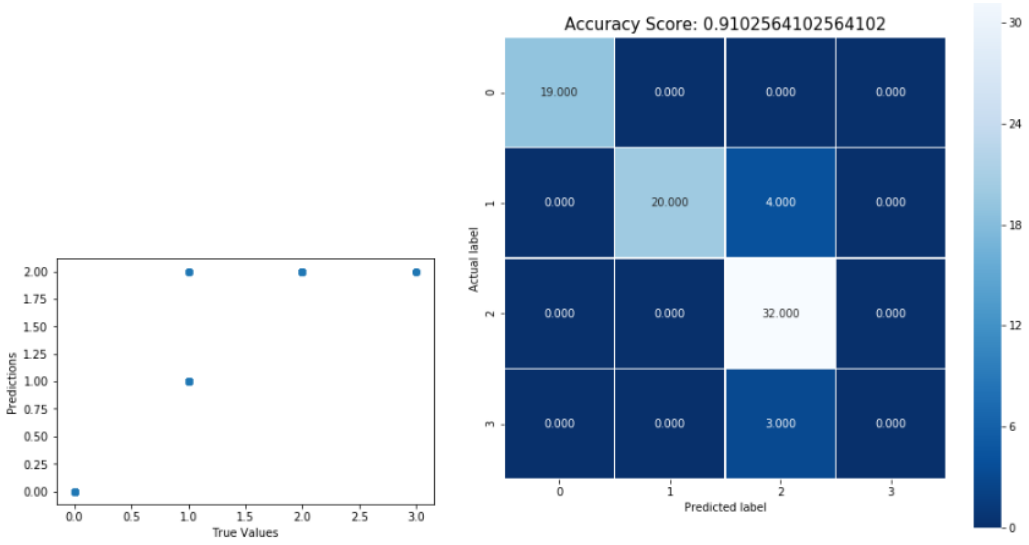

Figure 3.9: Evaluation of the prediction model: confusion matrix

3.1.4 Data Visualization

Data visualization is the presentation of data in a graphical format. Patterns and trends that might go undetected in text-based data can be exposed and recognized easier with data visualization.
Because the attributes 'Etablissement d'études' and 'Domaine d'études' have a large number of unique values, we reduce their number to 5 by aggregating the least frequent values to a single category called 'Other'.

Votre âge	3	Votre âge	3
Sexe	2	Sexe	2
Etat civil	2	Etat civil	2
Niveau d'étude	3	Niveau d'étude	3
Etablissement d'études	5	Etablissement d'études	83
Situation professionnelle	3	Situation professionnelle	3
Région	4	Région	4
Domaine d'études	2	Domaine d'études	44

Figure 3.10: The number of unique value for each of the 8 attributes non-text

Here are the plots visualization the correlation between the different features and topics.

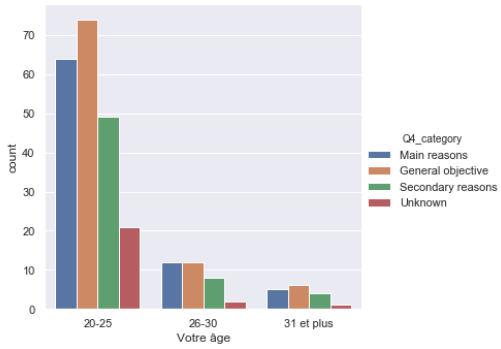

Figure 3.11: The categories according to the age

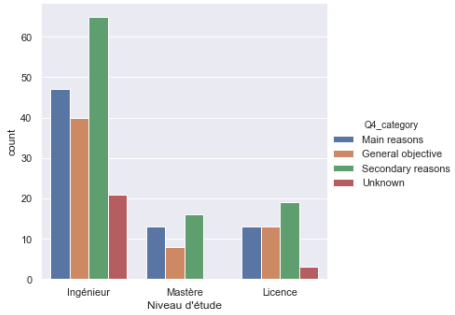

Figure 3.12: The categories according to the education level

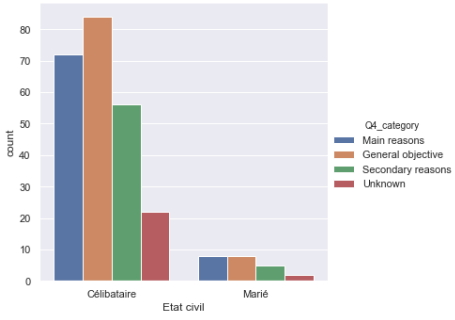

Figure 3.13: The categories according to the civil state

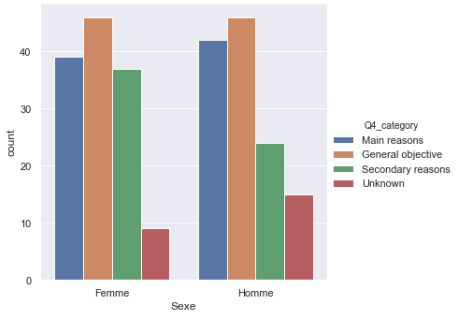

Figure 3.14: The categories according to the sexe

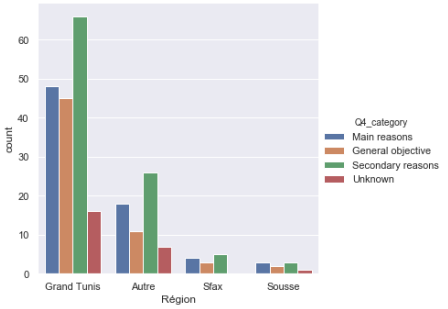

Figure 3.15: The categories according to the region

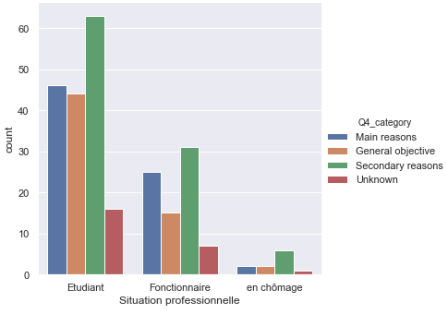

Figure 3.16: The categories according to the professional situation

We can see that none of these attributes is the best classifier, and this is due to the choice of the topics and the results of the text processing. A better data set can give much better results. However, for all the graphs we can conclude that both sexes, all the ages we have asked in this study with their different level of education and their different civil state agree on the same importance of topics.

3.2 Question 5: Socially, what's the difference between Tunisia and abroad in social life ?

This question can explain what the people are searching for when they are going abroad in the social area or the social reasons for the *Brain Drain*

3.2.1 Word frequency

The first step is to highlight the frequent or the words with the highest occurrence in the answers.

IDF	Word
2.582146	mentali
2.641569	tunisie
2.726727	vie
2.819817	respect
2.844510	plus
2.978041	étrange
3.275293	niveau
3.760801	gens
3.825339	liberté
3.825339	esprit

Figure 3.17: Words frequency

We can see that here the terms the most frequent in our dataset for this question.

R The words are missing some letters due to the stemming we have done while the data processing.

We can see almost the same results (words) in this plot.

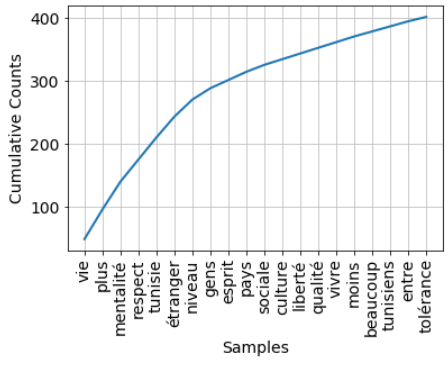

Figure 3.18: Cumulative counts of words

3.2.2 Topic Modeling

The three topics found are:

- Topic 1: Social level.

Topic 0

	term	topic_weight
91	tunisie	9.130309
99	étrange	4.725792
95	vivre	4.394485
53	pays	4.322487
78	société	4.060719
69	rien	3.934885
46	moins	3.475742
87	tous	3.104078
57	plus	3.083979
10	chaleur	3.065705

Figure 3.19: Topic one wordcloud

We can see in the wordcloud graph some other words that might be important for our study such as *racisme, service....*

- Topic 2: Mentality.

Topic 1

	term	topic_weight
43	mentali	32.417529
94	vie	18.463196
66	respect	12.978759
48	niveau	10.977995
39	liberté	8.127610
25	esprit	7.335497
30	gens	6.846118
99	étrange	6.571620
57	plus	6.100631
70	sais	4.819991

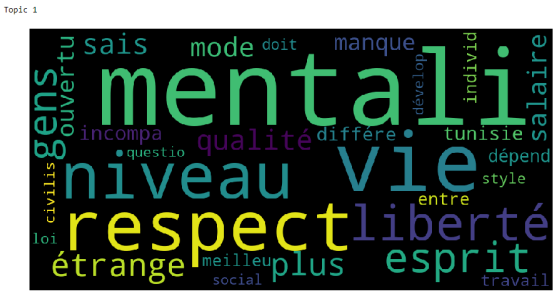

Figure 3.20: Topic one wordcloud

- Topic 3: Culture.

Topic 2

	term	topic_weight
14	culture	7.418605
57	plus	6.817226
66	respect	6.799016
86	toléran	6.589568
50	organis	5.089803
76	sociale	4.430552
29	famille	4.304616
89	travail	4.006111
85	sécurit	3.554464
5	autre	3.082284

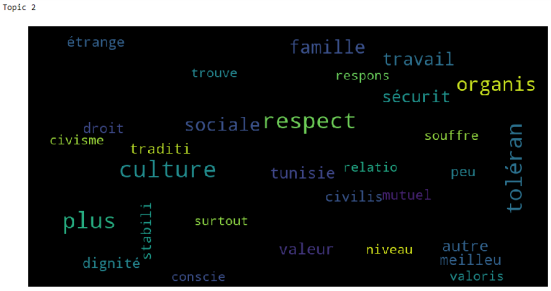

Figure 3.21: Topic one wordcloud

We can also see the distribution of the documents relatively to the topics we have put. We can see that a lot documents are related to one topic O, 1 or 2. But also, a lot of them are related to both

topics 1 and 0 or 2 and 0.

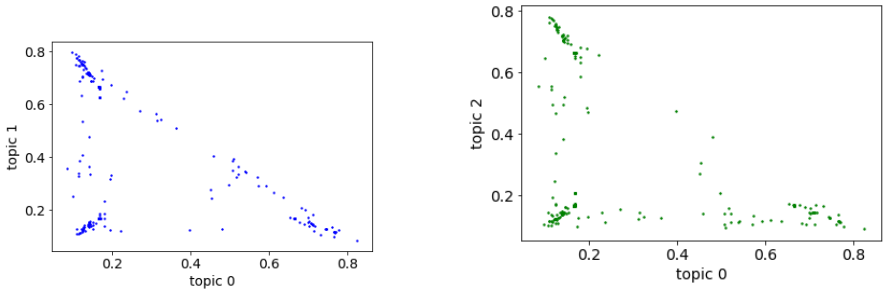

Figure 3.22: The documents distribution relatively to the topics

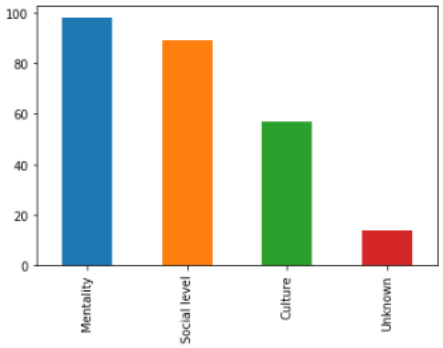

Figure 3.23: Plot showing the trend of the people toward the topics

3.2.3 Predection Model

We extracted a predictive model of the categories according to the other non-text attributes which will help us understand how the answers vary based on the characteristics of the respondent.

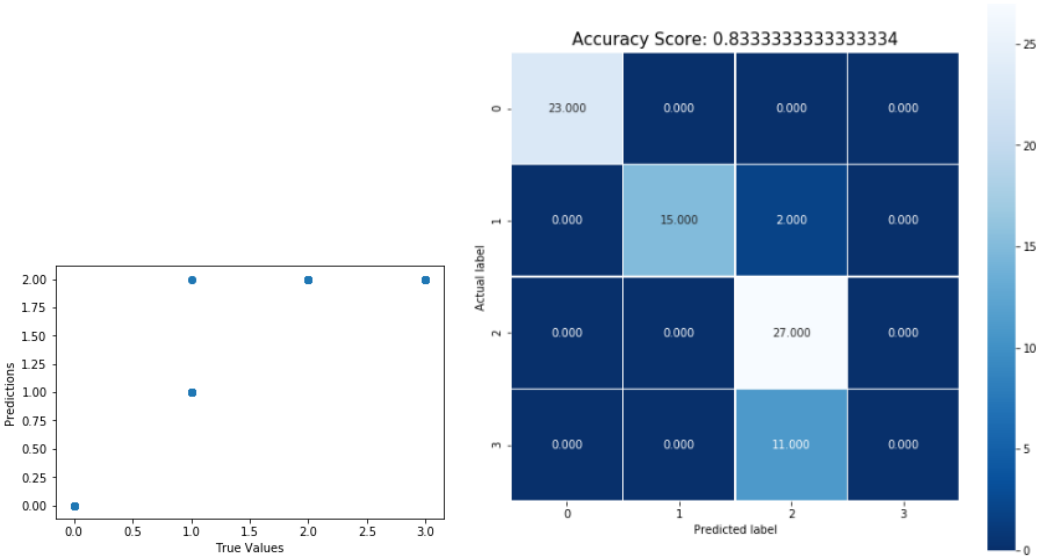

Figure 3.24: Evaluation of the prediction model: confusion matrix

3.2.4 Data Visualization

Data visualization is the presentation of data in a graphical format. Patterns and trends that might go undetected in text-based data can be exposed and recognized easier with data visualization.

Always after reducing the number of 'Etablissement d'études' and 'Domaine d'études' to 5, we would like to extract the relationships between the categories and the other non-text attributes.

Here too, we can see that none of these attributes is the best classifier, and this is due to the choice of the topics and the results of the text processing. A better data set can give much better results. However, for the civil state we can see that for the married people the mentality is really important and more relevant than the other general differences. For all the other graphs we can conclude that both sexes, all the ages we have asked in this study with their different level of education and their different civil state agree on the same difference.

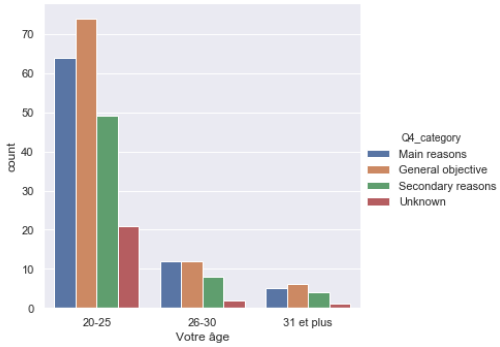

Figure 3.25: The categories according to the age

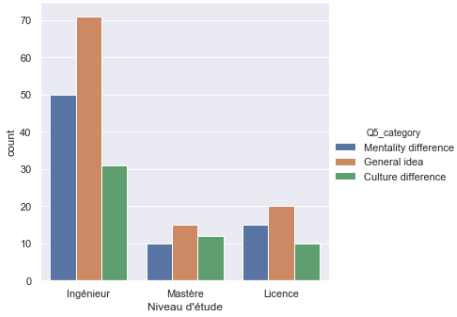

Figure 3.26: The categories according to the education level

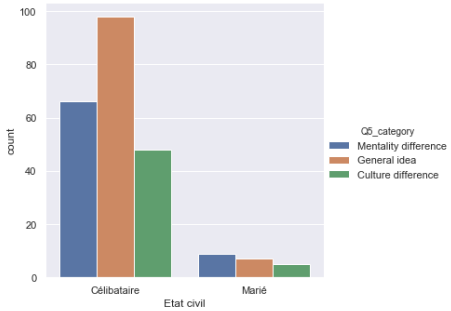

Figure 3.27: The categories according to the civil state

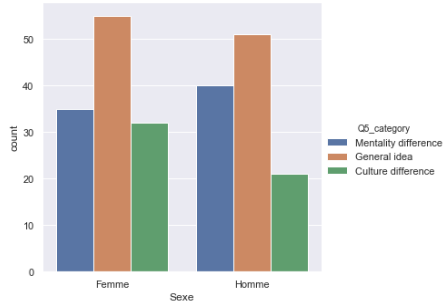

Figure 3.28: The categories according to the sexe

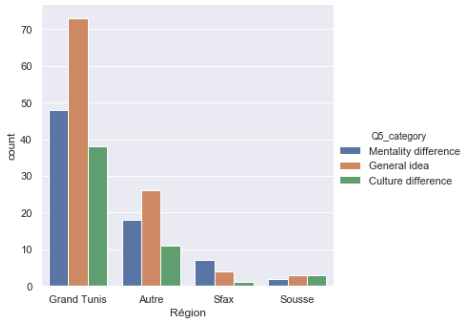

Figure 3.29: The categories according to the region

4. Conclusion

4.1 Commenting the Results

The results found shows the real main reasons encouraging the skilled and talented students and workers to try to find their way abroad not in Tunisia. This results can be resumed in this points:

- Seeking higher salary
- Seeking modern lifestyle and better comfort conditions
- Seeking better quality of education
- Seeking the respect and the freedom

The social difference is also one of the main reasons of the Brain drain. In fact, the Tunisians find many differences between the Tunisian society and the developed countries societies. The most important differences are related to the mentality and the culture of tolerance and freedom.

4.2 Further work and some interesting ideas

We can see that this study is just the start to search deeper in this subject. With a better dataset, this study can lead to highlight all the reasons and find the real solution to reduce this phenomena which can be really dangerous for our country.

Among the ideas to get a better dataset is to make the interviewees use a unique language. More question also can be interesting, such as the solution that the interviewees can suggest and the future they can predict for our country as a consequence for this phenomena. This study can also be interesting to be hold on different ages such as children, old people from different backgrounds.

I love my country... Don't just say it... Execute it!