National School of Engineering
of Tunis

TICV

Text Mining Project

Submitted By:

Mayara Elatrach
Fourat Thamri

Contents

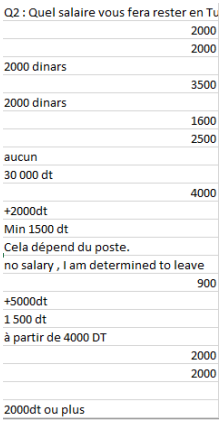
1 Introduction

Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interest. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).

2 Data preparation

2.1 Manual data preprocessing

Before going through the pipeline of data processing, we manually cleaned our corpus by:
- Eliminating arabic text.
- Translating english documents to french.
- Correcting grammatical errors for a homogeneous content.
- Homogenising salaries.



Q2 : Quel salaire vous fera rester en Tu
2000
2000
2000 dinars
3500
2000 dinars
1600
2500
aucun
30 000 dt
4000
+2000dt
Min 1500 dt
Cela dépend du poste.
no salary , I am determined to leave
900
+5000dt
1 500 dt
à partir de 4000 DT
2000
2000
2000dt ou plus

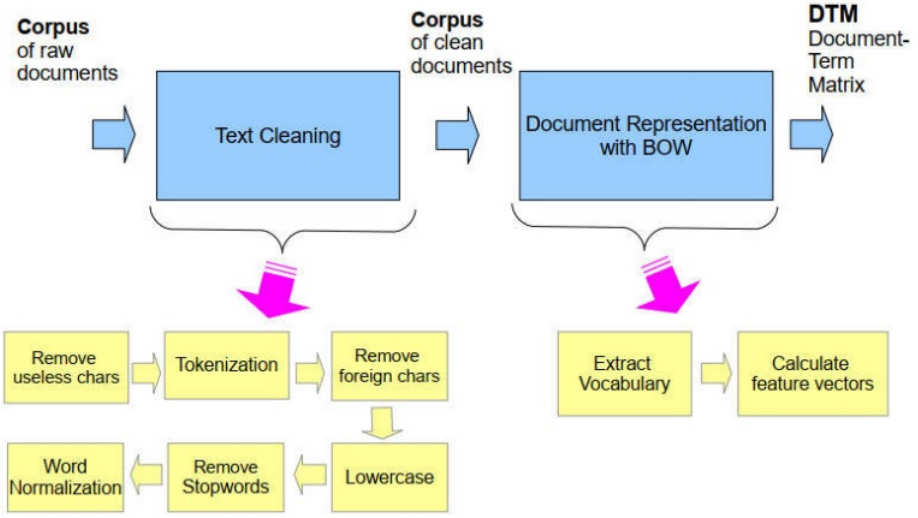
This image represents the column Q2 before modifications.

Q2 : Quel salaire vous fera rester en T
2000
2000
2000
3500
2000
1600
2500
aucun
30000
4000
>2000
>1500
Cela dépend du poste.
aucun
900
>5000
1500
>4000
2000
2000
>2000

This image represents the column Q2 after modifications.

2.2 Automatic data preprocessing

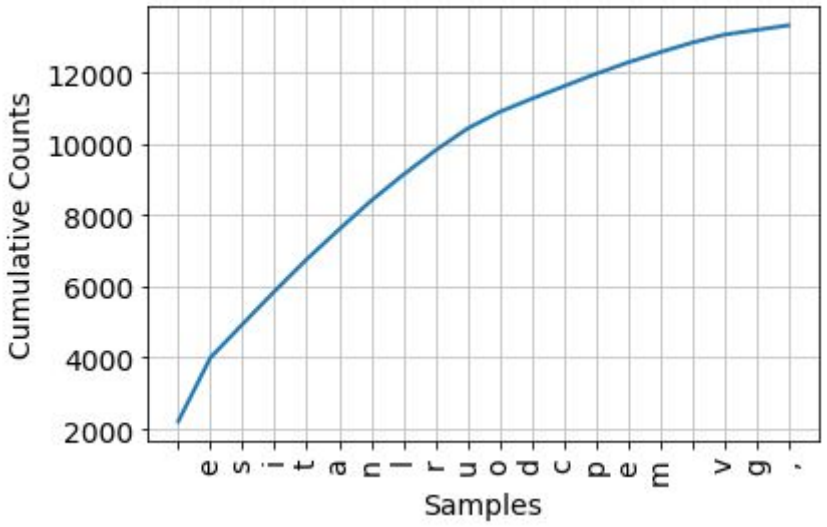
In order to process data we need to go through the following pipeline.



Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded. The tokens become the input for another process like parsing and text mining.

After manually cleaning our data: deleting Arabic lines and translating English

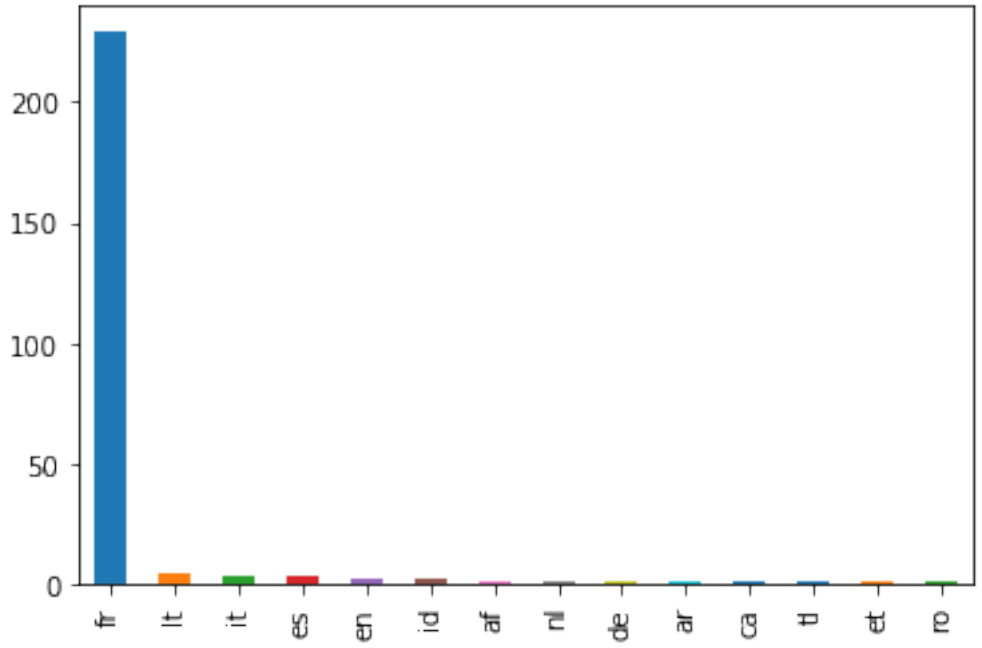
comments to french, we prepared our corpus for analyzing Q5 and Q6.
Our corpus contains characters that occur more frequently than others. The following graph shows the top 20 characters, they cover more then 95% of all character occurrences in the corpus.



Now that we have our corpus, we are going to clean it using NLTK. We followed the pipeline:
Language identification
Remove useless characters
Convert to lowercase
Tokenization
Remove stop words
Stemming

2.3 Language identification

Using the library NLTK, we identified many languages in the corpus. For example, these are the languages apparent in Q6-corpus with their frequency.



The most frequent language is French. The other languages could be basically neglected in front of French.
So now with our new French corpus, we deleted non-word characters. We then converted the whole text to lowercase and tokenized.

2.4 Removing stop words

In order to remove stop words from our corpus, we used the NLTK's default list of stop words for the French language.
However that list was not enough, so we manually added other stop words. So our final list of stop word list for Q6 became.

```
['au', 'aux', 'avec', 'ce', 'ces', 'dans', 'de', 'des', 'du', 'elle', 'en', 'et', 'eux', 'il', 'je', 'la', 'le', 'leur', 'lui',
'ma', 'mais', 'me', 'même', 'mes', 'moi', 'mon', 'ne', 'nos', 'notre', 'nous', 'on', 'ou', 'par', 'pas', 'pour', 'qu', 'que',
'qui', 'sa', 'se', 'ses', 'son', 'sur', 'ta', 'te', 'tes', 'toi', 'ton', 'tu', 'un', 'une', 'vos', 'votre', 'vous', 'c', 'd',
'j', 'l', 'à', 'm', 'n', 's', 't', 'y', 'été', 'étée', 'étées', 'étés', 'étant', 'étante', 'étants', 'étantes', 'suis', 'es',
'est', 'sommes', 'êtes', 'sont', 'serai', 'seras', 'sera', 'serons', 'serez', 'seront', 'serais', 'serait', 'serions', 'serie
z', 'seraient', 'étais', 'était', 'étions', 'étiez', 'étaient', 'fus', 'fut', 'fûmes', 'fûtes', 'furent', 'sois', 'soit', 'soyo
ns', 'soyez', 'soient', 'fusse', 'fusses', 'fût', 'fussions', 'fussiez', 'fussent', 'ayant', 'ayante', 'ayantes', 'ayants', 'e
u', 'eue', 'eues', 'eus', 'ai', 'as', 'avons', 'avez', 'ont', 'aurai', 'auras', 'aura', 'aurons', 'aurez', 'auront', 'aurais',
'aurait', 'aurions', 'auriez', 'auraient', 'avais', 'avait', 'avions', 'aviez', 'avaient', 'eut', 'eûmes', 'eûtes', 'eurent',
'aie', 'aies', 'ait', 'ayons', 'ayez', 'aient', 'eusse', 'eusses', 'eût', 'eussions', 'eussiez', 'eussent', 'car', 'les', 'êtr
e', 'quelque', 'chose', 'comme', 'ici', 'aussi', 'beaucoup', 'mieux', 'entre', 'surtout', 'avoir', 'très', 'parce', 'où', 'si',
'chaque', 'donc', 'dont', 'encore', 'faut', 'quand', 'tant', 'tel', 'tout', 'va', 'vient', 'peut', 'veut', 'ils', 'plus', 'moin
s', 'ainsi']
```

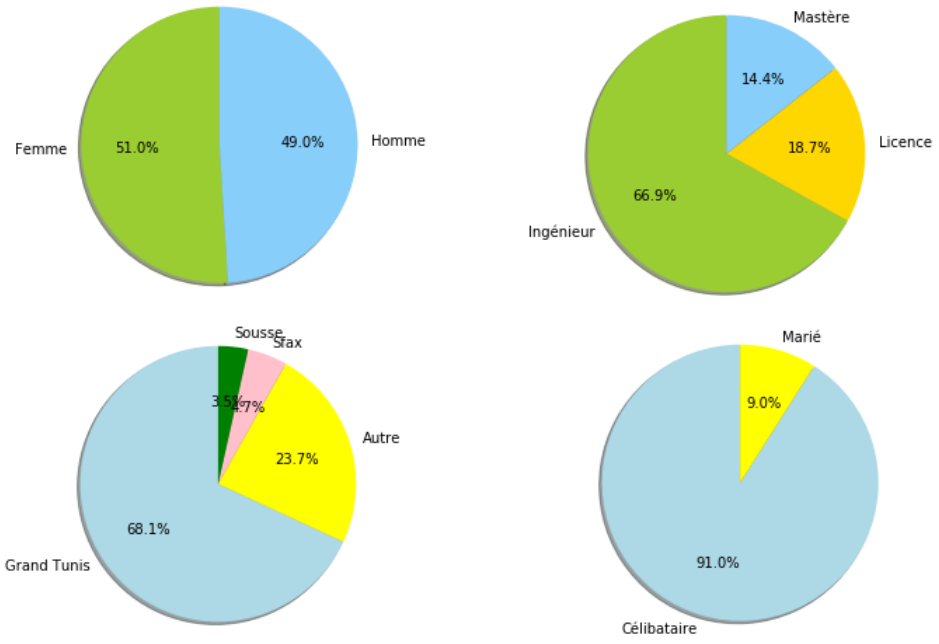
So the corpus is now composed of only non stop words and it's all in French.

2.5 Stemming

In order to normalize words we used two different methods. The first one is the Snowball method from the package NLTK.
the second one is Dr. Chiraz BenAbdelkader's method which consists of: removing s at the end of the word if the word contains at least 6 characters and keeping only the first x characters of the word. x will vary from Q5 and Q6. We slightly modified the second method by adding new rules.
Now that we have a clean corpus, we will transform it into a Document Term Matrix (DTM). DTM is feature vectors stored in rows of matrix.
In order to do that, we are going to use the Bag-Of-Words model.In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity.

3 Statistics

We were able to collect 257 responses. Answers consist of having responses to 6 questions and providing personal information that are represented by the following pie charts.



Salaries

This is the distribution of the most frequent answers in the question 2 about the desired salary to stay in Tunisia, that we cleaned manually .

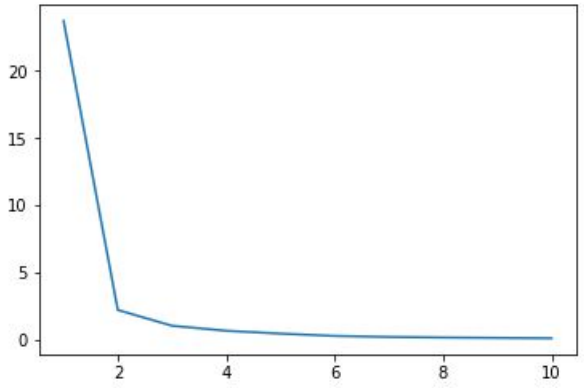
	Frequency
2000	44
3000	37
2500	19
5000	18
1500	13
4000	11
>2000	9
>1500	7
ce n'est pas une question de salaire	6
1800	5
1200	5
10000	4
>2500	4
3500	4
Aucun	3

We observed that approximately 45% of people are ready to stay in Tunisia for a salary between 2000 Dinars and 3000 Dinars.

3.1 Pattern extraction and evaluation for Q6

3.2 Clustering with K-means

Now that we have our corpus ready. We divided the corpus into 2 clusters for Q6. In fact, using the K-means method the elbow of the graph indicated that we have two clusters.



The curve drastically changes in $K = 2$ thus we choose two clusters.
Hence, these are the cloud of words of our corpus for the two clusters.

Cluster 1:

Cluster 2:

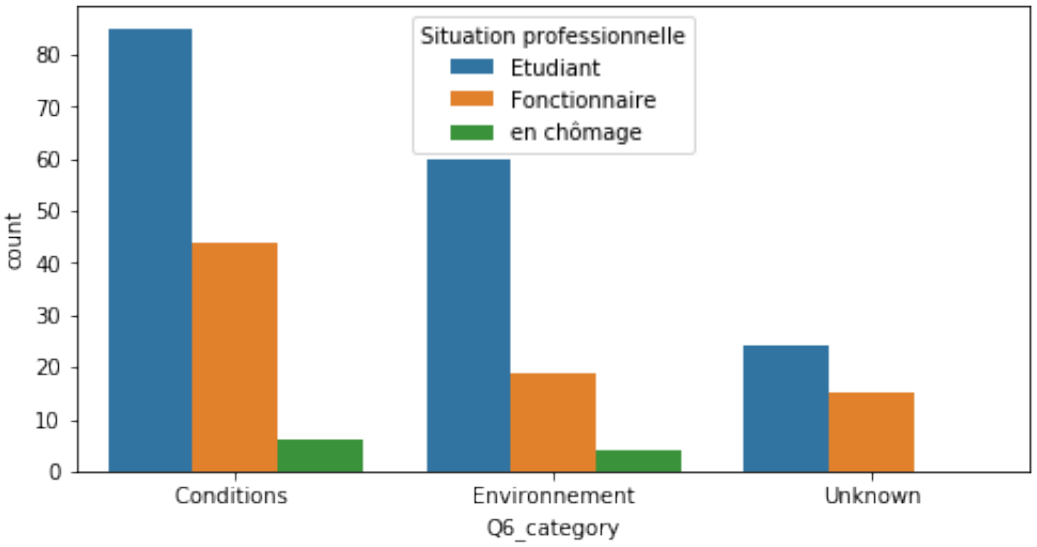
3.3 Topic Modelling with LDA

Using the Latent Dirichlet Allocation method we obtained two topics.
- Environnement
- Conditions

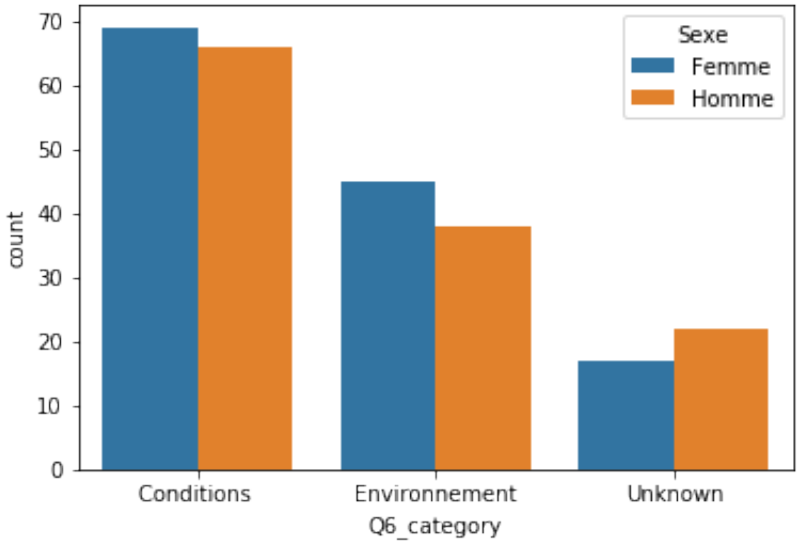
This is the cloud of words for the first topic: **Environnement**

This is the cloud of words for the second topic: **Conditions**

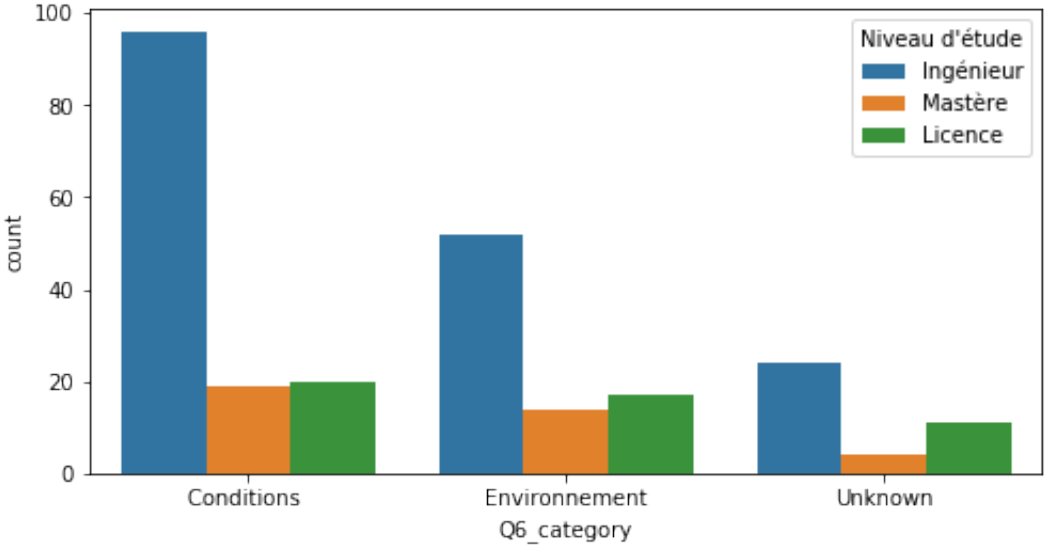
Now we want to know how personal information such as the civil state, age and studies affect the response to the questions.



We notice that the students would like to leave Tunisia mostly because of work conditions.



There is no significant difference between the responses of men and woman to Q6.

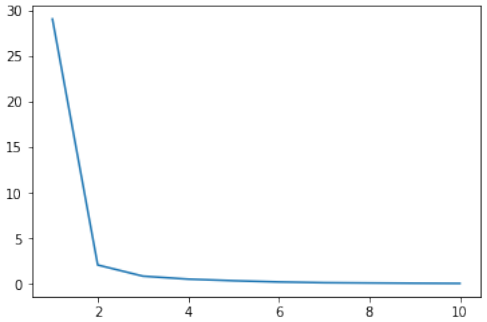


Engineers are the most prone to leave Tunisia because of work conditions.
However in this case, we notice that the Licence graduates seem more prone to leave because of the environment rather than the conditions.

3.4 Pattern extraction and evaluation for Q5

3.5 Clustering with K-means

Same as Q6, we used K-means to cluster our corpus using the LDA document weights.



The curve drastically changes in $K = 2$ thus we choose two clusters.

Hence, these are the cloud of words of our corpus for the two clusters.

Cluster 1:

Cluster 2:

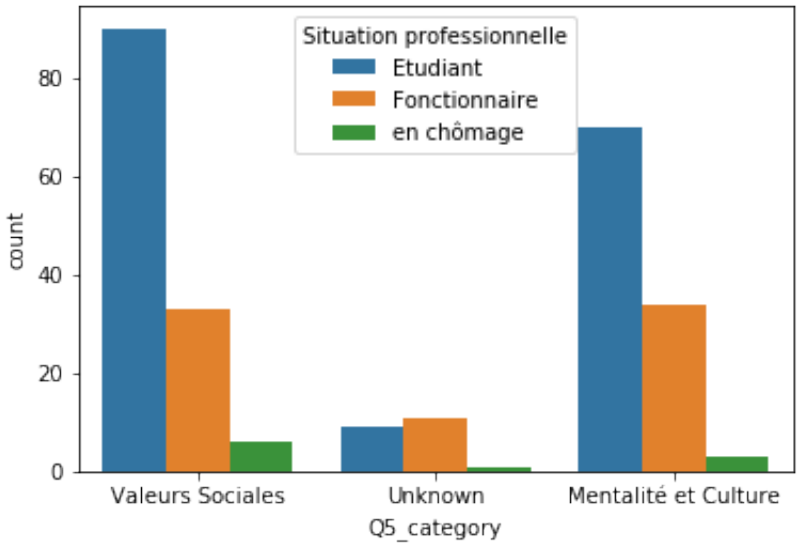
3.6 Topic Modelling with LDA

Using the Latent Dirichlet Allocation method we obtained two topics.
- Valeurs Sociales
- Mentalité et culture

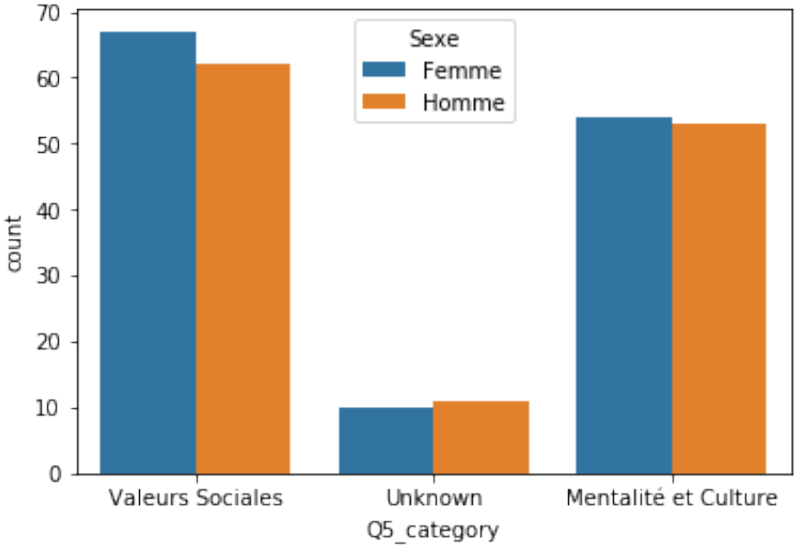
This is the cloud of words for the topic: **Mentalité et culture**

This is the cloud of words for the topic: **Valeurs Sociales**

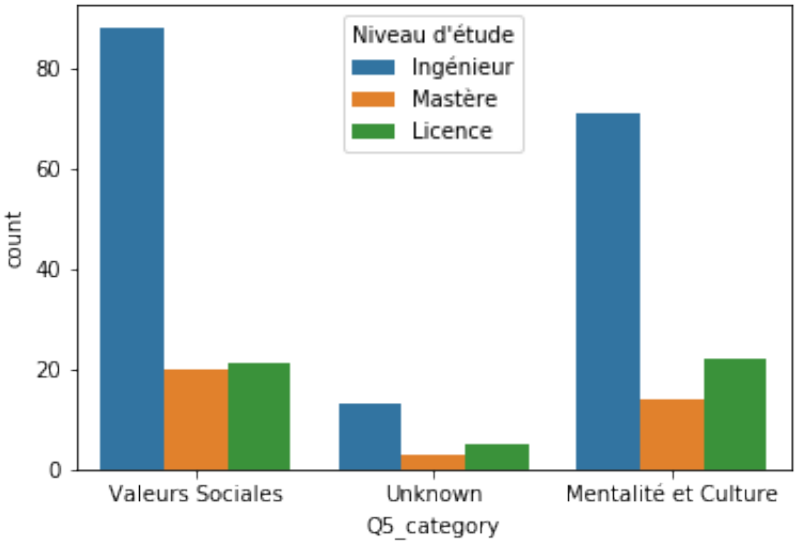
Now we want to know how personal information such as the civil state, age and studies affect the response to the questions.



We notice that the students would like to leave Tunisia mostly because of social values.



There is no significant difference between the responses of men and woman to Q5.



Engineers are the most prone to leave Tunisia because of social values.
However in this case, we notice that the Licence graduates seem more prone to leave because of the mentality rather than the social values.

4 Conclusion

In this project, we had the opportunity to work with text data that we collected. We cleaned the data using a pipeline in which we removed stop words and used the Bag-Of-Word model. After that we clustered these data and used a linear regression to classify them and commented on the results.