Report of SD210 Grand débat

Analysis for the topic clustering of ecological transition in Grand Débat (Based on different regions)

Author:
Chong SUN
Fengli LIN
Jizhe LIU
Yutong ZHAI

Supervisor:
Alexandre Garcia

April 28, 2019

Abstract

The main purpose of our task is to analyse the topic difference of ecological transition in Grand Débat on different regions and it is separated into three parts: firstly, we do some basic contents analysis based on the original data set and pre-processing on our data; Secondly, we do word split and dimensionality reduction; Thirdly, we realize three kinds of clustering algorithm based on the processed data: Tf-idf,LDA topic model and Word2Vec based K-means clustering.

The basic content analysis contains the comparative analysis of four topics and the detailed analysis on the topic of ecological transition. For the detailed analysis in this chosen topic, we process it into two step: The component analysis of closed questions in this topic and pre-processing and the component analysis of user's region information.

Word split and dimensionality reduction is done in order to analyse the open questions. in the reduced form which will be done in our second step. In the part of dimensionality reduction, we would use different methods to embed the phrase or the words into the vector space, and then apply the dimensionality reduction methods like PCA,t-SNE and compare their results in the clustering tasks.

We using Tf-idf to vectorize the answers and then we apply the LDA models to analyse the main topics. We choose one certain question and train our model by all the answers. The result reflects the global situation for this question. Then we compare the different regions and we aim to find the top topics in these regions. We adopt several visualization tools to help us understand the result intuitively.

For the part of Word2Vec based K-means clustering, we first use pre-trained french Word2Vec model to transform preprocessed words into vectors and represent sentence by the average of all word vectors. Next, we use K-means algorithm to clustering the sentence vectors. Here, we find the best K through drawing the elbow curve. After that, we choose those responses which are nearest to the cluster centroids as the representing sentences of each class. Finally, we perform tf-idf analysis on the representing sentences to get the key words.

1 Introduction

2 Basic content analysis

In this part, we show our result into three part. In the first section, we demonstrate the comparative analysis of four topics. In the second section, we detail the component analysis of close question about the topic of ecological transition. In the last section, we discuss the region information about the users in this topic and the way to split our original data set.

2.1 Comparative analysis of four topics

We calculate the number of people who answered questions for every topic and for every day between 2019-01-22 and 2019-03-08. This figure 1 shows our results. Among all the topics, the topic "Taxation and public expenditure" gains the highest popularity during the whole period. The second most popular topic is "Ecological transition".

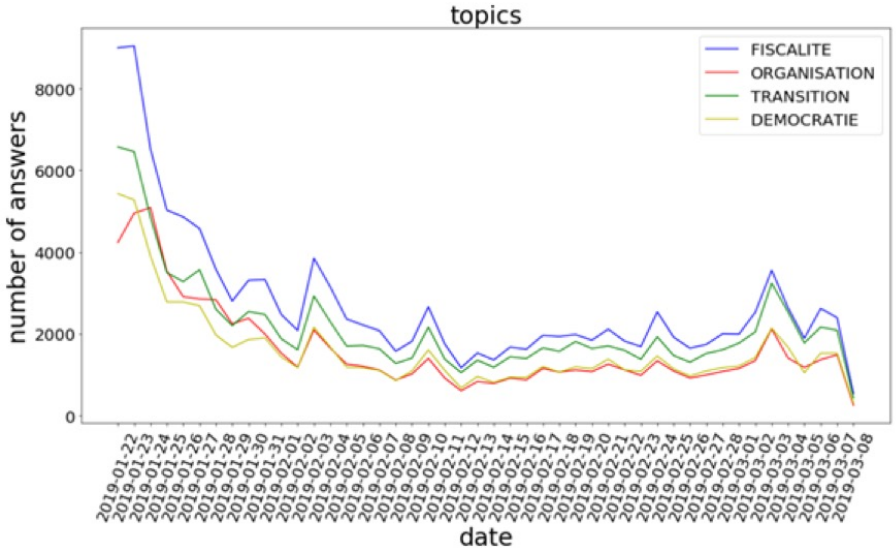
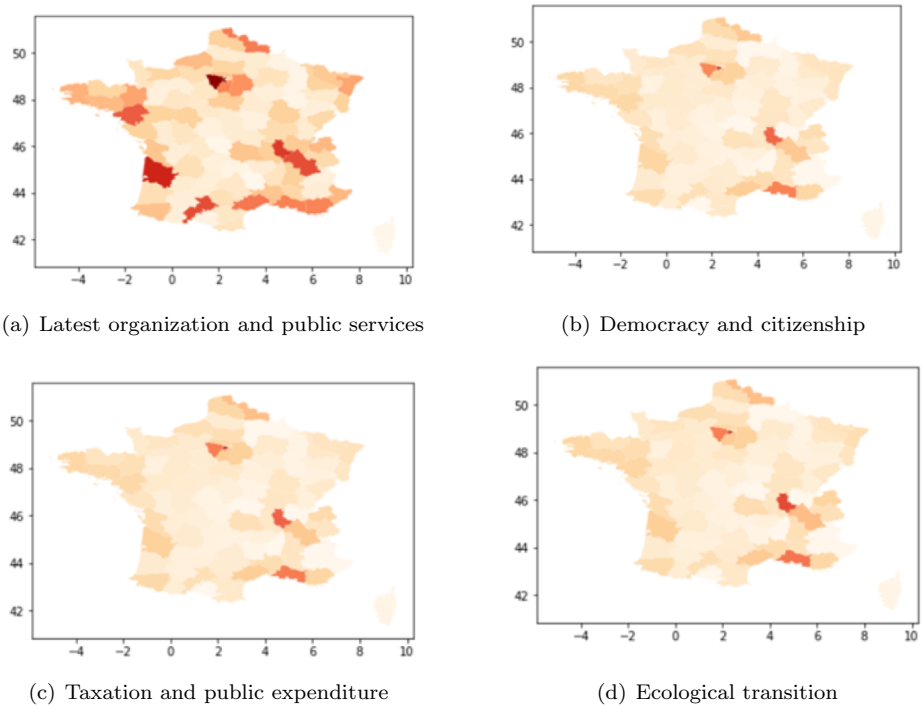


Figure 1: **Number of answers of four topics**

As for the last two topics, "Latest organization and public services" and "Democracy and citizenship", during this period they get close numbers of answers nearly every day. Furthermore, the number of answers for every topic reaches its peak at the first or second day. Then it dramatically decreases in the following few days and remains between 2000 and 4000 with some instabilities. It is understandable since people are curious when they meet something new.

For every topic, we also analyzed their popularity geographically. The following

four figures reflect the great differences that exist in different provinces. Situations for the "Latest organization and public services", "Democracy and citizenship", "Taxation and public expenditure" and "Ecological transition" are nearly the same, where there exist three places in deep color: Paris, Lyon and Marseille. However, the topic "Latest organization and public services" owns the highest participation geographically. Figure 2 indicates there are at least eight places in significant deeper color. Moreover, figure 2(a)-(d) illustrates that for all the topics, it is always Paris where they were widely discussed.



(a) Latest organization and public services

(b) Democracy and citizenship

(c) Taxation and public expenditure

(d) Ecological transition

Figure 2: **Answers distribution in different region**

2.2 Component analysis of close question

Then, we choose the topic "Ecological transition" to study. Firstly, we focus on studying the closed questions.

"Par rapport à votre mode de chauffage actuel, pensez-vous qu'il existe des solutions alternatives plus écologiques?", after filtering the NaN answer, we draw a pie chart of the answer Yes and No. From Figure 3, we can see that there are more people having other solution for heating.

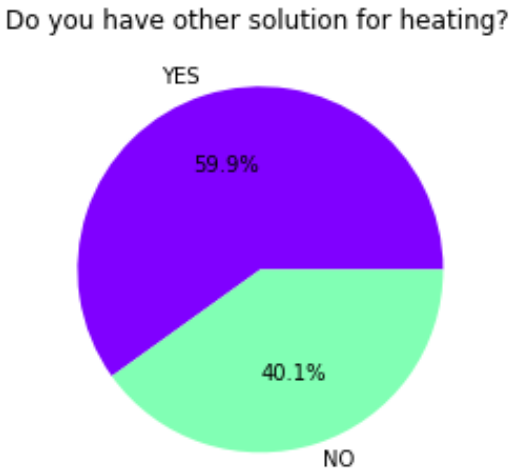


Figure 3: **"Compared to your current heating mode, do you think there are alternative solutions more ecological?"**

We did the same thing at the question "Avez-vous pour vos déplacements quotidiens la possibilité de recourir à des solutions de mobilité alternatives à la voiture individuelle comme les transports en commun, le covoiturage, l'auto-partage, le transport à la demande, le vélo, etc?". From figure 4, there are almost one half of people having other solution for the commuting rather than car.

Following the question above, we continue to study the question "Si non, quelles sont les solutions de mobilité alternatives que vous souhaiteriez pouvoir utiliser ?". Then we count the number of each alternative solution and draw the figure 5. From the graph, we see that the common transports takes the biggest proportion.

2.3 Discussion on the region information

In this part, considering the number of different department maybe too large, we use the latest region split criterion to divide our original "transition ecological. csv" file into 15 different regions files.

Because the original information about "authorZipCode" contains human mistakes and sometimes are incomplete. So the first step I did is to process them into a standard format. Here I standard the "authorZipCode" into department postal code and then according to the latest region partition criterion, I pick out the same region user's information into it's appropriate csv file.

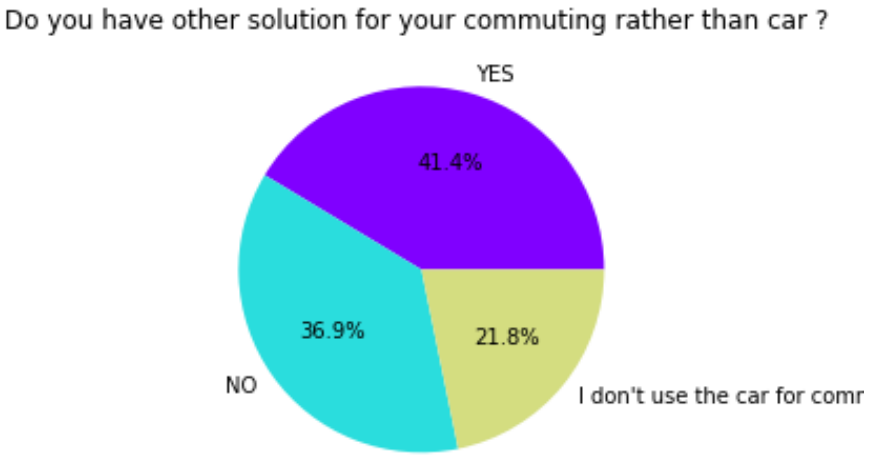


Figure 4: **Do you have the possibility for your daily commute to use alternative mobility solutions to the private car such as public transport, carpooling, car-sharing, transport on demand, cycling, etc.?**

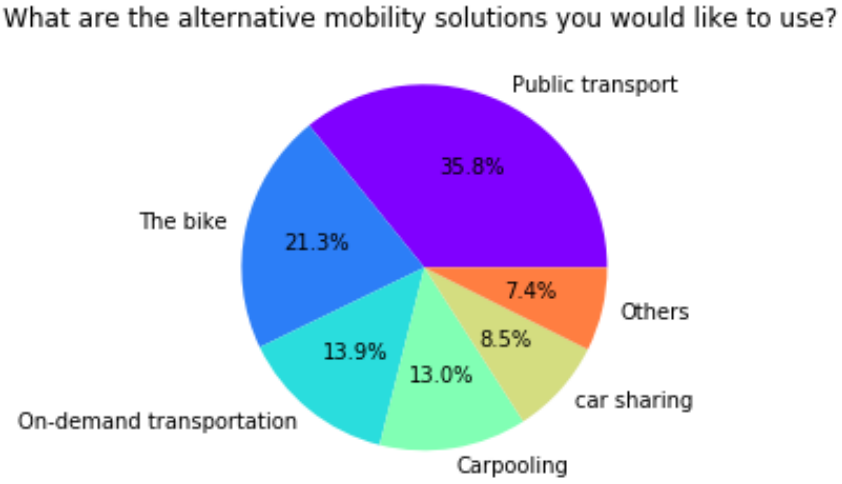


Figure 5: **If not, what alternative mobility solutions would you like to use?**

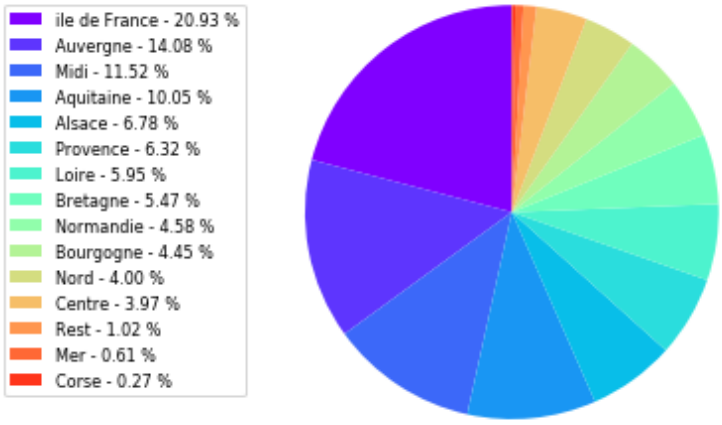


Figure 6: **Answers on the topic of ecological transition according to different regions**

After pick out the users from Alsace-Champagne-Ardenne-Lorraine, Aquitaine-Limousin-Poitou-Charentes, Auvergne-Rhône-Alpes, Bourgogne-Franche-Comté, Bretagne, Centre-Val de Loire, Corse, Île-de-France, Languedoc-Roussillon-Midi-Pyrénées, Nord-Pas-de-Calais-Picardie, Normandie, Pays de la Loire, Provence-Alpes-Côte d'Azur and Régions d'outre-mer, I foud out that we still have 889 items with "authorZip-Code" not in any department. For the sake of perseverance of the original answers, I combine them into a "Rest.csv" file.

In the figure 6, we can see the different regions has different contribution to the topic of ecological transition. Apparently, people from ile de France participate most actively in this topic, people from Auvergne-Rhône-Alpes place the second. The third and forth biggest participant group from Languedoc-Roussillon-Midi-Pyrénées and Aquitaine-Limousin-Poitou-Charentes. The rest regions' degree of participation are far lower than the first four regions. However, the population density is one reason for this result.

3 Latent Dirichlet Allocation (LDA)

3.1 Introduction

We would like to analyse the topic from the answers and here we will use LDA method to cluster some main topics and extract key words.

LDA is widely used in the topic model and its a generative statistical model which help us group the data through the similarity. Here we also use the tf-idf to vectorizer the documents and the result could directly apply for the transform process.

3.2 Analysis

We choose the question:" Y a-t-il d'autres points sur la transition écologique sur lesquels vous souhaiteriez vous exprimer ?" to apply the algorithm and we will show the result through some visualization tools.

The data has been preprocessed and we define the topic number as 10 and we show the top words then we could see the main topic results as below:

```
Topic #0:
produire agriculture bio pesticide agriculteur interdire glyphosate favoriser aider production
Topic #1:
voiturer véhiculer électrique batterie diesel électriques moteur hydrogéner polluer essence
Topic #2:
transition écologique devoir politiquer pouvoir faire falloir citoyen entreprendre action
Topic #3:
nucléaire énergie éolien centrale solaire développer renouvelables rechercher électricité production
Topic #4:
logement aider ville pouvoir isolation travail bâtiment devoir énergie panneau
Topic #5:
faire falloir arrêter pouvoir écologie aller taxer pays dire transition
Topic #6:
emballage plastiquer déchet tri produire consigner recyclage bouteille interdire jeter
Topic #7:
animal chasser biodiversité eau espèce protection interdire protéger zoner animale
Topic #8:
falloir enfant pouvoir planète faire devoir vie écologique climatique changement
Topic #9:
taxer transport camion transports routier pollueur oui ferroutage router bateau
```

Figure 7: **10 Topics got from sklearn tools**

We find the words in the same topic are relevant and we could find some relationship, for example "glyphosate" and "bio" shows the topic 0 is the agriculture related topic. And people care about the products , word "aider" might mean to help the poor people or to increase the standard of food supervision.

However we haven't gotten the specific coefficient of these topic and we want to know the degree of importance through the numbers. Here we adopt another library "gensim". We define 5 topic then we could get the result:

```
[(0,
  '0.021*"produire" + 0.011*"agriculture" + 0.011*"aider" + 0.009*"bio" + 0.008*"logement"'),
 (1,
  '0.017*"pouvoir" + 0.017*"faire" + 0.015*"falloir" + 0.015*"devoir" + 0.014*"écologique"'),
 (2,
  '0.015*"voiturer" + 0.014*"véhiculer" + 0.012*"déchet" + 0.011*"faire" + 0.009*"plastiquer"'),
 (3,
  '0.027*"énergie" + 0.018*"nucléaire" + 0.015*"taxer" + 0.013*"éolien" + 0.011*"développer"'),
 (4,
  '0.009*"devoir" + 0.007*";" + 0.006*"effet" + 0.006*"européen" + 0.005*"pouvoir"')]
```

Figure 8: **5 Topics got from library "gensim"**

Now we are easier to identify these topics, the topic 0 is about the agricultural product and the discuss about the "bio" product and social welfare for food and residence. Topic 1 is about the ecological protection and the actions we should apply. Topic 2 is about the problem of waste and the pollution which are leaded by the cars and plastic. Topic 3 is about the energy and sustainable development, people talk much about the nuclear energy and also the tax policy about the energy. Topic 4 is about some effects we got in the european and the next step we should consider.

We find our result seems good and we could abstract the hot topic that people discussed, then we need some evaluation tools to support our idea. We draw the distance and relevant relationship as below:

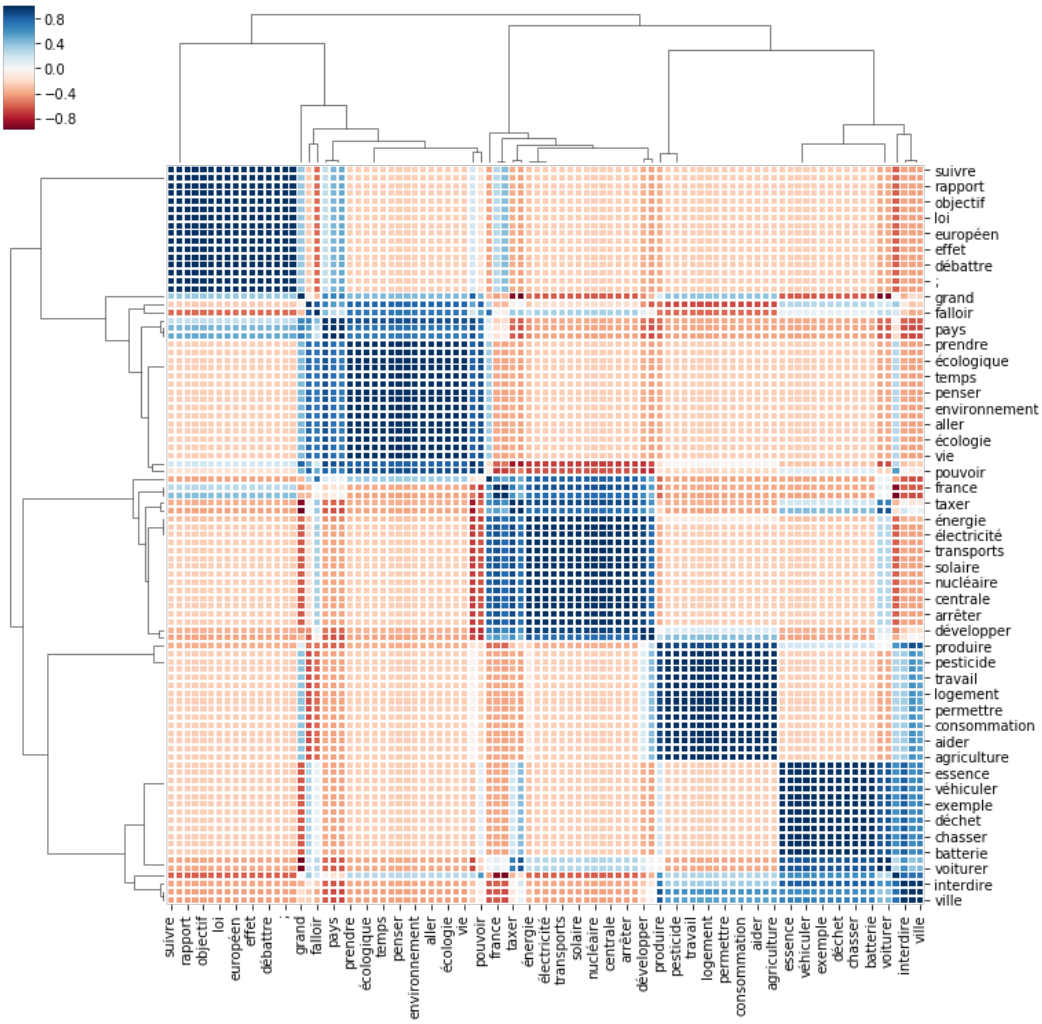


Figure 9: **Hierarchically-clustered heatmap**

The performance is good and we could clearly observe these 5 topics, amount the same group the relevance is very high and the words between different topics have little relationship. It support that these groups have been divided well and we find the main topics successfully.

3.3 Compare the region "Île-de-France" and "Bretagne"

The former result show the popular topics for that discussion, we would like to compare the difference between the region "Île-de-France" and "Bretagne" for

the question:"Que pourrait faire la France pour faire partager ses choix en matière d'environnement au niveau européen et international ?".

We process the answers by the same process and we could get the topics:

```
[(0,
  '0.028*"climatique" + 0.026*"problème" + 0.015*"ensemble" + 0.015*"réchauffement" + 0.012*"falloir"'),
 (1,
  '0.012*"environnement" + 0.012*"ressource" + 0.011*"énergie" + 0.011*"vie" + 0.009*"nucléaire"'),
 (2,
  '0.038*"problème" + 0.038*"lier" + 0.032*"importer" + 0.024*"eau" + 0.019*"sol"'),
 (3,
  '0.233*"dérèglement" + 0.227*"climatiques" + 0.216*"sécheresse" + 0.215*"croire" + 0.002*"interdépendants"'),
 (4,
  '0.217*"pollution" + 0.198*"air" + 0.162*"biodiversité" + 0.154*"disparition" + 0.153*"espèce"')]
```

Figure 10: **5 Topics in Île-de-France**

We could visualize the results using the pyLDAvis and have a intuitive understand(Due to incompatibility,we couldn't show the graphs in the jupyterlab while we could executate in jupyter notebook):

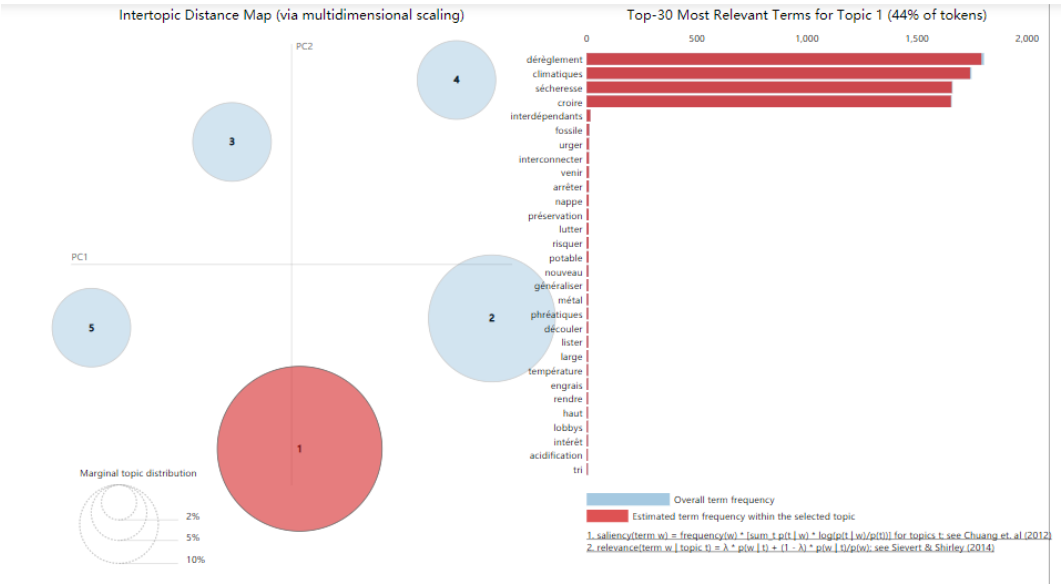


Figure 11: **Intertopic Distance Map in Île-de-France**

The topics are ordered by the amounts of relevant terms, the most representative topic in Île-de-France is topic 3 which is about the drought problem and the climate change. In some terms we could believe that Parisian care more about these problems.

Then we have a look for the performance of dividing the groups:

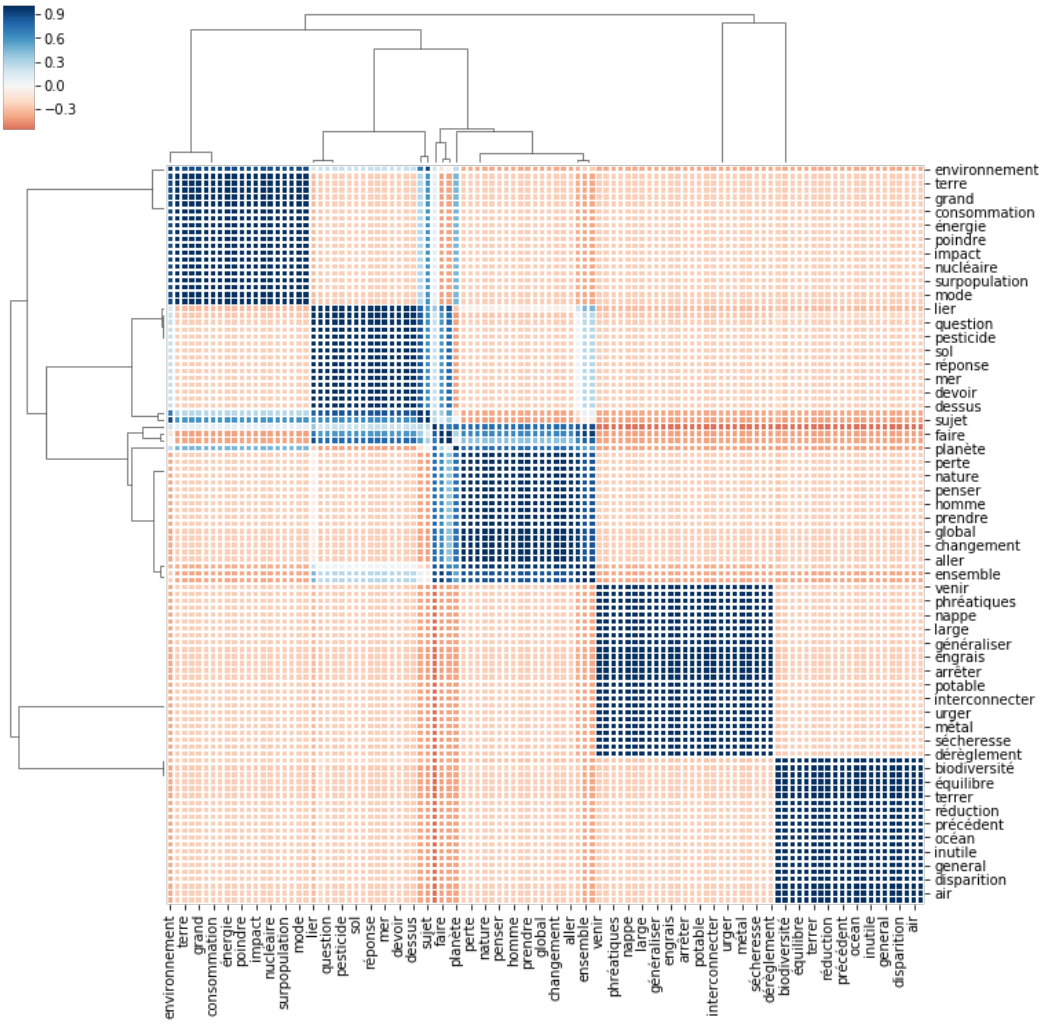


Figure 12: **Hierarchically-clustered heatmap in Île-de-France**

Because this problem is about the subject "environment", we find the scores are nearly positive. The five topics are divided well and we could abstract the idea from these words.

Then let's have a look at the results in Bretagne:

```
[(0,
  '0.142*"biodiversité" + 0.140*"disparition" + 0.138*"espèce" + 0.123*"pollution" + 0.103*"air"'),
 (1,
  '0.031*"lier" + 0.013*"environnement" + 0.011*"problème" + 0.008*"niveau" + 0.007*"vie"'),
 (2,
  '0.035*"biodiversité" + 0.033*"problème" + 0.025*"disparition" + 0.023*"espèce" + 0.020*"climatique"'),
 (3,
  '0.017*"problème" + 0.015*"pollution" + 0.010*"ensemble" + 0.010*"proposition" + 0.008*"compter"'),
 (4,
  '0.204*"dérèglement" + 0.201*"climatiques" + 0.193*"sécheresse" + 0.193*"croire" + 0.004*"importantes"')]
```

Figure 13: **5 Topics in Bretagne**

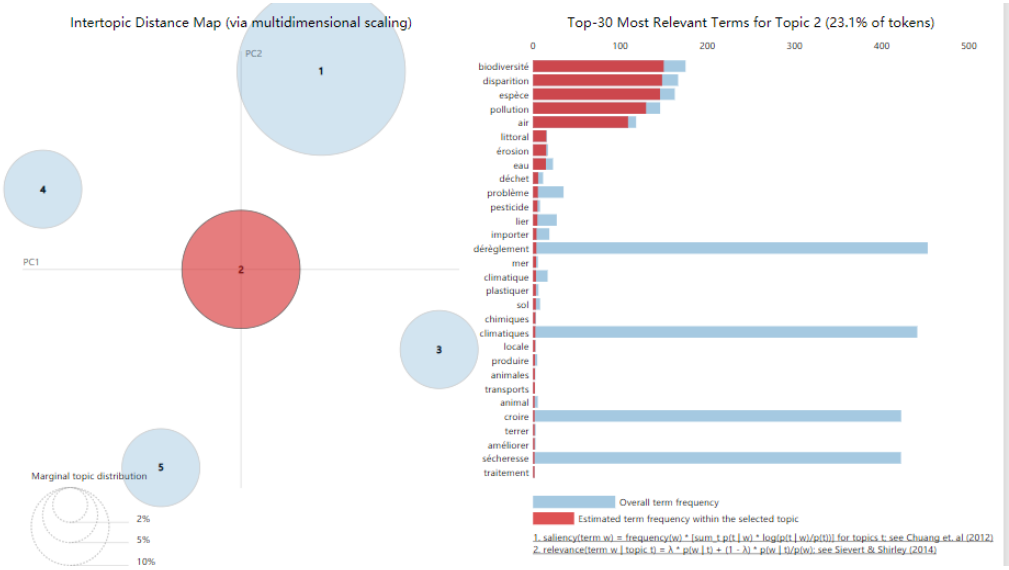


Figure 14: **Intertopic Distance Map in Bretagne**

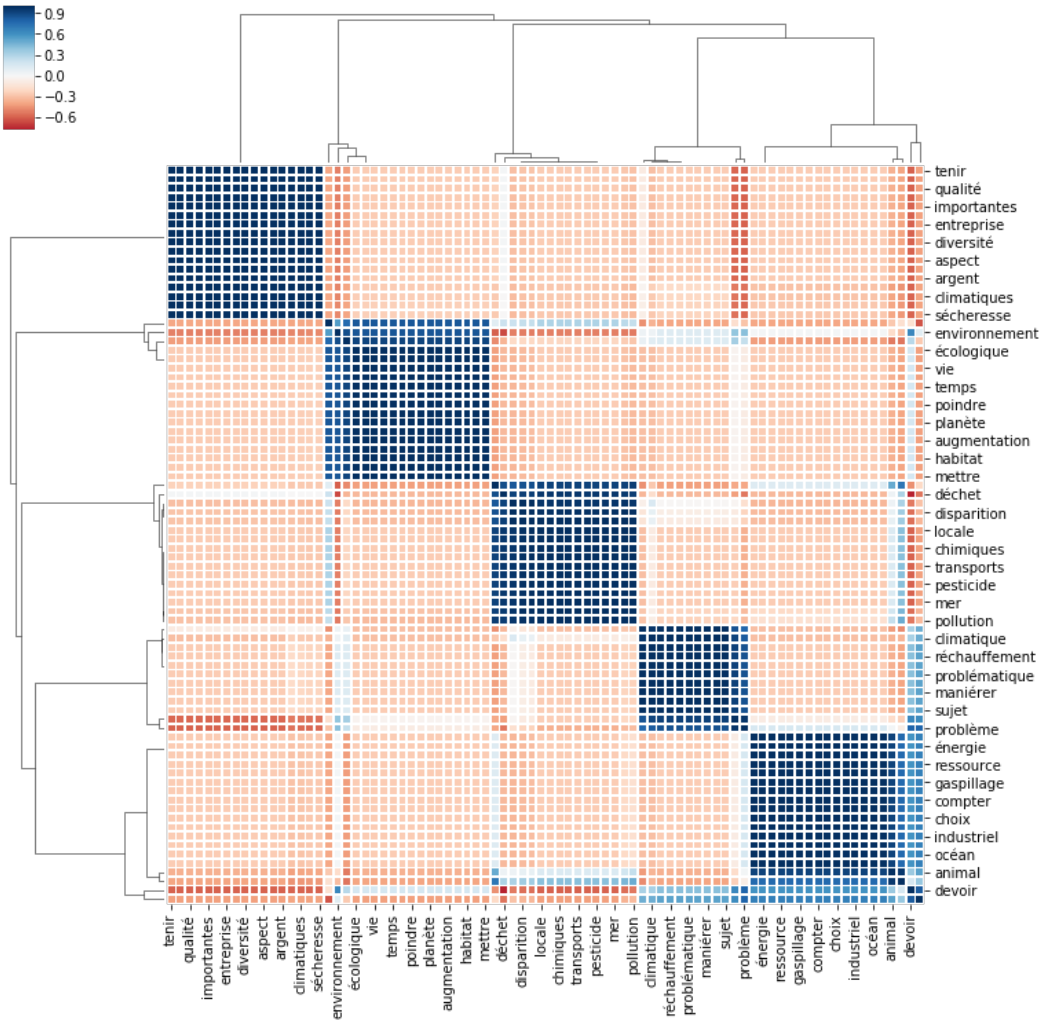


Figure 15: **Hierarchically-clustered heatmap in Bretagne**

We notice that the main topics are similar to the topics of Île-de-France, however we find the results are not as good as the former results through the performance of heatmap graph. They also much care the biodiversity, pollution and species extinction problem.

Then we will list the top 3 topics for the following regions: Loire, Centre,Alsace,Nord and Bourgogne.

```
[(0,
  '0.133*"biodiversité" + 0.126*"disparition" + 0.124*"espèce" + 0.122*"pollution" + 0.104*"air"'),
 (1,
  '0.025*"problème" + 0.016*"lier" + 0.015*"importer" + 0.010*"pouvoir" + 0.009*"ensemble"'),
 (2,
  '0.183*"dérèglement" + 0.179*"climatiques" + 0.175*"croire" + 0.175*"sécheresse" + 0.003*"économie"')]
```

Figure 16: **Top 3 topics in Loire**

```
[(0,
  '0.121*"biodiversité" + 0.119*"espèce" + 0.117*"disparition" + 0.103*"pollution" + 0.099*"air"'),
 (1,
  '0.151*"climatiques" + 0.151*"dérèglement" + 0.147*"croire" + 0.145*"sécheresse" + 0.022*"pollution"'),
 (2,
  '0.024*"problème" + 0.017*"importer" + 0.017*"pollution" + 0.011*"lier" + 0.008*"sol"')]
```

Figure 17: **Top 3 topics in Centre**

```
[(0,
  '0.166*"pollution" + 0.143*"air" + 0.013*"eau" + 0.012*"déchet" + 0.007*"terrer"'),
 (1,
  '0.178*"climatiques" + 0.178*"dérèglement" + 0.171*"croire" + 0.171*"sécheresse" + 0.008*"importer"'),
 (2,
  '0.135*"biodiversité" + 0.132*"espèce" + 0.131*"disparition" + 0.018*"lier" + 0.015*"problème"')]
```

Figure 18: **Top 3 topics in Alsace**

```
[(0,
  '0.137*"biodiversité" + 0.132*"espèce" + 0.131*"disparition" + 0.006*"pollution" + 0.006*"environnement"'),
 (1,
  '0.201*"pollution" + 0.179*"air" + 0.013*"eau" + 0.013*"climatique" + 0.012*"biodiversité"'),
 (2,
  '0.148*"dérèglement" + 0.141*"climatiques" + 0.137*"croire" + 0.136*"sécheresse" + 0.013*"lier"')]
```

Figure 19: **Top 3 topics in Nord**

4 Clustering analysis

4.1 Word2Vec based K-means clustering

In this part, we try to use Word2Vec based K-means method to find out the underlying topics of the answers in Grand Débat.

Word2Vec is one of the popular methods in language modeling and feature learning techniques in natural language processing (NLP). This method is used to create word embeddings in machine learning whenever we need vector representation of data.

In our project, we use pre-trained french word embedding result, which is trained on frWac corpus using Skip-gram method. The dimension of word vector is 700. The reason why we choose Skip-gram result instead of CBOW is that Skip-gram is better for infrequent words than CBOW. Because in CBOW the vectors from the context words are averaged before predicting the center word. In skip-gram there is no averaging of embedding vectors. It seems like the model can learn better representations for the rare words when their vectors are not averaged with the other context words in the process of making the predictions.

Next, having word vectors, we need to generate a vector representation for each answer. Here, we use the simple baseline method: Average of Word2Vec vectors. In the future, we plan to use more advanced method such as Average of Word2Vec vectors with TF-IDF or CNN based methods.

After transforming sentences into vectors, we apply the K-means method to perform clustering. For the initialization of K-means, we use the Kmeans++ algorithm, which has the advantage in achieving nearly optimal initialization. To find the best K parameter, we apply elbow method as shown in Fig 21.

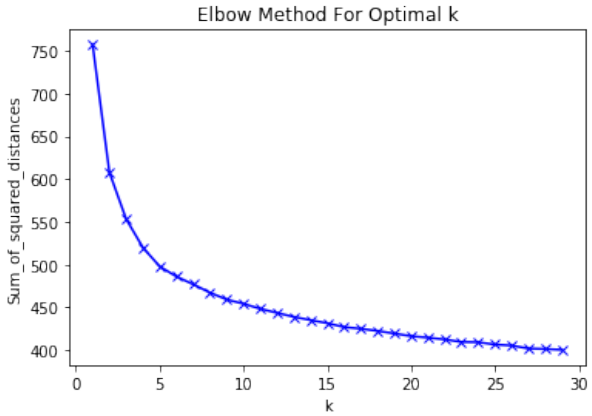


Figure 21: **Elbow method**

ACKNOWLEDGMENT

References

[1] Chuang J, Manning C D, Heer J. Termite: Visualization techniques for assessing textual topic models[C]//Proceedings of the international working conference on advanced visual interfaces. ACM, 2012: 74-77.

[2] Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics[C]//Proceedings of the workshop on interactive language learning, visualization, and interfaces. 2014: 63-70.

[3] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[4] Teh Y W, Newman D, Welling M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation[C]//Advances in neural information processing systems. 2007: 1353-1360.

[5] Krestel R, Fankhauser P, Nejdl W. Latent dirichlet allocation for tag recommendation[C]//Proceedings of the third ACM conference on Recommender systems. ACM, 2009: 61-68.