Python Programming for Data Scientists Final Report: Measuring the Political Motivation behind French Government's Endorsement for Electric Vehicles

Xiaoyan HU
x.hu@mpp.hertie-school.org

Abstract

In 2013, Bradley W. Lane and others identified two motivations for why governments seek to promote the electric car: risk management in response to ecological or energy concern versus industrial policy seeking an economic upgrade. Lane's framework of analysis provided useful guidelines, but it has been seven years and two government changes since their research. It is time to apply the theory on more recent policy texts to update interpretation of the intention of this French government on the French electric vehicle (EV) industry.

Instead of repeating the conventional way of generating policy positions by clicking, reading and downloading all the search result on the open web, this team project wishes to draw inspiration from computer sciences and explore a more efficient way of collecting and classifying text. After reviewing some recent case studies, this project decided to adopt Scrapy as the web scraping framework for collecting government publications from its official website and then process the text corpus using Latent Dirichlet Allocation (LDA).

From the 159 articles scraped from the government website, our results echoes Lane's conclusion in 2013. The political motivation of the French government behind its electric vehicle policy remains mixed judging from publications of this government. However, comparing to Lane's half-half conclusion, this project believes that the eager for industrial transition upgrade outweighs environmental concerns. The fact that écologie/écologique (ecology/ecological)" can be observed in many top-ranking topics possibly implies that environment issues is associated with many topics and thus is maybe a more prevalent concern of the French government. However, on concrete working projects, policymakers in France probably still let national industrial transition occupy much of their mind.

From the policy evaluation methodology perspective, the experiment carried out in this project successfully points out an alternative to traditional human reading, which is gaining trend now in social sciences. Both Scrapy and LDA are easily scalable and can be adapted to future research of a similar nature with a much bigger corpus. However, there is still much room for improvement in this project. Future work will concentrate on how to improve the coherence score of the model and interpreting the quantitative results with a more accurate narrative.

1. Introduction[1]

The French government proposed an ambitious plan in promoting the electric vehicle industry. It aims at multiplying the sales of electric vehicles (EV) by five times between 2018 and 2022. To boost the diffusion of electric vehicles, France employed various policy tools ranging from direct public subvention to maintaining car factories, RD subsidies, to consumer incentives, and parking fee exemptions. State leaders such as Emmanuel Macron and Édouard Phillippe presented in international conferences and visited production site, sending supporting signals to the EV market.

At first glance, one may argue that the close relationship between carbon emission reduction and electric vehicle promotion should infer that the French government push on this industry is mainly an environmental concern. After all, the first government plan on "véhicules décarbonés" (decarbonized vehicles) in the 21st century was raised by the French president of that time, Nicolas Sarkozy, as a response to the Grenelle Agreement. Full deployment of electric vehicles yields promising a picture for harnessing the speed of global warming. However, text analysis of more recent politician's speech and government policy papers confirm that the intention of financial instruments also includes enforcing the competitiveness of the French automobile industry, such as the commission letter from Édouard Phillippe on a government consulting report. In the engagement letter of a government-commissioned report, French

[1] GitHub page link:`https://github.com/Xiaoyan-Adele/Political-Motivation-behind-French-EV-Promotion`

premier minister Edouard Phillippe wrote: ”it is clear that my industry is suffering a degree of erosion, and that the degradation of my trade balance in this sector has accelerated. There are, however, many opportunities, with ongoing revolutions in the field of clean mobility and driverless and shared vehicles.” (Edouard Phillippe, 2018) [8].

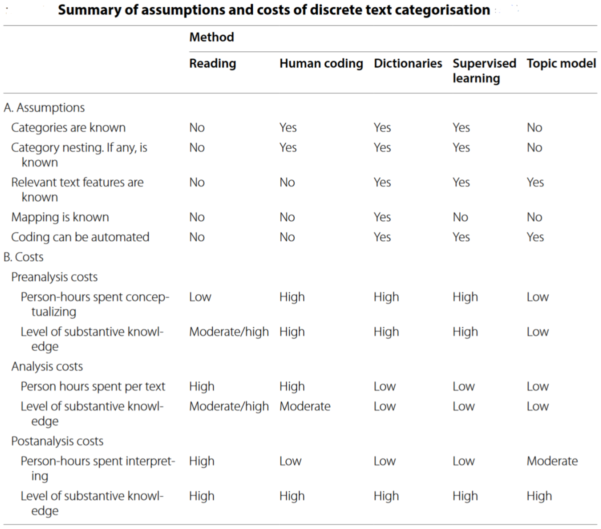
In 2013, Bradley W. Lane and others identified two motivations for why governments seek to promote the electric car: risk management in response to ecological or energy concern versus industrial policy seeking an economic upgrade. By reading government documents, Lane concluded that France was an intermediate case with a substantial blend of industrial policy and risk management [6].

This intuitive summary was based on the human reading comprehension skill, and many policy pieces of research rely on this method when reviewing past literature. However, this project wished to draw inspiration from computer sciences and explore a more efficient way of collecting text and a more objective manner of analyzing them. With this motivation, this project developed into two sub parts: 1) data collection using web scrapy framework Scrapy; 2) topic modelling using unsupervised learning method inspired by the application of Latent Dirichlet Allocation (LDA) in natural language processing. It found out that although the motivation of this French government remained mixed, policymakers valued industrial transition more when they designed concrete plans.

2. Related Work

This project hsa consulted a number of case studies in order to build its ow web scraper. There are multiple packages available online that allow us to scrape data from the open web such as Scrapy, Selenium, and BeautifulSoup. Scrapy is an extensive architecture that can manage the whole scraping task within its framework. However, it is not the easiest to master for beginners. This project invested considerable efforts to understand and model one from scratch. BeautifulSoup is said to be a much more user-friendly package and can manage just as well as Scrapy when it comes to simple tasks. However, BeautifulSoup often requires specific modules to support its operation. Selenium is not designed for web scraping initially. It is a popular package when it comes to browser automation and extracts updating data . After completing a course on web scraping on DataCamp, this project finds it more comfortable to use Scrapy and thus Scrapy becomes the library for constructing this project.

The official website of Scrapy provided an excellent introduction to the rationale behind its framework. The Medium article *A Minimalist End-to-End Scrapy Tutorial* written by Harry Wang [5] guided much of the self-exploration in constructing this project. I found Wang’s demonstration on how to follow URLs to extract data from each corresponding page particularly helpful and relevant.



Summary of assumptions and costs of discrete text categorisation

	Method				
	Reading	Human coding	Dictionaries	Supervised learning	Topic model
A. Assumptions					
Categories are known	No	Yes	Yes	Yes	No
Category nesting. If any, is known	No	Yes	Yes	Yes	No
Relevant text features are known	No	No	Yes	Yes	Yes
Mapping is known	No	No	Yes	No	No
Coding can be automated	No	No	Yes	Yes	Yes
B. Costs					
Preanalysis costs					
Person-hours spent conceptualizing	Low	High	High	High	Low
Level of substantive knowledge	Moderate/high	High	High	High	Low
Analysis costs					
Person hours spent per text	High	High	Low	Low	Low
Level of substantive knowledge	Moderate/high	Moderate	Low	Low	Low
Postanalysis costs					
Person-hours spent interpreting	High	Low	Low	Low	Moderate
Level of substantive knowledge	High	High	High	High	High

Figure 1. Summary of assumptions and costs of discrete text categorisation by Asmussen and Møller (2019)

Collecting a considerable-sized corpus with a few lines of python code is not the end goal of this project. Eventually, we want to explore topic generation method other than personal intuitive reading. In natural language processing, this research puzzle belongs to the domain of topic modelling. In a summary table presented by Asmussen and Møller (2019) [3] on the assumptions and costs of different text categorization method, topic modelling involved rather little human comprehension efforts and was easily scalable (see figure 1).

In the 2000s, topic modelling has already attracted the attention of academia on its potential application in social sciences (Ramage et al. 2009) [12]. Ramage and his co-authors focused mainly on Latent Dirichlet Allocation (LDA), which later became the most used, state-of-the-art and most straightforward method. Muskan and Mukesh (2016) [4] described its potentials in detection of emerging event in social multimedia. Nikolenko et al. (2017) [9] showed how LDA could be complied with qualitative studies by adapting into a semi-supervised approach on the case of Russian LiveJournal dataset aimed at ethnicity discourse analysis. Prabhsimran et al. (2019) [14] discussed how LDA could be combined with other techniques such as cloud computing, sentiment analysis to monitor and control government policies. Maybe outside the social sciences domain, a more vibrant topic modelling user community locates in the private sector. There are ample use cases of how to do business intelligence with LDA on consumers reviews and due diligence reports (e.g. Marcello et al. on the hospitality and tourism sector (2018) [7]).

Besides exploring the potentials of LDA on traditionally qualitative research domains, this project also consulted several practical tutorials and code books. Shashank Kapadia demonstrated the LDA in great details with the sklearn library. Due to the language nature of the scraped text, this project also investigated several topic modelling projects in French on GitHub (Abbes, 2019 [1] ; Pitanga, 2019 [10]). Since a dominant number of topic modelling projects are in English, these two repositories gave much inspiration on how to preprocess the corpus properly and convinced me to use gensim library instead of sklearn to build the topic model of this project.

3. Proposed Method

This project proposed to adopt Scrapy as the web scraping framework for collecting data and LDA for generating under-observed topics from the corpus collected.

3.1. Web Scraping

This project decided to scrape the French government website (https://www.gouvernement.fr). First, it is representative. This site is the official platform for publishing opinion piece, press release and legislation record of the French prime minister office. Thus it represents the official standing of the French government. Secondly, comparing to the website of the French president, prime minister website yields more content during the exploratory research, and it is much more up-to-date as it focuses on the position of the current administration.

It was vital to decide which keyword to use for navigating the search engine of the target websites (https://www.gouvernement.fr/search/). While being able to extract the maximum information possible, this keyword also needed to remain neutral so that the search result will not be too biased towards either ”industrial policy” or ”ecological concerns”. During the initial inspection, this project tested several keywords such as ”automobile”, ”PSA Renault”, ”véhicules électrique” and ”batterie (battery)”. Each yielded a slightly different assortment of results. According to the first result page of each word, ”automobile”, ”PSA Renault” seemed to be closer with industry probably due to the contextual usage of those words. Between ”batterie” and ”véhicules électrique”, although both seemed to be somewhat unbiased, this project decided to proceed with ”véhicules électrique” since it is closer to the core content that project’s author wanted to examine.

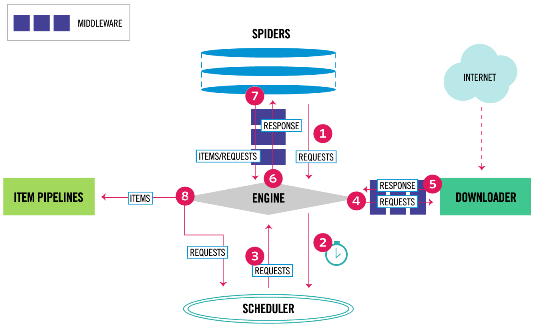


Figure 2. Scrapy Architecture

```
import scrapy
from scrapy import Request
from evarticles.items import EvArticleItems


class EvArticleSpider(scrapy.Spider):

    name = 'EvArticleSpider'
    allowed_domains = ["www.gouvernement.fr"]
    start_urls = ['https://www.gouvernement.fr/search/site/vehicules%2520electriques']

    def parse(self, response):
        # use xpath to parse elements

        def parse_content(response):
            content = response.xpath('//*[@id="block-system-main"]/div/div')
            content = content.xpath('string(.)').extract()[0]
            item = EvArticleItems()
            item['content'] = content
            item['title'] = ''
            item['link'] = ''
            item['publication'] = ''
            yield item

        body = response.xpath('//*[@id="block-system-main"]/div/div/ol/li')
        for url in body:
            #get the content of the url link
            content_url = url.xpath('div[2]/h2/a/@href').get()
            self.logger.info('scrape article content')
            yield response.follow(content_url, callback=parse_content)

        # next page
        # //*[@id="block-system-main"]/div/div/div[2]/ul/li[7]/a
        pagers = response.xpath('//*[@id="block-system-main"]/div/div/div[2]/ul/li')
        next_page = pagers[-1]
        if next_page.xpath('a') and 'suivant' in next_page.xpath('a/text()').extract()[0]:
            next_url = "https://" + self.allowed_domains[0] + next_page.xpath('a/@href').extract()[0]
            yield Request(url=next_url)
```

Figure 3. Code Design for URL Content Extraction

The diagram (figure 2) shows an overview of the Scrapy architecture. This project invested more efforts in inserting an auto page-turner. After writing a request to get the next page and used a call back function to call the same parse function to get the URL from the new page, this project combined the page turning and URL matching code with an additional auto-following code which instructed the spider to crawl the individual content page. When scraping the articles, I described the xpath in a way that it only encompassed the main body of the text while leaving out titles and summaries. The following code exert illustrates the python mechanism I just described:

3.2. Topic Modeling

LDA is an unsupervised, probabilistic modelling method which extracts topics from a collection of papers. A topic is defined as a distribution over a fixed vocabulary. LDA analyses the words in each paper and calculates the joint probability distribution between the observed (words in the paper) and the unobserved (the hidden structure of topics) (Asmussen and Møller 2019) [3]. On a practical level, topic modelling starts with text preprocessing. In this project, I mixed a series of corpus preprocessing libraries (e.g. re, nltk, and spacy) according to the characteristics of our dataset. Besides standard practices such as transforming to lowercase, removing numbers, splitting, deleting punctuations, and lemmatization, this project took into consideration that it would apply the unsupervised learning model on a corpus that had excessive phrases such as ”n”, and the text was in French. Thus, I adjusted the stop words accordingly. When constructing word dictionaries, this project switched from the initially chosen sklearn library to gensim, since I did not figure out how to build the word dictionary in French with sklearn.

The code implementation of LDA topic modelling can be described with the following bullet points. More detialed informtaion can be seen in my github repository:

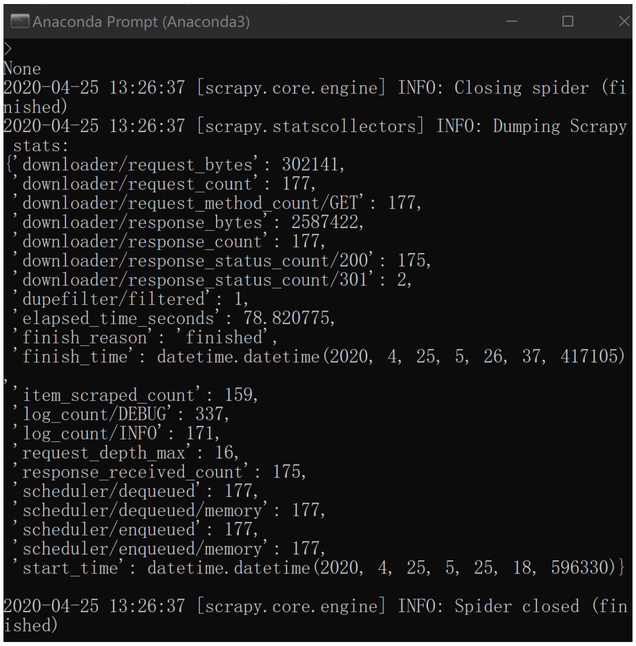
- Library import;
- Data cleaning;
- Preparing data for LDA analysis;
- Model training and result visualization
- Model tuning and final results presentation

This project used the code written by Selva Prabhakaran on *Topic Modeling with Gensim (Python)* [11] as the template and built our own LDA model based on it. I visualized the result by building an interactive chart with pyLDAvis package and computed the coherence score to get an easy-to-read judgement on the quality of the first LDA model. pyLDAvis is a powerful library to build web-based interactive visualization of topics estimated with LDA. Built initially on R and D3, it also has a python version which is the one I used for our project. pyLDAvis plots the topics as circles in the two-dimensional plane whose centres are determined by computing the distance between topics, and then by using multidimensional scaling to project the inter-topic distances onto two dimensions. pyLDAvis encodes each topic’s overall prevalence using the areas of the circles, where it sorts the topics in decreasing order of prevalence (Sievert and Shirley, 2014) [13].

4. Experiments

4.1. Web Scraping with Scrapy

Using the spider this project set up, we scraped 159 items from the French government website and exported it to a CSV file sized 302141 bytes. In total, the process took 78.82 seconds. Although the learning and adjusting process took a considerable amount of time of the author, the spider performance was satisfactory. Figure 4 shows the diary of the spider code execution.



```
>
None
2020-04-25 13:26:37 [scrapy.core.engine] INFO: Closing spider (fi
nished)
2020-04-25 13:26:37 [scrapy.statscollectors] INFO: Dumping Scrapy
 stats:
{’downloader/request_bytes’: 302141,
 ’downloader/request_count’: 177,
 ’downloader/request_method_count/GET’: 177,
 ’downloader/response_bytes’: 2587422,
 ’downloader/response_count’: 177,
 ’downloader/response_status_count/200’: 175,
 ’downloader/response_status_count/301’: 2,
 ’dupefilter/filtered’: 1,
 ’elapsed_time_seconds’: 78.820775,
 ’finish_reason’: ’finished’,
 ’finish_time’: datetime.datetime(2020, 4, 25, 5, 26, 37, 417105)
,
 ’item_scraped_count’: 159,
 ’log_count/DEBUG’: 337,
 ’log_count/INFO’: 171,
 ’request_depth_max’: 16,
 ’response_received_count’: 175,
 ’scheduler/dequeued’: 177,
 ’scheduler/dequeued/memory’: 177,
 ’scheduler/enqueued’: 177,
 ’scheduler/enqueued/memory’: 177,
 ’start_time’: datetime.datetime(2020, 4, 25, 5, 25, 18, 596330)}

2020-04-25 13:26:37 [scrapy.core.engine] INFO: Spider closed (fin
ished)
```

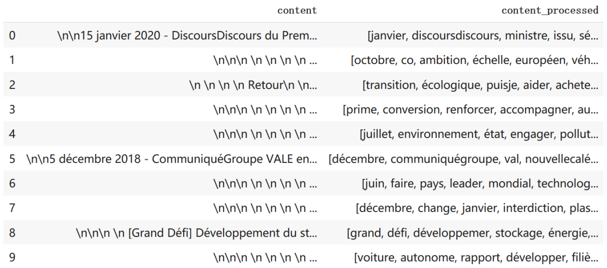
Figure 4. Spider Crawling Result

Results Analysis: After glancing through the scraped CSV content, I realized that this file could not be fed into the LDA model directly. Although during the xpath adjustment, I described it in a way to avoid some useless but recurrent phrases text such as the social network sharing buttons or titles and teasers, I discovered that in the text, there was still some highly repetitive but useless content such as press contact, email addresses or telephone numbers. Moreover, some search results were by nature not compatible with our LDA modelling such as prime minister official visit agendas. Although in the meeting note, keywords were mentioned, the rest of the text was purely about his itinerary of the day. Lastly, since we could not search for articles with the two words ”véhicules électrique” combined only, the website also yielded results that contained only one of the two words. It might better to delete those less complete results however this project decided to keep the list. Human languages are not a rigid formula that only allows one unique pattern for the same conveying messages. Some web search results mentioned only ”électrique” but use ”voiture” instead of its synonym ”véhicules”. Therefore, deleting only ”one-word-matched” research results could miss relevant information for topic modeling.

4.2. LDA Topic Modeling

After taking out irrelevant text such as prime minister agenda and manually removing all the press contact information from the CSV file, we treated the text with natural language processing techniques. The following screenshot can give a good intuitive comparison of the preprocessed text and the processed ones (see figure 5).

Then we convert the word into vectors and tune the initial hyperparameters of the LDA model into 20 topics, automated the alpha number and set chuck into 100.



	content	content_processed
0	\n\n15 janvier 2020 - DiscoursDiscours du Prem...	[janvier, discoursdiscours, ministre, issu, sé...
1	\n\n\n \n \n \n \n ...	[octobre, co, ambition, échelle, européen, véh...
2	\n \n \n \n Retour\n \n...	[transition, écologique, puisje, aider, achete...
3	\n\n\n \n \n \n \n ...	[prime, conversion, renforcer, accompagner, au...
4	\n\n\n \n \n \n \n ...	[juillet, environnement, état, engager, pollut...
5	\n\n5 décembre 2018 - CommuniquéGroupe VALE en...	[décembre, communiquégroupe, val, nouvellecalé...
6	\n\n\n \n \n \n \n ...	[juin, faire, pays, leader, mondial, technolog...
7	\n\n\n \n \n \n \n ...	[décembre, change, janvier, interdiction, plas...
8	\n\n\n \n [Grand Défi] Développement du st...	[grand, défi, développemer, stockage, énergie,...
9	\n\n\n \n \n \n \n ...	[voiture, autonome, rapport, développer, filiè...

Figure 5. Text Preprocessing Comparison

Results Analysis: The initial LDA model had a coherence score of 0.4523, which was moderately low. As advised by Prabhakaran, this project also included a model optimization section. However, after the fine-tuning, I improved coherence score to 0.4784, which would e a qualified result for further analysis. I suspect that further improvement should take place in data preprocessing and probably involved human judgement and editing of sticky words (e.g. such as discoursdiscours in the preprocessing results comparison).

Figure 6 shows improved pyLDAvis representation. On the left panel, a global perspective of topics is provided by two features. The area of circles is proportional to the relative prevalence of the topics in the corpus. As we can see, topic one accounts for 49.7% of the tokens, which is also the biggest circle in the graph. In addition, left pane also shows inter-topic differences. The more distanced the two circles are, the more different the two topics are. In this project, topic with high proportion of the tokens are rather dispersed. The biggest topic, Topic One, stands along and clearly distinguishes itself from the rest of the topics. There are some overlapping between Topic Two and Topic Three. Topic Five and the rest rather close. It worth noting that in pyLDAvis representation as well as the topic-terms summary of LDA model, topics are only shown as index instead of a text title (e.g. industrial policy, environment concerns etc). It implies that LDA models still needs human interpretation to make sense of those topic numbers.

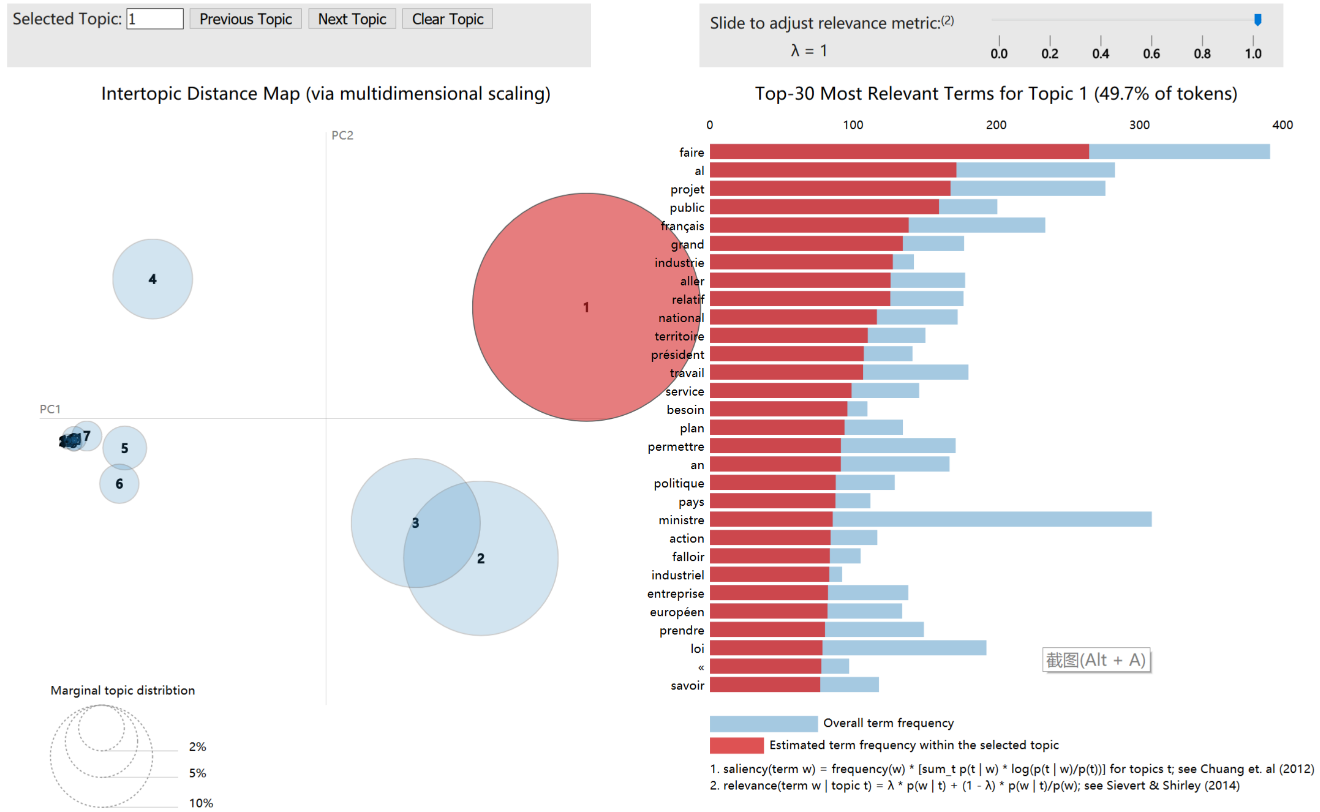
Moving to the right panel, the widths of the grey bars represent the corpus-wide frequencies of each term, and the widths of the red bars represent the topic-specific frequencies of each term. λ controls the ranking of terms in each topic. By comparing the widths of the red and grey bars for a given term, users can quickly understand whether a term is highly relevant to the selected topic be-cause of its lift (a high ratio of red to grey), or its probability (absolute width of red) When I set λ to 1, the terms are ranked solely by the probability of term for the selected topic. From Topic One, the relevant terms include mostly general government related expressions and seems to focus on action (e.g. the French verb ”faire” (to do, to make) ranks top). The first topic cluster has terms such as ”industrie (industry)”, ”grand”, ”industriel (industrial)”, ”enterprise”, ”français (French)” etc. This suggests that Topic One could be highly industry/business relevant.

Besides industry, this project is interested in exploring the French government position on ecology in respect to its endorsement of the electric vehicle industry. Topic two (22.9% of the tokens) seems to be rather environment related as it contains keywords such as ”transition”, ”énergie (energy)” and ”écologique (ecological)”. However, compared to an environmental concern, topic two mentioned ”mobilité (mobility)”, ”énergétique (energy-related)”, ”port”, or ”transport” too often. This project cannot conclude now that the French government does not associate environmental concerns with electric vehicles. Among the total of this corpus, terms such as ”air”, ”pollution/polluant (pollution/polluted)” or ”écologie/écologique (ecology/ecological)” can also be observed in many top-ranking topics. Those terms do not concentrate as much as other topics such as ”industrie (industry)”.

5. Analysis

The LDA model on search results from a French government website with the keyword ”véhicules électrique” reveals several findings that can help us to understand the political motivation behind state endorsement of the electric vehicle industry. As the most represented theme in the corpus, Topic One suggested that much of the narratives from the French government concentrates on ”français” ”industrie/industriel (industry)” which ”besoin (needs)” ”politique (policy)” and ”plan” at the ”national” level. Second to industrial policy, the French government also concerns about ”energie (energy)” ”transition” at the ”transport” sector and involves many ”loi (law)” and ”measure(s)”. This is linked to not only ”écologie” but also ”employeurs (employers)”, ”travail (work)” and ”entreprise”.

This project echoes Lane’s finding in 2013. The political motivation of the French government behind its electric vehicle policy remains mixed judging from publications from official website of this government. However, comparing to Lane’s half-half conclusion, this project believes that the eagerness for industrial upgrade outweighs environmental concerns. The fact that écologie/écologique (ecology/ecological)” can be observed in many top-ranking topics could imply that environment issues is associated with many topics and thus is maybe a more wide-spread concern of the French government. However, on concrete working projects, policymakers probably still let national industrial transition occupy much of their mind. At the same time, this project observes a number of ”action” related terms in the ”industry heavy” topic, Topic One. Those ”faire (to do, to make)”, ”project”, ”besoin (need)”, ”plan”, ”politique (policy)”, and ”falloir(should)” appear together with ”public”, ”national”, ”president”, ”ministre” may indicate a strong



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

2

Figure 6. pyLDAvis Visualization of the LDA model

presence/high voice of the French government in its industry performance. The dirigisme that long characterized the French economic model seems to still exist according to this project. The fact that French government often considers an ”européen (european)” aspect in its industry related topic suggests French government’s involvement in its economy could be no longer a national capacity to unilaterally determine economic practices and outcomes. French industrial policy can relate to an European level or possibly investing in the European scale (Ansaloni Smith, 2018) [2].

6. Conclusions and Reflection

On using topic modelling, Ramage et al. (2009) [12] once commented that: ”*the social sciences’ demand for methodological rigor must also be satisfied: free parameters’ choices must have warrants, topics and their usage should be characterizable, and results should be easily communicated visually.*” In this project, I successfully employed two computer science techniques on a traditionally human reading qualitative research. Both Scrapy and LDA are easily scalable and can be adapted to future research of a similar nature with a much bigger corpus. In fact, this idea is already shared by many researchers, especially when it comes to topic generation with social media posts. However, there is still much room for improvement in this project. Firstly, one can still discuss whether the keyword search is the best way to collect the most representative and comprehensive corpus for later analysis. Secondly, the final coherence score is far from the state-of-art benchmark (often believed to be on average 0.6-0.7). It could be that the preprocessing part can be improved including the manual deletion of some irrelevant text.

One of the motivation for conducting this research is to explore the possibility of replacing human comprehension with machine learning. From this perspective, this project achieved its goal. I realize that domain-specific knowledge is still indispensable. LDA results do not provide human-readable description on each topic but yield probabilistic relations of terms. Thus, we still depend on domain experts to label each topic number and with a theme and validate the reliability of the model.

Another example to demonstrate the importance of domain knowledge is data preprocessing. Firstly, the primary inspection of the CSV file revealed that although all of them were theoretically relevant to the research question of this project, not all of them were suitable to construct a topic model. Thus, I still had to process the text manually to filter out irrelevant text. Secondly, the rough output from web spider poses the question on how involved human should be when it comes to automated topic modelling. In this project, human knowledge is still required to select the right text for machine analysis, which diminish the efficiency gain that we can have from unsupervised learning models. Therefore, it worth asking the question of how much noise we can allow in our unsupervised model. It should be a case by case judgement or as long as the corpus is big enough, ”truth” will ultimately reveal itself.

7. Acknowledgements

I would like to thank many people for supporting me throughout my struggle with the difficulties encountered in this course and this project. My classmates Maximilian and Sebastian helped me to dial in the class via Skype in the first half of this semester when I was quarantined alone in Wuhan. Huy answered some of my question on piazza regarding the spider code. Vic Cheng, a friend in Wuhan, helped me debugging the request and parse code for article content extraction when I was completely lost in the Scrapy architecture. Without their encouragement and generous support, I would not be able to go as far by myself in this team project.

References

[1] Siwar Abbes. Topic modeling project, 2019. `https://github.com/siwaar/Topic_Modeling_Project`.

[2] Matthieu Ansaloni and Andy Smith. The neo-dirigiste production of french capitalism since 1980: the view from three major industries. *French Politics*, 16(2):154–178, 2018.

[3] Claus Boye Asmussen and Charles Møller. Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, 6(1):93, 2019.

[4] Muskan Garg and Mukesh Kumar. Review on event detection techniques in social multimedia. *Online Information Review*, 2016.

[5] Wang Harry. A minimalist end-to-end scrapy tutorial, 2019. `https://towardsdatascience.com/a-minimalist-end-to-end-scrapy-tutorial-part-i-11e350bcdec0`.

[6] Bradley W Lane, Natalie Messer-Betts, Devin Hartmann, Sanya Carley, Rachel M Krause, and John D Graham. Government promotion of the electric car: Risk management or industrial policy? *European Journal of Risk Regulation*, 4(2):227–245, 2013.

[7] Marcello Mariani, Rodolfo Baggio, Matthias Fuchs, and Wolfram Höepken. Business intelligence and big data in hospitality and tourism: a systematic literature review. *International Journal of Contemporary Hospitality Management*, 2018.

[8] Xavier Mosquet and Patrick Pélata. Reinforcing the attractiveness and competitiveness of france in tomorrow’s automotive industry and mobility, 2019.

[9] Sergey I Nikolenko, Sergei Koltcov, and Olessia Koltsova. Topic modelling for qualitative studies. *Journal of Information Science*, 43(1):88–102, 2017.

[10] Jana Pitange. topic modeling, 2019. `https://github.com/JanaPitanga/topic_modeling`.

[11] Selva Prabhakaran. Topic modeling with gensim (python). `https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#8tokenizewordsandcleanuptextusingsimple_preprocess`.

[12] Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D Manning, and Daniel A McFarland. Topic modeling for the social sciences. In *NIPS 2009 workshop on applications for topic models: text and beyond*, volume 5, page 27, 2009.

[13] Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.

[14] Prabhsimran Singh, Yogesh K Dwivedi, Karanjeet Singh Kahlon, Ravinder Singh Sawhney, Ali Abdallah Alalwan, and Nripendra P Rana. Smart monitoring and controlling of government policies using social media and cloud computing. *Information Systems Frontiers*, pages 1–23, 2019.