TEXT AND SOCIAL MEDIA ANALYTICS

MBA544B

CIA 2

FOSSIL FUEL INDUSTRY

Submitted by

VINEET NAIK – 2128130

CHRIST
UNIVERSITY
BENGALURU, INDIA
Declared as Deemed to be University under Section 3 of UGC Act 1956

Submitted to

Dr. Bindhia Joji

MBA PROGRAMME

SCHOOL OF BUSINESS AND MANAGEMENT

CHRIST (DEEMED TO BE UNIVERSITY), BANGALORE

December 2022

TABLE OF CONTENT

TOPIC MODELLING FOR THE FOSSIL FUEL INDUSTRY

1. INTRODUCTION

A method for locating hidden subjects in vast amounts of text is topic modeling. The method I'll be outlining falls under an unsupervised machine-learning algorithm. Latent Dirichlet Allocation (LDA) is the name of the algorithm, which is a component of Python's Gensim module. Topic Modelling is a part of the NLP.

Natural language processing is the process of cutting, extracting, and transforming languages utilized in thelibrary of nltk into new data so that we may gain useful insights from it. It only speaks the languages found in the library because it contains NLP-related items and cannot comprehend anything outside of what is
found there.
You must include that language in the existing library if you process on another language. For instance, NLP is used in email spam filtering to transform input data into new data that the system can comprehend. A model is then developed using this new data to predict if a message is a spam or not. NLP is mostly utilized in text processing, and it can be used to simplify a variety of jobs. e.g., chatbots, email filtering, speech recognition, social media monitoring, hiring and recruitment, language translation, and so forth. Recognizing words from subjects in a documentor data corpus is known as topic modeling. This is helpful since it is considerably more difficult and time-consuming to extract words from a document than it is to extract them from subjects that are present in the content.
In an oil & Gas industry with natural language processing, technicians can engage in a full dialog with machine applications to troubleshoot unexpected problems swiftly and accurately. This not only makes maintenance work safer and easier but can greatly reduce asset downtime due to unexpected issues as well.

2. DATA SET DESCRIPTION

The Dataset contains 50 articles about Fossil fuels. The data has been scraped online from various blogs and articles. The Dataset talks about various aspects of the fossil fuel industry. It includes topics such as fossil fuel consumption, trends in the fossil fuel industry, government regulations on fossil fuel consumption, the advantages of fossil fuels, the negative impact of fossil fuels on the environment and humans, etc.

3. PROBLEM STATEMENT

Excessive burning and consumption of fossil fuels are leading to global climate change and suggesting the automobile industry switch to electric vehicles. Researchers have analyzed and predicted India would soon be facing scorching heat waves that could break the human survivability limit. This is forcing governments all over the world to take regulatory actions over automobile and manufacturing industries in many countries. Oil/Gas industry (Fossil Fuel Industry) stakeholders must be up to date with the rising risk of regulations over the world to decide optimal production of fossil fuels.

1. METHODOLOGY

There is 'n' a number of different approaches that could be taken to address the problem statement of "How the dataset can be used to identify and analyze the recent trends, regulations, and sentiments of the people and government over the consumption of fossil fuels."

LDA - Topic modeling

Latent Dirichlet Allocation (LDA) is an unsupervised technique that allocates a value to each document based on the topic. Dirichlet is a probability distribution, while latent is another word for hidden. LDA views each document as a collection of topics and each topic as a collection of words. It goes over all of the topics and each word one by one. It will assign each word to a topic at random and evaluate how frequently the word appears in that topic with which other words.

Topic modeling is an unsupervised machine learning technique that can scan a collection of documents, find word and phrase patterns within them, and automatically cluster word groupings and related expressions that best define the collection of documents. The Dataset has blogs and articles about the fossil fuel industry, its adaption, challenges, etc. The topic modeling algorithm is applied to understand what are the topics which are discussed in all 50 blogs, which will be helpful for people to understand its advantages and implications.

Importing packages

The packages like Numpy, Pandas, Gensim, and spacy, which are necessary for running a topic modeling algorithm, are imported into python.
For Visualization pyLDAvis package is used.

With the aid of pyLDAvis, users can better understand the themes in a topic model that has been tailored to a corpus of text data. A fitted LDA topic model is used to extract data that the software then uses to inform an interactive web-based visualization.

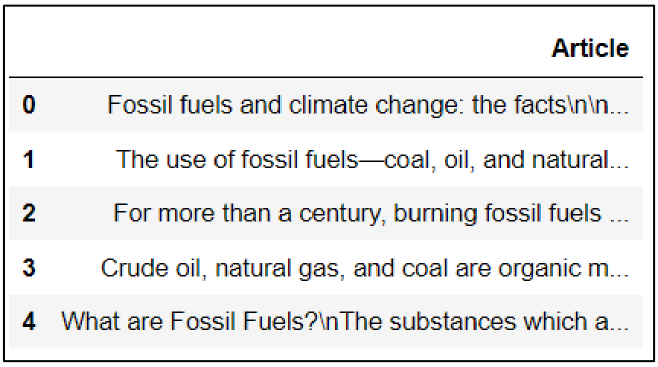
Gensim: is an open-source framework that uses contemporary statistical machine learning to do unsupervised topic modeling, document indexing, retrieval by similarity, and other natural language processing functions.

SpaCy: is a Python NLP library that is free and open source. It is intended for use in the development of information extraction and natural language comprehension systems.

Nltk: is a Python toolbox for working with natural language processing (NLP). It offers us a large number of test datasets for various text-processing packages. Tokenizing, parse tree visualization, and other operations can be accomplished with NLTK.

Importing Dataset

The Dataset is imported into a python file from the excel sheet. Below the given picture is the Dataset viewed from a python file. After importing the Dataset, it is viewed, and various text pre-processing techniques are applied.



	Article
0	Fossil fuels and climate change: the facts\n\n...
1	The use of fossil fuels—coal, oil, and natural...
2	For more than a century, burning fossil fuels ...
3	Crude oil, natural gas, and coal are organic m...
4	What are Fossil Fuels?\nThe substances which a...

Removal of stopwords

This is the most important phase because such words don't help define the themes.

Taking out the punctuation.

Although the topic modeling stage does not at all require punctuation, we did consider some special characters (.,?!) as suitable for our future work.

The concept underlying topic modeling is that topics, which are composed of words, are what makeup texts. Punctuation is not necessary for this sentence.

Tokenization

To construct a dictionary and document term matrix for the topic model, the text is broken down into a list of tokens.

The outcome is a list of the input text. The next step is to filter the tokens based on the POS Tags. Using NLTK, this function will tag the portions of speeches that correlate to each token in the corpus.

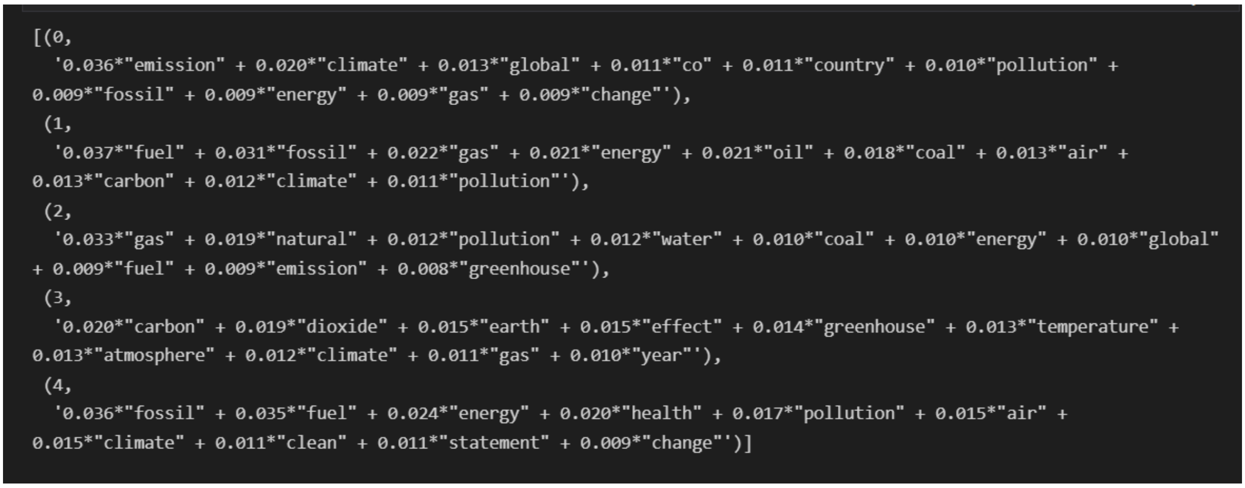
Text pre-processing

Count Vectorizer breaks down a sentence or any text into words by performing pre-processing tasks like converting all words to lowercase, thus removing special characters. Here tokenized input of data is used.

- **Max_df** is a function that removes the most probabilistic words in the document. For example, the, or, and, an, and so on. These words are removed from the documents since they don't add any value to the analysis.
- **Min_df** is a function that removes the words that occur only once or twice in the document.

Fit the LDA model

After setting the number of components as 5 at random, this is the output of topics generated. Every topic is distributed as the distribution over words.

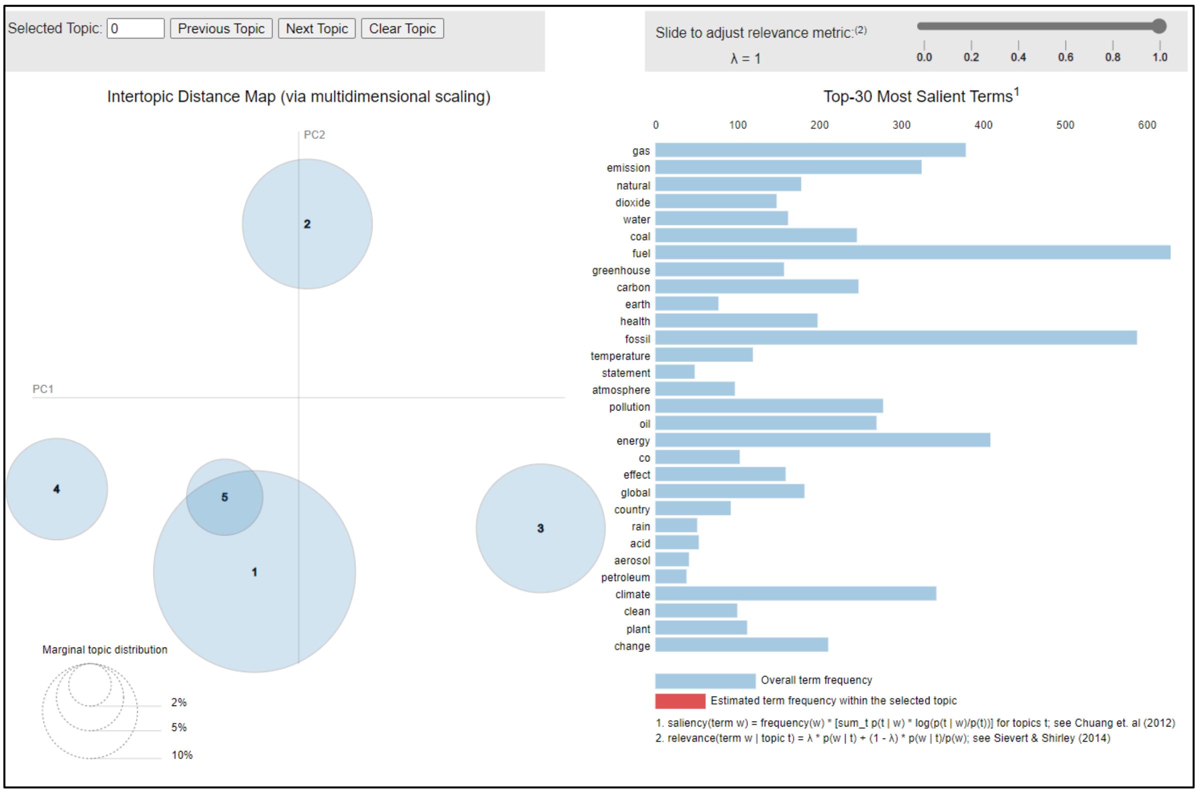


```
[(0,
  '0.036*"emission" + 0.020*"climate" + 0.013*"global" + 0.011*"co" + 0.011*"country" + 0.010*"pollution" +
0.009*"fossil" + 0.009*"energy" + 0.009*"gas" + 0.009*"change"'),
 (1,
  '0.037*"fuel" + 0.031*"fossil" + 0.022*"gas" + 0.021*"energy" + 0.021*"oil" + 0.018*"coal" + 0.013*"air" +
0.013*"carbon" + 0.012*"climate" + 0.011*"pollution"'),
 (2,
  '0.033*"gas" + 0.019*"natural" + 0.012*"pollution" + 0.012*"water" + 0.010*"coal" + 0.010*"energy" + 0.010*"global"
+ 0.009*"fuel" + 0.009*"emission" + 0.008*"greenhouse"'),
 (3,
  '0.020*"carbon" + 0.019*"dioxide" + 0.015*"earth" + 0.015*"effect" + 0.014*"greenhouse" + 0.013*"temperature" +
0.013*"atmosphere" + 0.012*"climate" + 0.011*"gas" + 0.010*"year"'),
 (4,
  '0.036*"fossil" + 0.035*"fuel" + 0.024*"energy" + 0.020*"health" + 0.017*"pollution" + 0.015*"air" +
0.015*"climate" + 0.011*"clean" + 0.011*"statement" + 0.009*"change"')]
```

Visualization of topics

The pyLDAvis package is used to create this interactive visualization dashboard. With the aid of pyLDAvis, users can better understand the themes in a topic model that has been tailored to a corpus of text data. A fitted LDA topic model is used to extract data that the software then uses to inform an interactive web-based visualization.

The graph here is created based on the intertopic Distance Map that, on interacting with any of the topics, returns the frequency of the top 30 most Salient terms in the bar graphs on the right side.



Selected Topic: 0

Slide to adjust relevance metric:(2)

λ = 1

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)

Model perplexity and coherence score

The relative distance between words within a topic is measured by coherence. The topic coherence score is always used to measure how well the topics are extracted when determining the optimal number of topics to be extracted using LDA. We can compare the goodness-of-fit of LDA models fit with varying numbers of topics to determine an appropriate number of topics. The perplexity of a held-out collection of documents can be used to assess the goodness-of-fit of an LDA model. The perplexity of the model reveals how well it describes a group of documents.

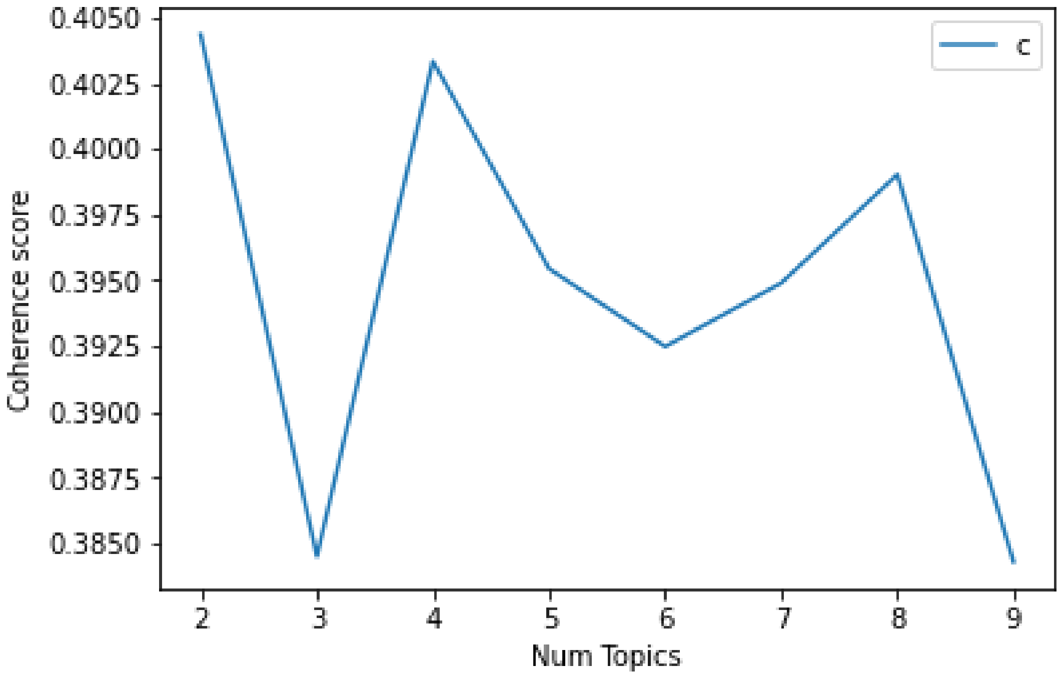
The model perplexity and coherence scores for our model are as below:

<u>Perplexity</u>: -6.34226697854453 65

<u>Coherence Score</u>: 0.39931751989069525

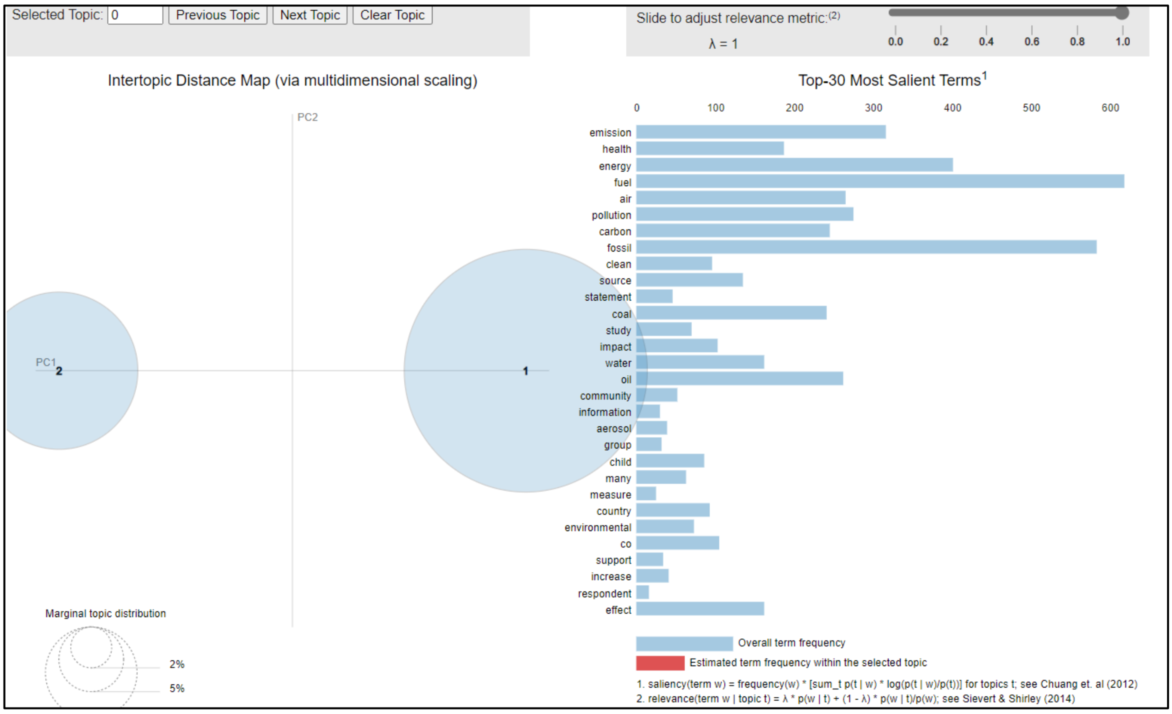
As the number of topics was selected at random, we cannot be sure if our model is optimal. Hence coherence score of various models at a different number of set topics has to be compared to get the optimal model.

Comparison of coherence score for different values of topics



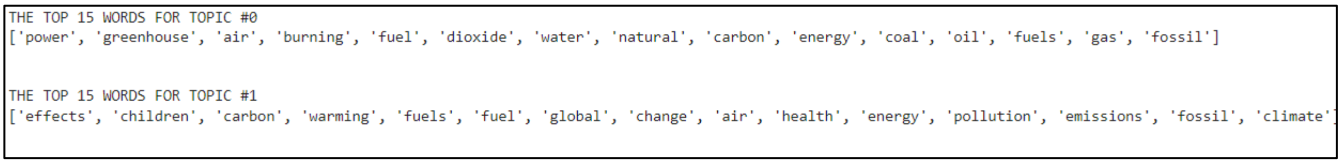
Through the graph, we can see that the coherence score is high for value=2. Hence the number of topics should be set at two and fitted with the LDA model.

New Model Visualization



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)

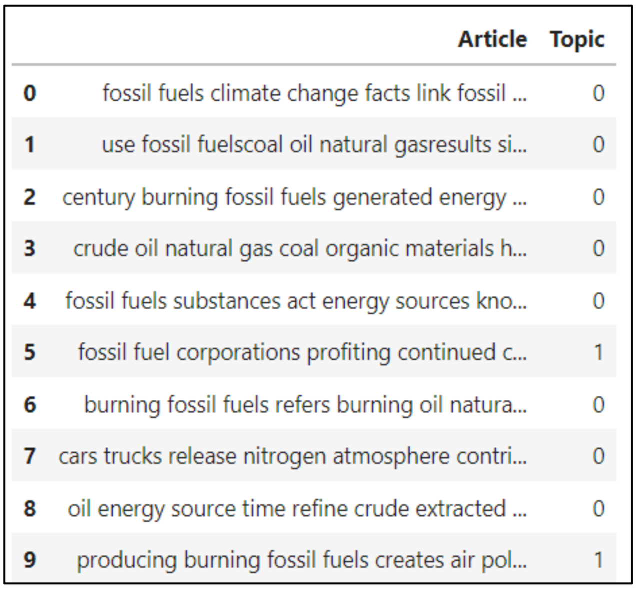
The top 15 words for each topic extracted and depicted from the model are shown below:



```
THE TOP 15 WORDS FOR TOPIC #0
['power', 'greenhouse', 'air', 'burning', 'fuel', 'dioxide', 'water', 'natural', 'carbon', 'energy', 'coal', 'oil', 'fuels', 'gas', 'fossil']


THE TOP 15 WORDS FOR TOPIC #1
['effects', 'children', 'carbon', 'warming', 'fuels', 'fuel', 'global', 'change', 'air', 'health', 'energy', 'pollution', 'emissions', 'fossil', 'climate'
```

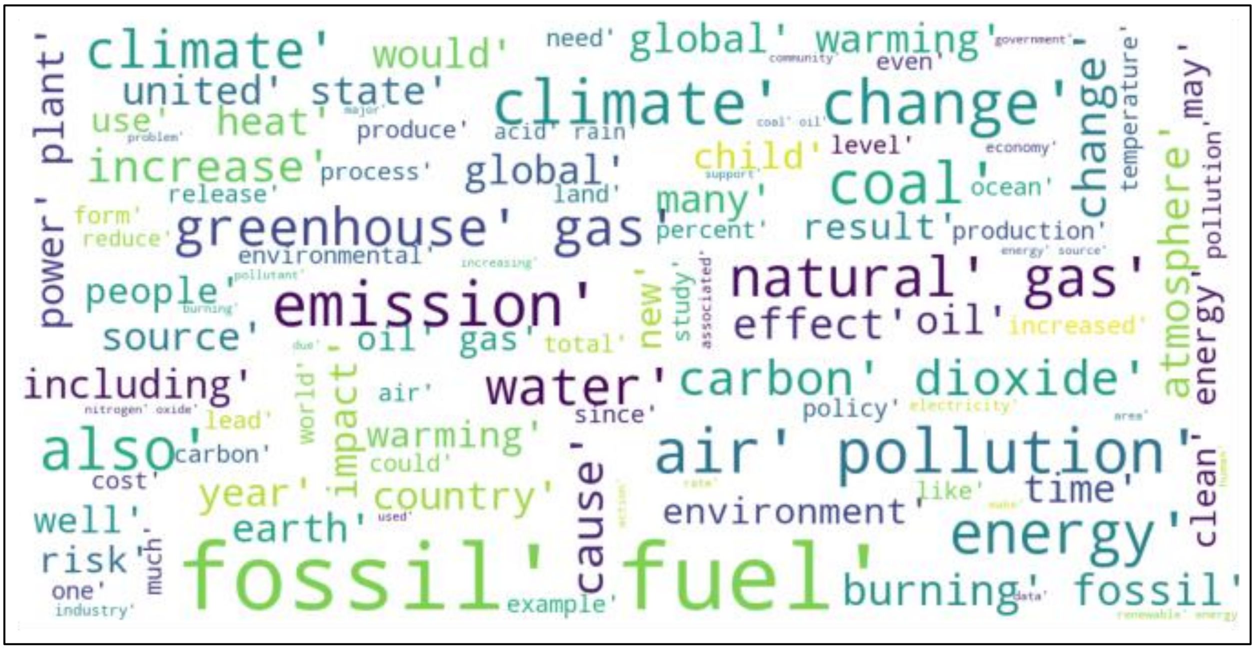
Lastly each article is assigned to either topic 1(coded as ‘0’) or topic2 (coded as ‘1’)



	Article	Topic
0	fossil fuels climate change facts link fossil ...	0
1	use fossil fuelscoal oil natural gasresults si...	0
2	century burning fossil fuels generated energy ...	0
3	crude oil natural gas coal organic materials h...	0
4	fossil fuels substances act energy sources kno...	0
5	fossil fuel corporations profiting continued c...	1
6	burning fossil fuels refers burning oil natura...	0
7	cars trucks release nitrogen atmosphere contri...	0
8	oil energy source time refine crude extracted ...	0
9	producing burning fossil fuels creates air pol...	1

WORDCLOUD

Wordcloud is a great way to represent text data. The size and color of each word that appearsin the wordcloud indicate its frequency or importance.



Here we can see that the terms associated with the Fuel industry are highlighted, which indicates that these words frequently occurred in the articles.

Results and discussion

From the above topic modeling analysis, we can conclude that the 2 topics that are extracted from the 50 articles are very useful for understanding the trends, and current affairs of the fossil fuel industry, its implementation, challenges, and benefits. It will be very helpful for internal stakeholders to know more about the latest trends happening in the industry. By being aware and up-to-date about the regulations, current affairs, and sentiments the stakeholders of the fossil fuel industry can use the insights for sentimental analysis, set production quantity for optimal storage and supply, be aware of the potential risks.

References

- https://pypi.org/project/gensim/
- https://towardsdatascience.com/topic-modelling-in-python-with-spacy-and-gensim-dc8f7748bdbf#:~:text=Topic%20Modelling%20is%20a%20technique,developed%20by%20Blei%20et%20al.
- https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#14computemodelperplexityandcoherencescore