# Topic Modelling

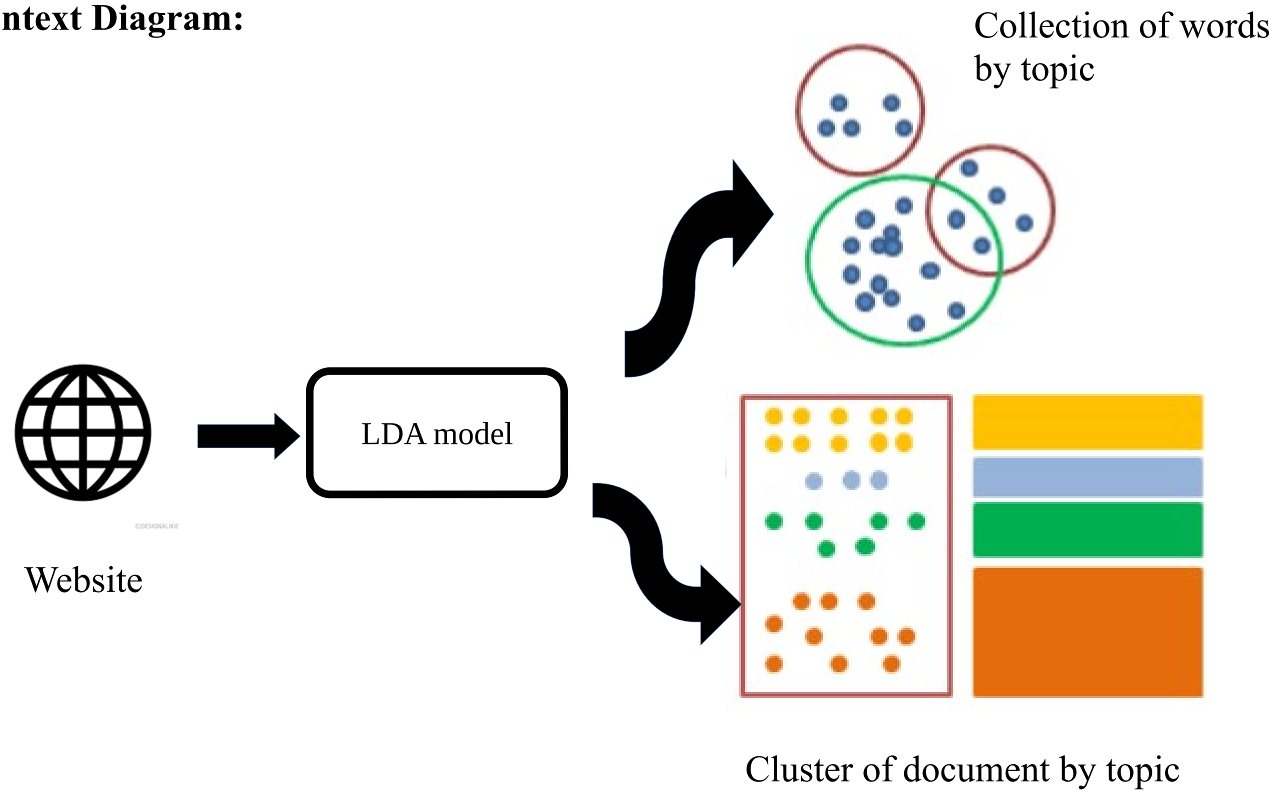
## Introduction:

Topic modelling is a technique used in natural language processing (NLP) to uncover hidden thematic structures in a collection of texts. This software report provides an overview and analysis of topic modelling performed on a web page using the given code snippet.

## Code Overview:

The provided code uses the Selenium library to scrape text from a webpage (https://www.studytoday.net/acid-rain/). The extracted text is then preprocessed and subjected to topic modelling using the gensim library. The code generates a visualization of the topics using pyLDAvis.

## Context Diagram:



## Software Used:

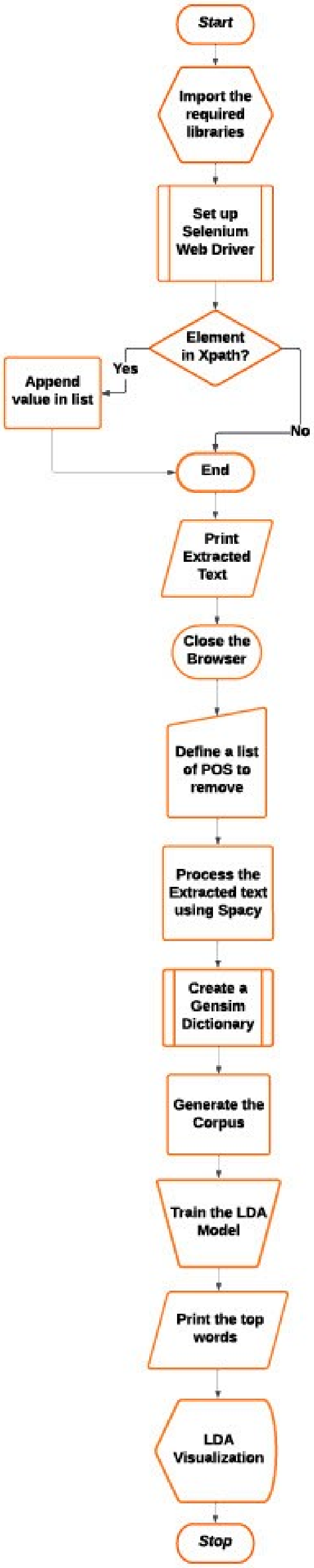
| Tools | Description |
|---|---|
| Python | A high-level, general-purpose programming language |
| Jupyter Notebook | A project to develop open-source software, open standards, and services for interactive computing across multiple programming languages |

| | |
|---|---|
| Selenium | Used for web scraping and interacting with web browsers |
| Pandas | Used for data manipulation and analysis |
| Matplotlib | Used for data visualization |
| Seaborn | Used for data visualization |
| Spacy | A library for natural language processing tasks |
| Gensim | A library for topic modelling and document similarity analysis. |
| pyLDAvis | A library for interactive topic model visualization |

**Algorithm:**

1. Import the required libraries: selenium, pandas, matplotlib.pyplot, seaborn, spacy, pyLDAvis.gensim_models, en_core_web_md, gensim, pickle.
2. Set up a Selenium WebDriver to control the web browser (Chrome) and open the desired website.
3. Find all elements on the page using XPath and store them in a list.
4. Iterate through the elements and extract the text from each element, excluding empty text.
5. Print the extracted text.
6. Close the browser.
7. Load the required libraries and models for natural language processing and topic modeling: en_core_web_md from spaCy, Dictionary and LdaMulticore from Gensim.
8. Define a list of POS tags to remove from the text.
9. Process the extracted text using spaCy to tokenize, lemmatize, and filter out unwanted POS tags.
10. Create a Gensim dictionary from the processed tokens and generate the document-term matrix (corpus).
11. Save the corpus and dictionary objects to disk.
12. Train an LDA (Latent Dirichlet Allocation) model using Gensim on the corpus, specifying the number of topics and passes.
13. Save the trained LDA model to disk.
14. Print the top words for each topic in the LDA model.
15. Load the saved dictionary, corpus, and LDA model from disk.
16. Prepare the LDA visualization using pyLDAvis and display it.

## Flowchart:


Start
Import the required libraries
Set up Selenium Web Driver
Element in Xpath?
Yes
Append value in list
No
End
Print Extracted Text
Close the Browser
Define a list of POS to remove
Process the Extracted text using Spacy
Create a Gensim Dictionary
Generate the Corpus
Train the LDA Model
Print the top words
LDA Visualization
Stop

**Working Code:**

```
from selenium import webdriver

from selenium import webdriver

# Create a new instance of the Chrome driver
browser = webdriver.Chrome()

# Open the website
browser.get("https://www.studytoday.net/acid-rain/")

# Get all elements on the page
elements = browser.find_elements("xpath", "//*")  # Using the correct method find_elements()

# Iterate through the elements and extract text
all_text = []
for element in elements:
    text = element.text.strip()
    if text:
        all_text.append(text)

# Print the extracted text
for text in all_text:
    print(text)

# Close the browser
browser.quit()

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import spacy
import pyLDAvis.gensim_models
pyLDAvis.enable_notebook()# Visualise inside a notebook
import en_core_web_md
from gensim.corpora.dictionary import Dictionary
from gensim.models import LdaMulticore
from gensim.models import CoherenceModel

# Our spaCy model:
nlp = en_core_web_md.load()
# Tags I want to remove from the text
removal= ['ADV','PRON','CCONJ','PUNCT','PART','DET','ADP','SPACE', 'NUM', 'SYM']
```

```
tokens = []
for summary in nlp.pipe(all_text[:]):
   proj_tok = [token.lemma_.lower() for token in summary if token.pos_ not in removal and not token.is_stop and token.is_alpha]
   tokens.append(proj_tok)

from gensim import corpora
dictionary = corpora.Dictionary(tokens)
corpus = [dictionary.doc2bow(text) for text in tokens]
import pickle
pickle.dump(corpus, open('corpus.pkl', 'wb'))
dictionary.save('dictionary.gensim')

import gensim
NUM_TOPICS = 5
ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics = NUM_TOPICS, id2word=dictionary, passes=15)
ldamodel.save('model5.gensim')
topics = ldamodel.print_topics(num_words=4)
for topic in topics:
    print(topic)

dictionary = gensim.corpora.Dictionary.load('dictionary.gensim')
corpus = pickle.load(open('corpus.pkl', 'rb'))
lda = gensim.models.ldamodel.LdaModel.load('model5.gensim')
import pyLDAvis.gensim
lda_display = pyLDAvis.gensim.prepare(lda, corpus, dictionary, sort_topics=False)
pyLDAvis.display(lda_display)
```

## Sample Input:

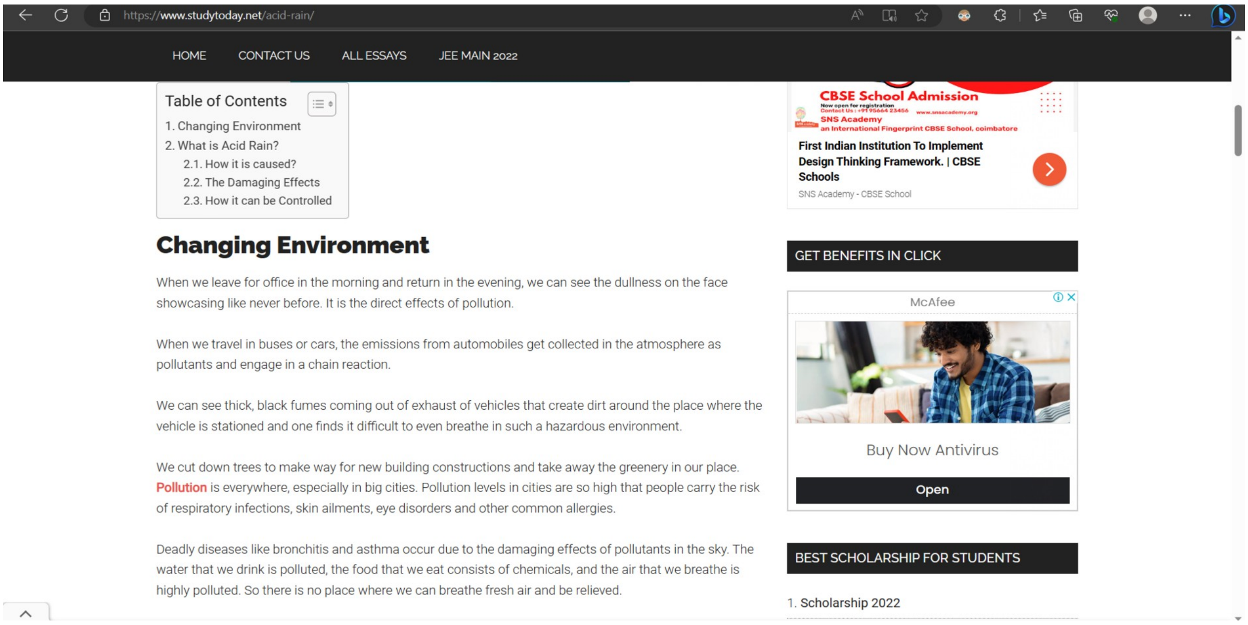


Table of Contents

### Changing Environment

When we leave for office in the morning and return in the evening, we can see the dullness on the face showcasing like never before. It is the direct effects of pollution.

When we travel in buses or cars, the emissions from automobiles get collected in the atmosphere as pollutants and engage in a chain reaction.

We can see thick, black fumes coming out of exhaust of vehicles that create dirt around the place where the vehicle is stationed and one finds it difficult to even breathe in such a hazardous environment.

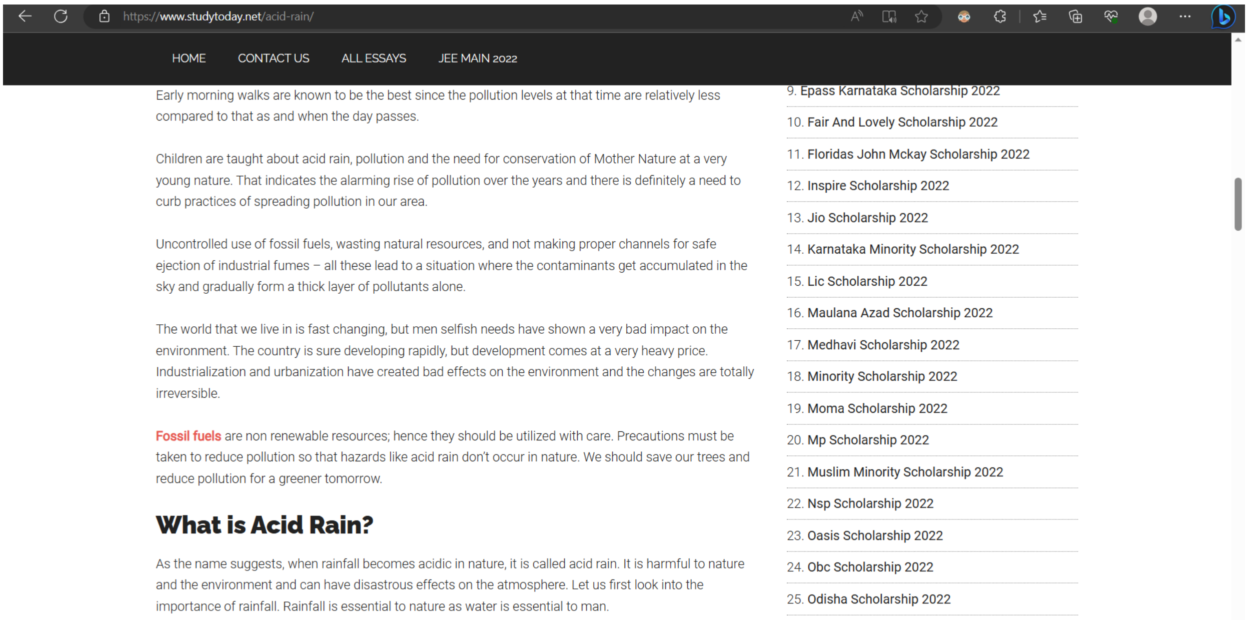
We cut down trees to make way for new building constructions and take away the greenery in our place. Pollution is everywhere, especially in big cities. Pollution levels in cities are so high that people carry the risk of respiratory infections, skin ailments, eye disorders and other common allergies.

Deadly diseases like bronchitis and asthma occur due to the damaging effects of pollutants in the sky. The water that we drink is polluted, the food that we eat consists of chemicals, and the air that we breathe is highly polluted. So there is no place where we can breathe fresh air and be relieved.

GET BENEFITS IN CLICK

BEST SCHOLARSHIP FOR STUDENTS

1. Scholarship 2022



Early morning walks are known to be the best since the pollution levels at that time are relatively less compared to that as and when the day passes.

Children are taught about acid rain, pollution and the need for conservation of Mother Nature at a very young nature. That indicates the alarming rise of pollution over the years and there is definitely a need to curb practices of spreading pollution in our area.

Uncontrolled use of fossil fuels, wasting natural resources, and not making proper channels for safe ejection of industrial fumes – all these lead to a situation where the contaminants get accumulated in the sky and gradually form a thick layer of pollutants alone.

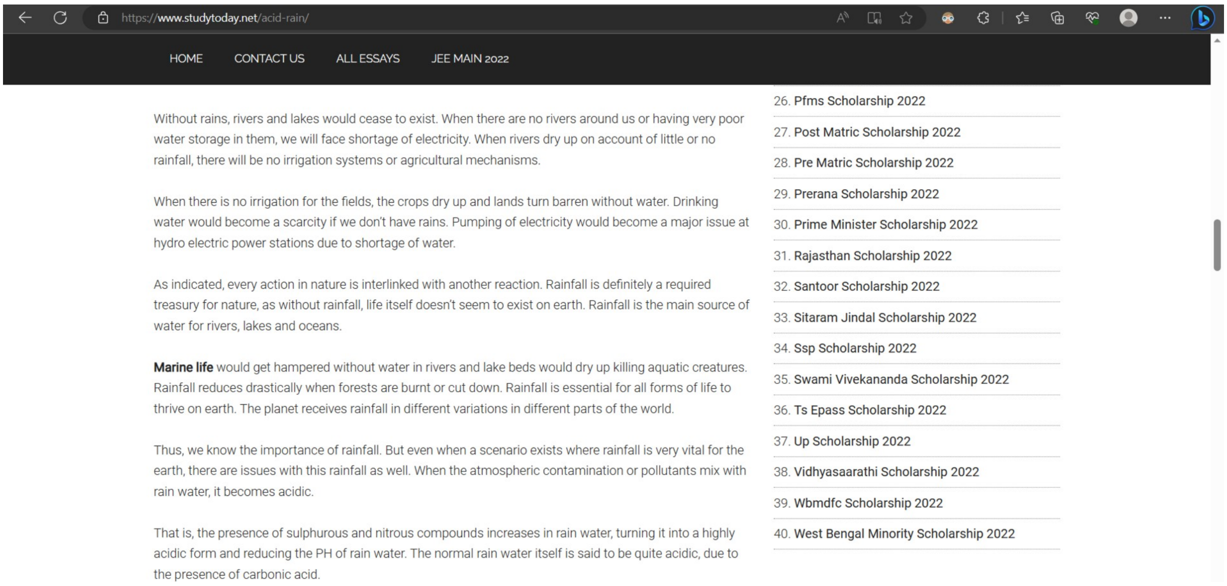
The world that we live in is fast changing, but men selfish needs have shown a very bad impact on the environment. The country is sure developing rapidly, but development comes at a very heavy price. Industrialization and urbanization have created bad effects on the environment and the changes are totally irreversible.

Fossil fuels are non renewable resources; hence they should be utilized with care. Precautions must be taken to reduce pollution so that hazards like acid rain don't occur in nature. We should save our trees and reduce pollution for a greener tomorrow.

## What is Acid Rain?

As the name suggests, when rainfall becomes acidic in nature, it is called acid rain. It is harmful to nature and the environment and can have disastrous effects on the atmosphere. Let us first look into the importance of rainfall. Rainfall is essential to nature as water is essential to man.

9. Epass Karnataka Scholarship 2022
10. Fair And Lovely Scholarship 2022
11. Floridas John Mckay Scholarship 2022
12. Inspire Scholarship 2022
13. Jio Scholarship 2022
14. Karnataka Minority Scholarship 2022
15. Lic Scholarship 2022
16. Maulana Azad Scholarship 2022
17. Medhavi Scholarship 2022
18. Minority Scholarship 2022
19. Moma Scholarship 2022
20. Mp Scholarship 2022
21. Muslim Minority Scholarship 2022
22. Nsp Scholarship 2022
23. Oasis Scholarship 2022
24. Obc Scholarship 2022
25. Odisha Scholarship 2022



Without rains, rivers and lakes would cease to exist. When there are no rivers around us or having very poor water storage in them, we will face shortage of electricity. When rivers dry up on account of little or no rainfall, there will be no irrigation systems or agricultural mechanisms.

When there is no irrigation for the fields, the crops dry up and lands turn barren without water. Drinking water would become a scarcity if we don't have rains. Pumping of electricity would become a major issue at hydro electric power stations due to shortage of water.

As indicated, every action in nature is interlinked with another reaction. Rainfall is definitely a required treasury for nature, as without rainfall, life itself doesn't seem to exist on earth. Rainfall is the main source of water for rivers, lakes and oceans.

**Marine life** would get hampered without water in rivers and lake beds would dry up killing aquatic creatures. Rainfall reduces drastically when forests are burnt or cut down. Rainfall is essential for all forms of life to thrive on earth. The planet receives rainfall in different variations in different parts of the world.

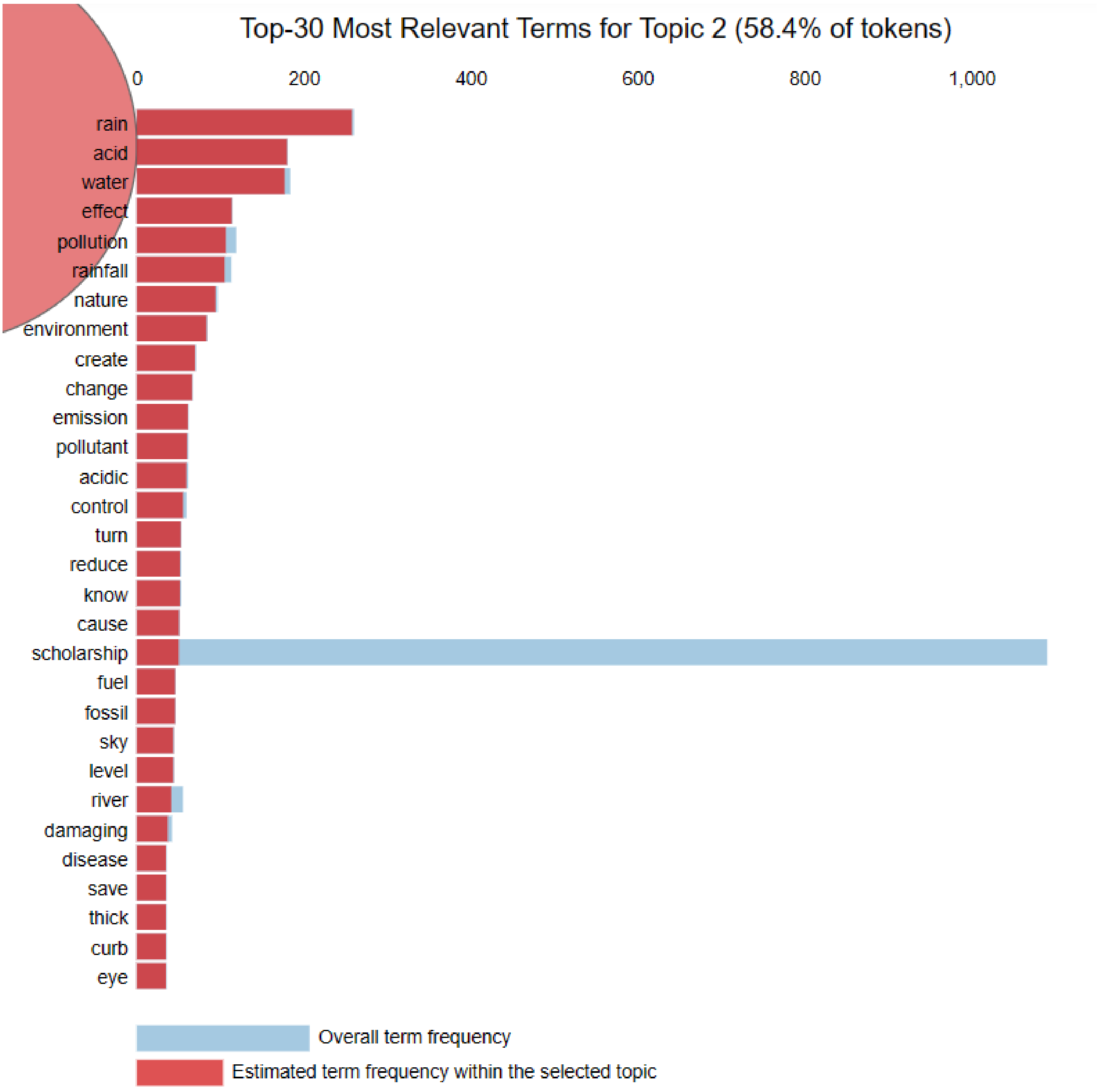
Thus, we know the importance of rainfall. But even when a scenario exists where rainfall is very vital for the earth, there are issues with this rainfall as well. When the atmospheric contamination or pollutants mix with rain water, it becomes acidic.

That is, the presence of sulphurous and nitrous compounds increases in rain water, turning it into a highly acidic form and reducing the PH of rain water. The normal rain water itself is said to be quite acidic, due to the presence of carbonic acid.

26. Pfms Scholarship 2022
27. Post Matric Scholarship 2022
28. Pre Matric Scholarship 2022
29. Prerana Scholarship 2022
30. Prime Minister Scholarship 2022
31. Rajasthan Scholarship 2022
32. Santoor Scholarship 2022
33. Sitaram Jindal Scholarship 2022
34. Ssp Scholarship 2022
35. Swami Vivekananda Scholarship 2022
36. Ts Epass Scholarship 2022
37. Up Scholarship 2022
38. Vidhyasaarathi Scholarship 2022
39. Wbmdfc Scholarship 2022
40. West Bengal Minority Scholarship 2022

## Output:

```
(0, '0.359*"scholarship" + 0.036*"minority" + 0.018*"karnataka" + 0.018*"matric"')
(1, '0.039*"rain" + 0.028*"acid" + 0.027*"water" + 0.017*"effect"')
(2, '0.249*"class" + 0.229*"scholarship" + 0.200*"students" + 0.063*"student"')
(3, '0.146*"essay" + 0.062*"favourite" + 0.036*"speech" + 0.028*"policy"')
(4, '0.047*"rainfall" + 0.037*"water" + 0.021*"control" + 0.021*"life"')
```



**Code Execution Flow:**

1. Web Scraping:

- The Selenium library is imported.
- An instance of the Chrome driver is created.
- The web page "https://www.studytoday.net/acid-rain/" is opened.
- All elements on the page are retrieved using the XPath expression.
- Text is extracted from the elements and stored in the all_text list.

2. Text Preprocessing:

- The spaCy library is imported, and the en_core_web_md model is loaded.
- Parts of speech (POS) tags to be removed from the text are defined.
- Tokenization, lemmatization, and removal of stop words and specified POS tags are performed using spaCy.
- The preprocessed tokens are stored in the tokens list.

3. Topic Modelling:

- The gensim library is imported.
- The tokens list is converted into a gensim dictionary and corpus.
- The dictionary and corpus are saved as files for future use.
- An LDA (Latent Dirichlet Allocation) model is trained using the corpus and dictionary.
- The trained model is saved as a file.
- The top words for each topic are printed.

4. Topic Visualization:

- The saved dictionary, corpus, and LDA model are loaded.
- The pyLDAvis library is imported and enabled for notebook visualization.
- The pyLDAvis visualization is generated using the loaded LDA model, corpus, and dictionary.
- The visualization is displayed.

**Summary and Analysis:**

The provided code demonstrates a complete workflow for topic modelling. It starts with web scraping to gather text data from a webpage. Then, the text is preprocessed by removing stop words, applying lemmatization, and filtering based on specific parts of speech. After preprocessing, the code uses the gensim library to train an LDA model on the preprocessed text.

The trained LDA model is then used to generate the top words for each topic. The number of topics is set to 5 in this example. The code prints the top words for each topic, allowing users to gain insights into the main themes present in the extracted text.

Additionally, the code employs the pyLDAvis library to create an interactive visualization of the topic model. The visualization provides an intuitive representation of the topics and their relationships, making it easier to interpret and explore the results.

The generated visualization can help users understand the underlying themes in the extracted text and explore the relationships between topics. It can be a useful tool for researchers, content analysts, and anyone dealing with large amounts of text data.

Overall, the provided code demonstrates an effective implementation of topic modelling using web scraping, text preprocessing, and the gensim library. The combination of these techniques enables the identification and exploration of latent topics in textual data.