


Topic Modeling using LDA and Text Summarizer web-app



Abstract

The project aims at building topic models using the LDA algorithm to simplify visualizing and analyzing a given long text data.

The dataset (sample) used contains reviews of a tyre company. The dataset has more than 10,000 lines. After feeding the dataset to the model it'll generate graphs and visuals depicting the topics discussed in the reviews.



Requirements

Functional requirements:

- Dataset for feeding the model
- Removing contractions
- Removing words with high frequency
- Generating graphs and visuals

System requirements:

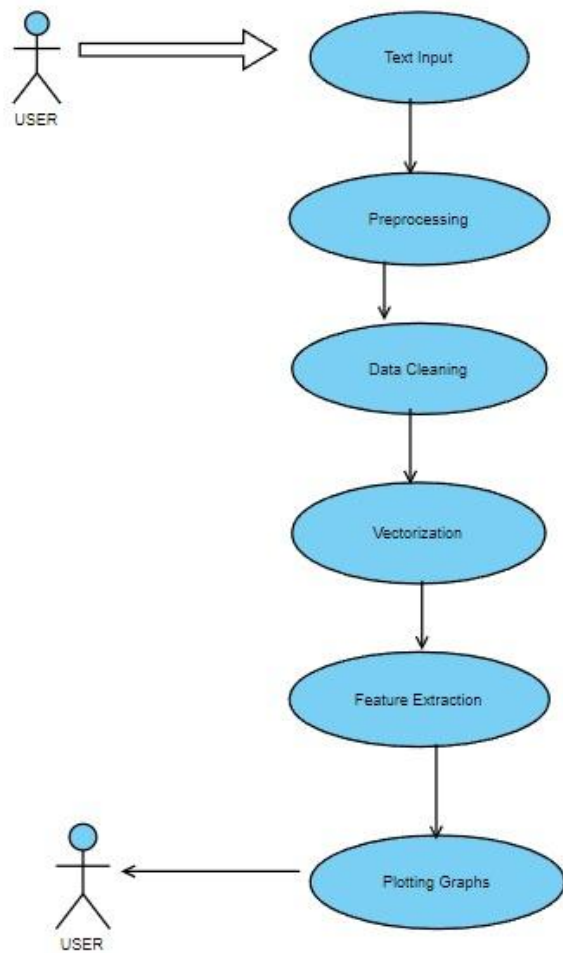
- Python 2.7 or newer
- Core i3 and 4GB RAM
- Web browser
- Jupyter Notebook



Approach

1. Set Up → get all the libraries and data needed in a workable format
2. Feature Extraction → Extract helpful information to understand the data
3. Data Cleaning → Clean the data to make it workable with models efficiently
4. Vectorizing → Make the dataset numeric
5. Exploration → Get insights from the data





Feature Extraction

The following features are extracted from the given text:

1. Number of words in a given review.
2. Number of characters in a given review
3. Average word length of a review
4. Number of numeric characters present
5. Number of words, which are in UPPER CASE



In reviews, it is seen that longer reviews are usually negative, and short reviews are positive. UPPER case words show some emotion and should be taken into account. Similarly, the numeric characters can explain whether the reviews are about a date or order ids.



Data Cleaning

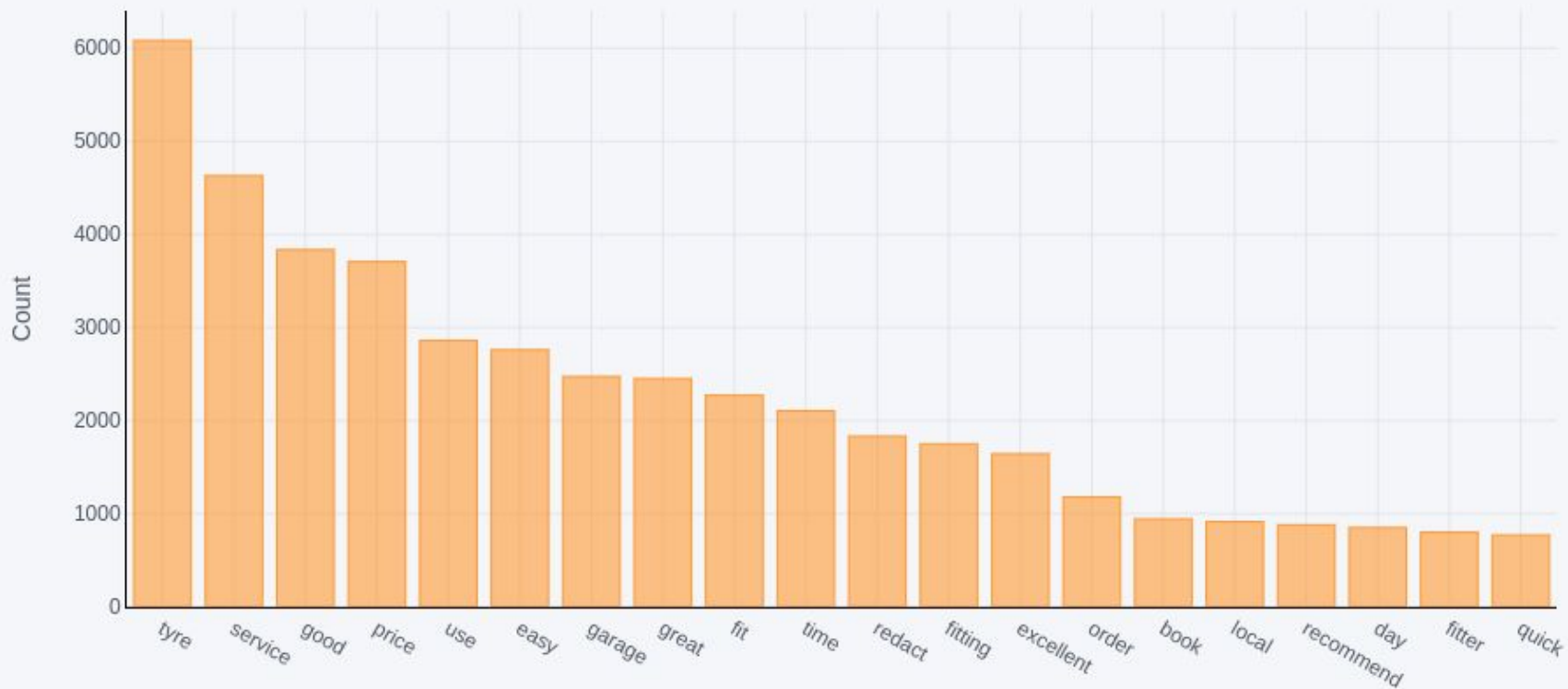
1. Change all words to lower case
2. Expand certain contractions
3. Remove punctuations and special characters
4. Remove extra spaces and trailing spaces
5. Remove accented characters
6. Remove stop words
7. Change the words to their base form



Exploration



Top 20 unigrams



Text Summarizer

Text summarization is the process of making synopsis from a given text document while keeping the important information and meaning of it.

Automatic Summarization has become an essential method for accurately locating significant information in vast amounts of text in a short amount of time with minimal effort.

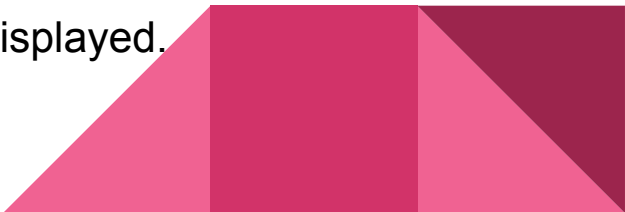
In this project, we propose to implement a web application that can summarize a text or a wikipedia link.

Text summarization is one of the natural language processing (NLP) applications that will undoubtedly have a significant influence on our lives.

With the rise of digital media and ever increasing publication, who has the time to read complete news items, documents, books to determine whether they are beneficial or not?



How does it work?

- Upon receiving a url, scraper.py scrapes the text present on the website. The text is then formatted and clean.
 - This formatted text is then passed to the summarizer.py which uses spacy tokenizes the text into sentences and words.
 - The frequency of each word is calculated and stored in a dictionary. The frequency of each word is then normalized by dividing by the maximum frequency (this is done in order to find the relative frequency of each word).
 - Next, the sentence scores are calculated by adding the word frequency of each word present in the sentence.
 - A heap queue is then used to get sentences with the highest sentence scores.
 - The sentences are then joined to get the summary.
 - Next, the estimated reading time is calculated.
 - Finally, the title, summary and estimated reading time are displayed.
- 

Libraries Used:

1.Spacy

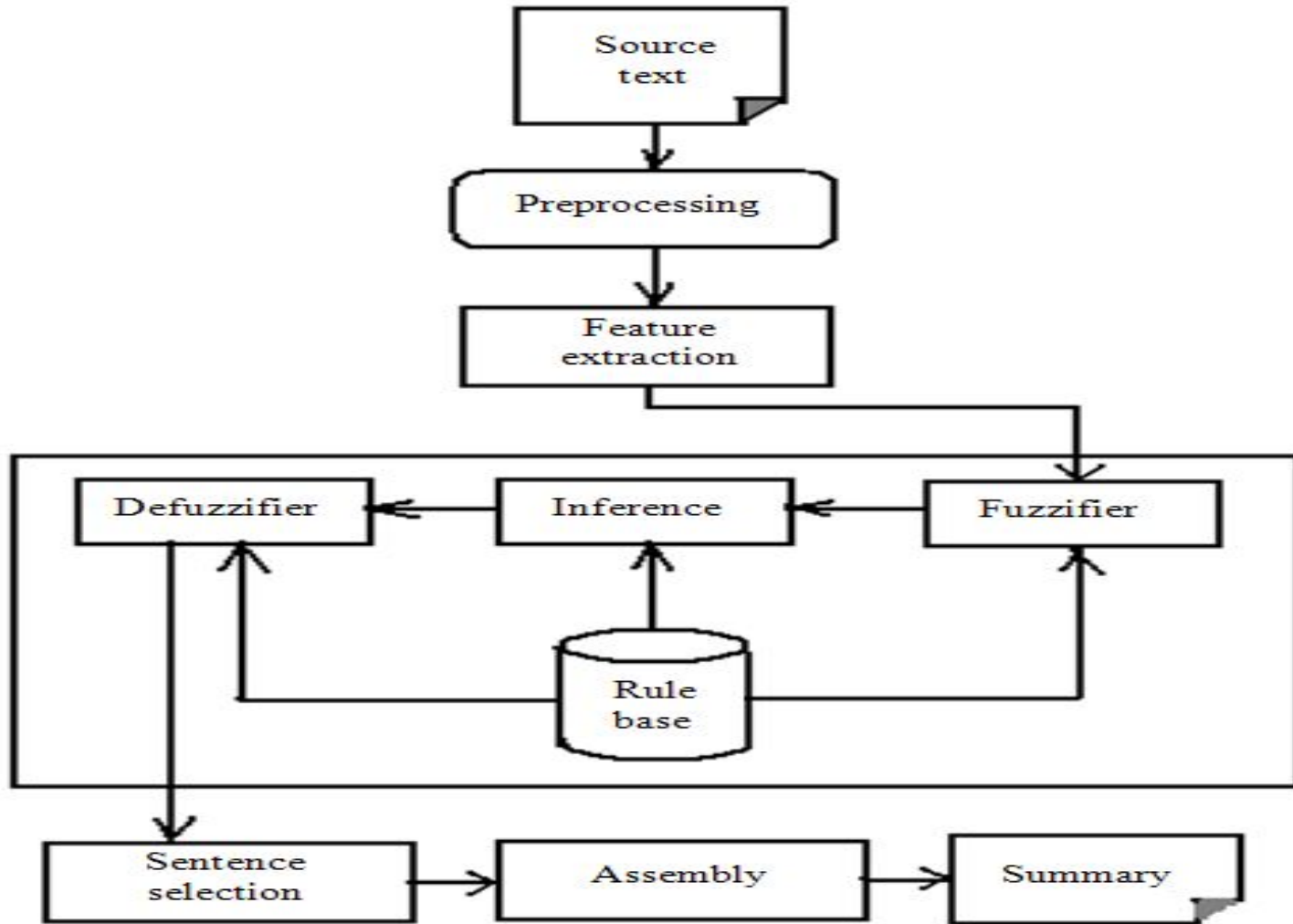
2.Flask

3.Regular Expression(re)

4.BeautifulSoup

5.Urllib





Industry Use-cases

- **Media Monitoring**

The issue of overload of information and content shock can be solved by automatic summarization as presents it can condense the continuous torrent of information into smaller pieces of information.

- **Search Marketing and SEO**

Multi-document summarization can be a powerful tool to quickly analyze dozens of search results, understand shared themes and skim the most important points.

- **Internal Document Workflow**

Summarization can enable analysts to quickly understand everything the company has already done in a given subject, and quickly assemble reports that incorporate different points of view.

- **Medical Cases**

Summarization can be a crucial component in the tele-health supply chain when it comes to analyzing medical cases and routing these to the appropriate health professional.

- **Books and Literature**

Summarization can help consumers quickly understand what a book is about as part of their buying



Text Summarizer

Paste url of article below

<https://en.wikipedia.org/wiki/Internet>

Summarize

Summary

Essay on Cow | Cow Essay for Students and Children in English - A Plus Topper (Estimated reading time: 1 mins, 26 seconds)

Cow: Cow is one of the most useful domestic animals and is of great use to humanity. In a rural area, cow dung is used to make dry cow dung cakes, which are used as fuel for burning and are used in the kitchen to providing a flame for cooking daily. Given below is an extended essay of approximately 400-500 words and is for the students of standards 7, 8, 9, and 10 and a short piece of nearly 100-150 words for the students of standard 1, 2, 3, 4, 5, and 6. A cow is one of the most innocent and loving domestic animals who are harmless. If the general physical description of a cow is to be given then, a cow is a four-legged animal with a large body and two horns, a mouth, two eyes, and two ears. The flesh of the cow is tanned to make cow leather, and it is the most widely used form of leather all over the world. A cow's milk can be used to make various dairy products, for example, butter, clarified butter, curd, cottage cheese,