Auto Tagging / Annotation of Documents

Major Project
In
Information Retrieval and Extraction

By-
Aditya Gupta (2019201067)
Aditya Mohan Gupta(2019201047)
Upinder Singh(2019201083)
Sai Raju Ramchander(2020900010)

18 November 2020

Abstract

The idea of this project is to generate reviews for the given document based on templates and important keywords.

We initially take a review dataset and mask a few words from each of the reviews from the given dataset and generate our templates. With the use of pre pre-trained model we try to predict the masked positions of the template using HRW(high ranking words) from the document for which we want to generate a review.our model tries to fit those keywords and generate the review.

TABLE OF CONTENT

Datasets:

The following are the datasets that are used.

- ❖ NY Datasets
- ❖ Amazon review Datasets

From the **NY Dataset** we have collected both the articles and reviews.
From the **Amazon Review Dataset** we have collected only reviews.

Data Preprocessing:

Usual preprocessing steps are used to take some text data in its raw form and transform it into text data that will be more useful for neural network processing.
The following are the steps used in the data preprocessing

- ➢ Removal of stopwords from the data.
- ➢ Removal of punctuations.
- ➢ Stemming the words.
- ➢ Converting the text to lower case.

❏ Removal of stopwords:

stopwords are **removed** or excluded from the given text so that more focus can be given to those **words** which define the meaning of the text.

❏ Removal of Punctuations :

All the **punctuations** and the special characters are being **removed** by using the **regular expression** as they don't make a meaningful sentence.

❏ Stemming of words :

Stemming is the process of getting the word into a root form. In this phase the given word is brought into the root word. This helps the model to identify the context of the word.

❏ Converting the text to lower case :

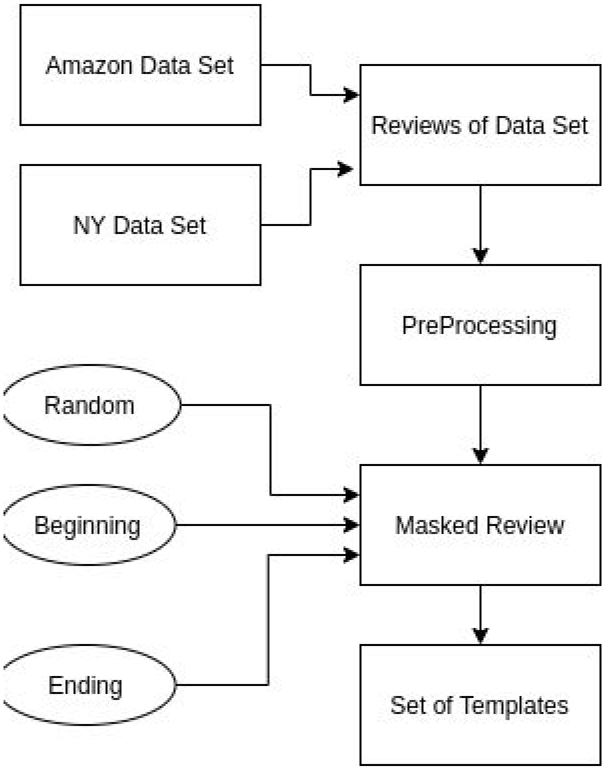
This is the final phase in the data pre-processing. In this phase all the data is being converted to lowercase. This is done to avoid the conflict with the words starting with upper case and to maintain the uniform between all the words.

Masking Reviews:

Once the Review data is completely cleaned. The Reviews are masked to generate templates. The data with masked tags are called the templates.

- Masking is carried in three different ways.
  - ❖ From beginning : few words are taken from beginning, and is replaced by ***mask***.
  - ❖ From End : few words are taken from End, and is replaced by ***mask***.
  - ❖ Randomly : few words are taken from in between and is replaced by ***mask***.

Generating templates from Reviews



Web Scraping:

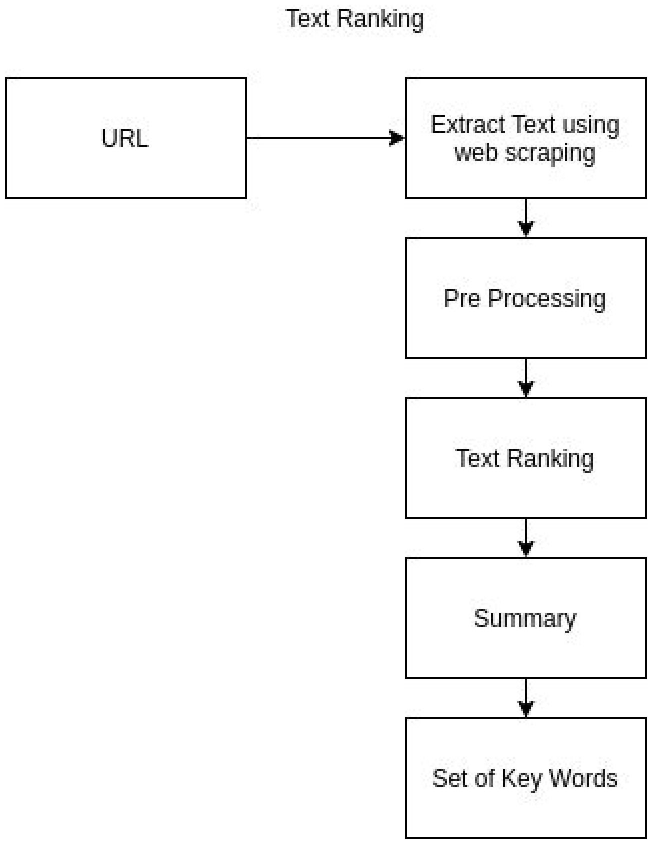
Web Data Extraction, Web Harvesting etc.) is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table.

Once we get the **URL** of articles from **NY dataset**, our work is reduced to fetch content from URL for respective article.

Web Scraping method is used to perform this task.

Text Ranking:

Text ranking graph based ranking algorithm for NLP. TextRank is an automatic summarisation technique. Graph- based ranking algorithms are a way for deciding the importance of a vertex within a graph, based on global information recursively drawn from the entire graph.



- From the given Url’s the text is extracted using web scraping.
- The preprocessing is done on the text.
- The actual text ranking is being done here.
- All the summary of text rankling is done here.
- Finally the set of key words are given as output.

Summary of Text:

Summary of Text is shortened version of a text. It contains the main points in the text and is written in your own words. It is a mixture of reducing a long text to a short text and selecting relevant information. We are using the **text ranking algorithm** to extract the summary of the given document.

Tokenization:

Tokenization is the process of turning a meaningful piece of data, such as an account number, into a random string of characters called a token.

Options:

To get the **list of words** that needs to be **filled** in the masked value. Words are chosen on the basis of their high scores that are fit in the suitable mask in a template.

Model Used : FitBert

FitBert model is used for **text-infilling** purpose.It takes two inputs, one is a list of keywords that need to be filled in a masked position and the other is a masked template and gives one output after filling masked positions.

FitBERT stands for “**F**ill **i**n **t**he Blanks, **BERT**”. Because of the training objective for BERT (masked language modeling), it is very good at filling in the blanks. In fact, that is one of the two tasks that BERT is trained on (the other is next sentence prediction, and the RoBERTa paper showed that performance on downstream tasks actually improves if you only train on the fill in the blank task).

Semantic Similarity:

- ❑ Pretrained model GloVe:
  Global Vectors, is a model for distributed word representation. The model is an unsupervised learning algorithm for obtaining vector representations for words. This is achieved by mapping words into a meaningful space where the distance between words is related to semantic similarity. As a log-bilinear regression model for unsupervised learning of word representations, it combines the features of two model families, namely the global matrix factorization and local context window methods.

- ❑ Cosine similarity:
  Cosine similarity is a metric used to determine how similar the documents are irrespective of their size. Mathematically, it measures the **cosine** of the angle between two vectors projected in a multi-dimensional space.

- ❑ **Semantic Similarity score** is a parameter used to evaluate two sentences based on their semantics.To calculate this score we are using pretrained GloVe Model and cosine similarity combinely.

- ❑ Both of the above methods are used combinedly to find **semantic similarity score** between a given article and each predicted review.The review which gives maximum similarity score that is considered as a final review for the corresponding article.

Technologies used:

- Python3

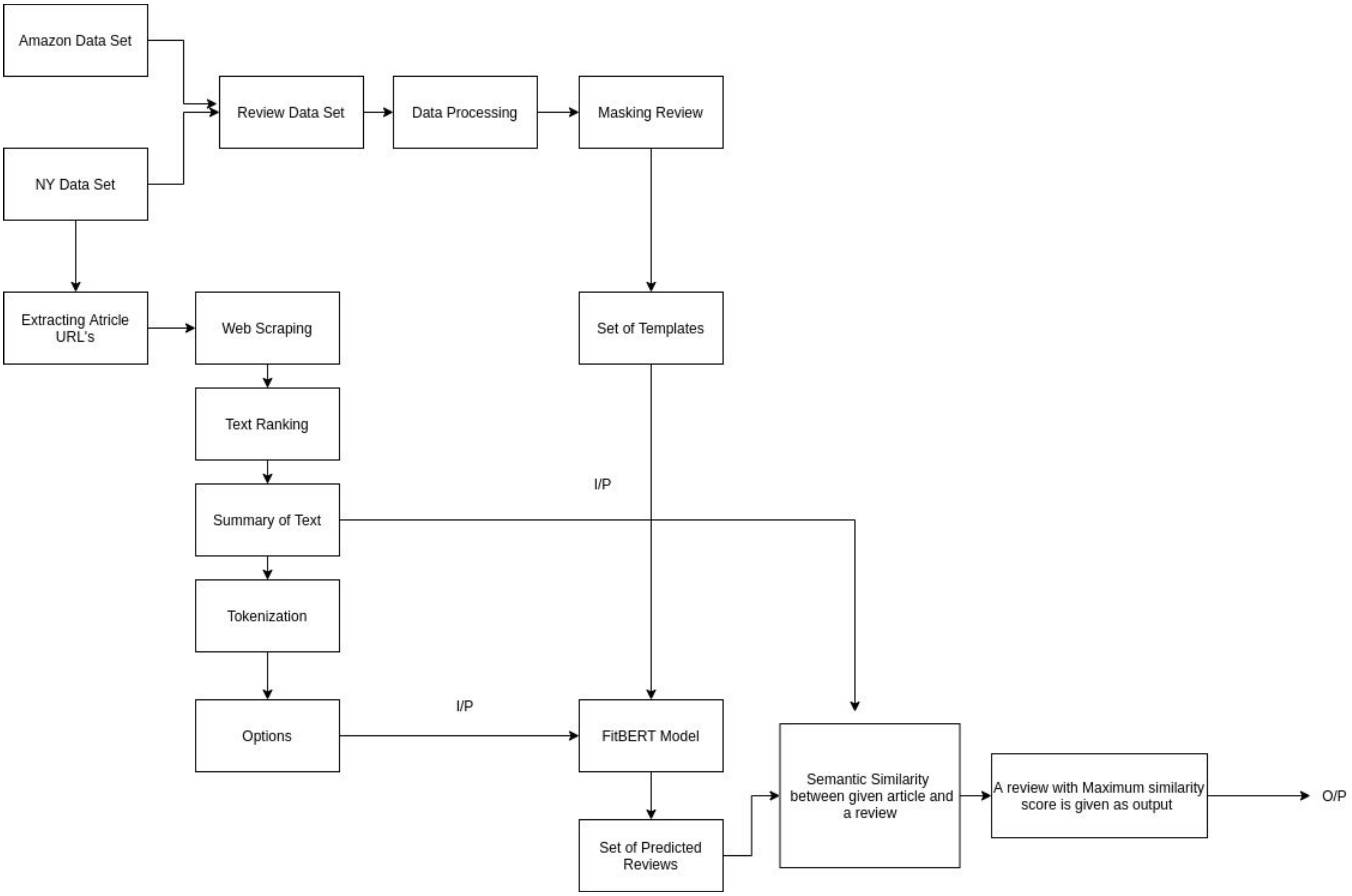
Libraries used:

- The following are the libraries that are being imported.
    - Libraries that are used for data cleaning and manipulation
        - Pandas
        - Numpy
        - Random

- Libraries that are used for tokenizing the given sentences
    - SKlearn
    - Nltk
    - Gensim
    - Sklearn

- Library used for pre trained model
    - Fitbert
    - GloVe

FlowChart


Amazon Data Set
Review Data Set
Data Processing
Masking Review
NY Data Set
Extracting Atricle URL's
Web Scraping
Set of Templates
Text Ranking
I/P
Summary of Text
Tokenization
I/P
Options
FitBERT Model
Semantic Similarity between given article and a review
A review with Maximum similarity score is given as output
O/P
Set of Predicted Reviews

References

- https://towardsdatascience.com/textrank-for-keyword-extraction-by-python-c0bae21bcec0
- https://pypi.org/project/pytextrank/
- https://stats.stackexchange.com/questions/312206/can-i-apply-word2vec-to-find-document-similarity
- https://github.com/AndriyMulyar/semantic-text-similarity
- https://medium.com/@adriensieg/text-similarities-da019229c894
- https://blog.ekbana.com/loading-glove-pre-trained-word-embedding-model-from-text-file-faster-5d3e8f2b8455
- https://medium.com/@samhavens/introducing-fitbert-4b047af860fd