Automatic Document Summarization using a Machine Learning approach

By

TEAM - JUSTICE LEAGUE

Department of Computer Science & Engineering
INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR

NOVEMBER 2016

JUSTICE LEAGUE

Sai Sriharsha Annepu - 13CS10012
Ananth Pranihith. P - 13CS10011
Thejesh Venkata - 13CS10013
Pavan Reddy. B - 13CS10015
Jyothi Swaroop. B - 13CS10016
Prithvi Raj Reddy - 13CS10029
Konda Akhil - 13CS10030
Sai Sambasiva. P - 13CS10034
Supradeep Allu - 13CS10050
Aswanth Kumar - 13CS30019
Suryateja Chunduru - 13EE10017
Prasanth Balaga - 09CS3031

OBJECTIVE

DEFINITION:

Automated Text Document Summarization is the process of reducing the content of a text document with a computer program in order to create a summary that retains the most important points of the original document.

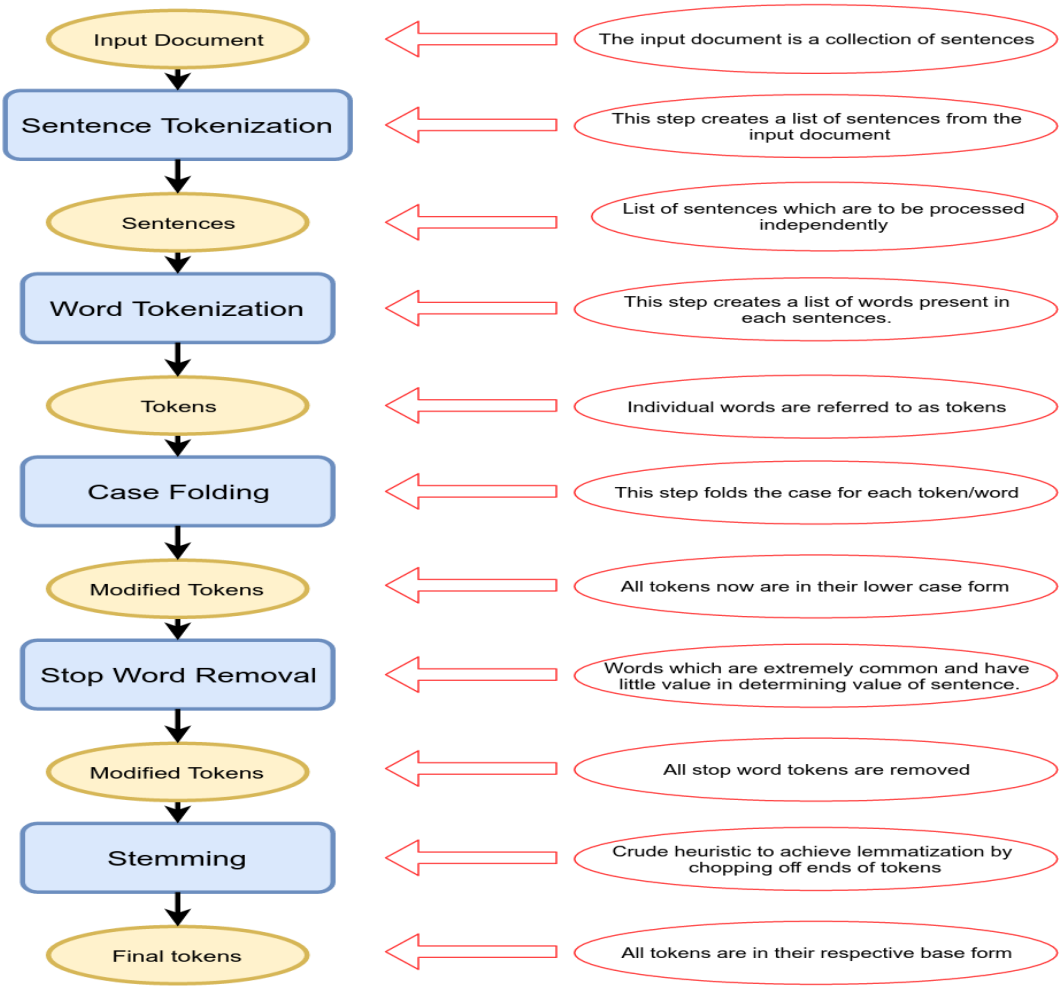
OBJECTIVE:

The objective of the project is to build an Automatic Text Document Summarization System to help users obtain essence of large documents with least effort using a Machine Learning Approach.

METHODOLOGY

The generation of a summary depends on the computed feature vectors of all the sentences. The pre-processing unit splits the text into word groups while attaining the text sentence structure.

Pre-Processing steps



Features Identified:

When a new document is given to the system, the "learned" patterns are used to classify each sentence of that document into either a "present-in-summary" or "not-present-in-summary" sentence, producing an extractive summary. A crucial issue in this framework is to obtain the relevant set of features. The summarizer produces summaries using the features extracted. The following features have been identified as important features in the developement of the system:

1) TF-ISF
2) TEXT RANK
3) WORDNET BASED RANKING
4) SENTENCE LENGTH
5) PROPER NOUNS
6) NUMERICAL DATA
7) SENTENCE POSITIONING

Feature Extraction:

TF-ISF

This feature is analogous to TF-IDF in the context of Information Retrieval. In IR, we have to select few documents which are most relevant from a given set of documents. Here, we have to select few sentences from the given document to be included in the extractive summary of the document. The used feature was calculated as the mean value of the TF-ISF measure for all the words of each sentence. The feature was calculated in the following way:

$$\frac{\sum_{t \in s} TF - ISF(t)}{\sum_{t \in s} 1}$$

Sentence Length

This feature is employed to penalize sentences that are too short, since these sentences are not expected to belong to the summary. sentences containing less than a pre-specified number of words are not included in the abstract. We use the normalized length of the sentence, which is the ratio of the number of words occurring in the sentence over the number of words occurring in the longest sentence of the document.

Sentence Positioning

Usually first and last sentence of first and last paragraph of a text document are more important and are having greater chances to be included in summary. This feature can involve several items, such as the position of a sentence in the document as a whole, its the position in a section, in a paragraph, etc., and has presented good results in several research projects.

Text Rank

The TextRank algorithm exploits the structure of the text itself to determine keyphrases that appear "central" to the text in the same way that PageRank selects important Web pages. TextRank is a general purpose graph-based ranking algorithm for NLP. For keyphrase extraction, we built a graph using some set of text units as vertices. Edges are based on some measure of semantic or lexical similarity between the text unit vertices. Unlike PageRank, the edges are typically undirected and can be weighted to reflect a degree of similarity. The proposed graph based text ranking algorithm consists of three steps:

1. Word frequency analysis
2. Word positional, string pattern based weight calculation algorithm
3. Ranking the sentences by normalizing the results above

Proper Nouns

Sentences containing proper nouns are having greater chances for including in summary. The motivation for this feature is that the occurrence of proper names, referring to people and places, are clues that a sentence is relevant for the summary. This is considered here as a binary feature, indicating whether a sentence s contains (value "true") at least one proper name or not (value "false"). Proper names were detected by a part-of-speech tagger.

Numerical Data

Sentences containing numerical data have higher chances for including in the extractive summary of the document. This feature emphasis on the numerical data present in the sentences. The numerical data generally correspond to some computed values or results which are important to be included in the summary. So, the sentences having numerical data, we add extra score to them.

Wordnet Ranking

We created list of Keywords by taking Adjectives, nouns and Proper nouns. We have calculated number of times a particular word or its synonyms occurred in document by using synset. Now if Keyword is a proper Noun a score of 1 is assigned or else a highest similarity score is assigned which is calculated with respect to synonyms of other words present in document. Now for each sentence a score is assigned which is (sum of scores of all keywords present in the sentence)/(Number of keywords present in sentence).

Training the summarizer

The summarizer was trained with a set of documents and their respective gold standard summaries using supervised learning algorithms. To accomplish this, a classification function that estimates the probability of a sentence being selected in the summary was developed. We were given with a set of training documents, their respective summaries and a fea-

ture vector file for each of those training documents. A feature vector file was containing feature vectors of all the sentences of the respective training document. A feature vector is a vector of feature values of a sentence. We combined the feature vectors from all those files and divided them into two classes. First one, containing the feature vectors of the sentences that are selected into the summary and the other containing the feature vectors of the sentences which are not selected into the summary. Using the feature vectors in a class, the variance and standard deviation of the values for each feature have been computed for that class.

Summary generation

A feature vector file is prepared for the given input document with a feature vector for each of its sentence. Now the feature vectors of these are classified into one of the two classes mentioned above, using naive bayes classifier. The variance and standard deviation values that were learned in the training phase are used for calculating the probabilities that involved in the naive bayes formula (probability formula of normal distribution involving variance and standard distribution is used). The sentences whose feature vectors fall into the first class make up the output summary.

RESULTS

The ROUGE(Recall-Oriented Understudy for Gisting Evaluation) toolkit is employed to evaluate the proposed algorithm. ROUGE, an automated summarization evaluation package based on Ngram statistics, is found to be highly correlated with human evaluations. The evaluations are reported in ROUGE-1 metrics, which seeks unigram matches between the generated and the reference summaries. The ROUGE-1 metric is found to have high correlation with human judgments at a 95% confidence level and hence used for evaluation.

The results obtained for the summarizer developed are shown below.

LARGE SUMMARIES

Rouge N-Gram	Avg_Precision	Avg_Recall	Avg_fScore
ROUGE - 1	0.50600	0.76950	0.59588
ROUGE - 2	0.46526	0.71521	0.55064
ROUGE - 3	0.46037	0.71552	0.54740
ROUGE - 4	0.46004	0.72440	0.54993

SMALL SUMMARIES

Rouge N-Gram	Avg_Precision	Avg_Recall	Avg_fScore
ROUGE - 1	0.53127	0.40905	0.41225
ROUGE - 2	0.42793	0.33680	0.33869
ROUGE - 3	0.41708	0.32892	0.33120
ROUGE - 4	0.41445	0.32786	0.33017

The training of the system was done on both large and small summaries using 120 documents having the gold standard summaries. The means and the standard deviations are calculated and stored. The tesing of the system was done on 90 documents and the obtained summaries are compared with the gold standard summaries. It is observed that this approach gives better results than systems using only a single feature to extract the summary. Essentially, we are generating the summary using the information obtained by all the features and hence this is a more novel approach to solve the problem.