Text Summarization Techniques

Text summarization is getting an idea of text in fewer words. In this we apply different techniques for text summarization:

1. Term-Frequency: This is the method of summarization based on the frequency of the words that are used in the text.

Idea:

- First method of constructing by taking each word and then assigning the frequency for that word as score of the word.
- Second method is similar but in this we remove the stop words (words that does not contribute in the meaning of sentence e.g. is, am, are) and special characters (e.g. '(', ']'). This method performs better than First One.

Steps:

- Load data.
- Perform sentence segmentation.
- Collect all words with their frequencies based on the method first or second.
- Score each sentence based on the frequency of each word in the sentences.
- Longer sentences are more likely to have greater score even they may not carry significant meaning. Thus, we normalize the score by dividing it with sentence length.
- Set the threshold of the sentences by taking average score of all the sentences.
- Generate summary by considering only the sentences whose score is more than threshold.

2. Term-Frequency Inverse-Document-Frequency (TF-IDF): It is extension of above method. In this method with the frequency of word, frequency of sentences relevant to words is also considered.

Idea:

- Consider two words having same frequency (e.g. 4) and not the part of stop words with the difference that first word is used in only two sentences twice and second word is used four times in four sentences. Now the word with four sentences has lower significance in the summarization than word with two sentences. So, we reduce the word significance by multiplying it with inverse frequency of sentences in which word lies.

Advantage:

- This method works well even without removal of stop words form the dataset, as the stop words are used throughout the text and directly get penalized by **IDF** factor.

Steps:

- Load data.
- Perform sentence segmentation.
- Collect all words with their frequencies. (Removing stop words give better performance).
- Collect log (IDF) for each word.
- Prepare the word-set with all words having score as product of word frequency and IDF.

- Score sentence based on the score of all the words.
- Longer sentence tend to have advantage over smaller sentences. So, we normalize by dividing by the number of words used in sentence.
- Threshold is found. Simplest way to do this is by calculating the average.
- Generate summary by checking whether score of sentence is greater than threshold or not.

3. Topic Signature: In Topic Signature, mid-range frequency words are collected than calculate the score based on length span of topic words in sentence.

Idea:

- Mid frequency words are supposed to carry more information i.e. stop words and very low frequency word do not influence summarization much. Thus, Frequency is used just to sort words into relevant or non-relevant categories but the frequency is not used in calculating sort. It is used just to filter the non-relevant words. For scoring each sentence we calculate the span of topic signatures (topic words) instead of IDF.
- Score can also be calculated just by taking the number of topic words each sentences has although the method above use is more efficient, this method is easy and provide good results.

Steps:

- Load data.
- Perform sentence segmentation.
- Collect all words. (Removing stop words and low frequency words).
- Find the proportion of sentence of each sentence (i.e. length from first to last topic word).
- In Luhn's idea, score is calculated by dividing sentence span by square of no. of topic words.
- Generate the summary by checking score with threshold value.

4. Centroid Based Clustering: This method uses similarity score to find the sentences having meaning similar to title of text, this work much better for the text with title. For text with no title, sentences with most relevant words tend to dominate.

Idea:

- Sentence with similar words tend to have similar meaning. Thus, cosine matrix is used where it compare each word of one sentence to each word of other and provide similarity for complete sentence.

Steps:

- Load data.
- Perform sentence segmentation.
- Collect all words. (Removing stop word).
- Tokenize all sentences with the relevant words.
- Convert all tokenized sentences into vectors.
- Find the cosine similarity matrix for each sentence.
- Compare all sentences with title or most relevant sentence of the text and generate summary.

5. Indicator Representation: This method is the transformation of sentence into features. It is not directly used in text summarization but still can provide good summarization result.

Idea:

- Here we focus more on the representation of sentences instead of the words. Since, the sentences carry information but only assigning weight to words and sentences based on frequency may lead to loss of information. E.g. Loss by 2 million is not frequency significant but data with numbers hold information or Company did well and Company made profit both carry positive message. So, instead we can use only one statement. Thus, considering only frequency does not deal with such cases. Thus, we develop each such idea into an individual feature.

  Features:

  - Title words: The words in the title of the article carry information about the article. So, sentence with the title words are more likely to contain information regarding article.
  - Sentence length: The longer sentences are more likely to carry more information about the text. Hence we prefer longer sentences for summary consideration
  - Sentence position: Every beginning sentence of the article or paragraph either contain new idea or justification or explanation of the previous sentences mentioned. This implies the position of sentence held is important.
  - Sentence similarity: Sentences with similar meaning are more important to represent a single idea. Thus, articles with four clusters having multiple sentences can be represented in only four sentences. Or in depth knowledge of particular type is required it can be explored by looking only all the sentences of cluster.
  - Term weight: This is the representation of words used in the sentences into numbers. Generally, a method from any of the above is used to represent weight of words. (E.g. TF-IDF).
  - Numerical data: Sentences with numbers are likely to have importance as quantifying a data is much better representation than words (E.g. Company made Hugh profit this year vs. Company had a 50 million turnover more in 2017 than in year 2016).

Advantage:

- This method has proven to be much better at performing summarization.
- The number of features does not with the very with the length of article thus low memory is used.
- Model can be easily trained and modified.

Steps:

- Load data.
- Perform sentence segmentation.
- Collect all words. (Removing stop word).
- Tokenize all sentences with the relevant words.
- Find features of all the sentences in the article. Now to remove any bias we divide the each feature by maximum of the feature values.
- Normalize all the features. (E.g. divide the each sentence length by max sentence length).
- Calculate the score based on threshold.

6. Scoring Indicator Representation:

- Fuzzy Logic Set: This is just the scoring method for Indicator Representation of all the features. Dividing the complete range into 5 sets and then based on the score obtained, the score of each feature can be labeled as L1-L5 (5 levels). Also, the range can be customized based on the need of article to be inclined (E.g. based on numerical data, exploring one area (i.e. single cluster) or number of sentences to be included in the summary).
- Normalizing data: Once the data is normalized (i.e. we set the range between 0 and 1) score can be calculated. If for all features score is between fixed ranges then the summary can be created directly adding the scores and comparing it with the threshold.

7. Machine Learning:

Idea:

- In Machine Learning the summarization concept is modified to classification. There is a need for summary to perform classification over text. This summary can be obtained both by humans or any of the above method. Now, By training the model over existing summary, the testing is done and then for next articles with each sentences we get whether the sentence is part of the summary or not.

Steps:

- Convert all the sentences into vectors either by Indicator representation or TF-IDF or TF.
- A dependent variable is created whether a sentence is a part of summary or not.
- Create the classification model with Naïve Bayes or Logistic Regression or SVM.
- For any target convert the article in the vectors and run model.
- For each sentences the prediction equal 1 sentences are collected and then combined to form a summary.