PUNE INSTITUTE OF COMPUTER TECHNOLOGY,
DHANKAWADI PUNE-43.

A Project Report On

Automatic Keyword Extraction using TextRank Algorithm

BY

Prajakta Mane -4151
Krishna Patel -4146
Anmol Khanapure -4142

CLASS: BE-1

ISO 9001:2015 Certified

COMPUTER ENGINEERING DEPARTMENT

Academic Year: 2019-20

PROBLEM STATEMENT : Automatic Keyword extraction using TextRank Algorithm (Graph Algorithm)

INTRODUCTION:

Keyword extraction is a text analysis technique that consists of automatically extracting the most important words and expressions in a text. It helps summarize the content of a text and recognize the main topics which are being discussed.

In keyword extraction with TextRank, the goal is to first rank each word in a passage based on its importance, and then return a certain number of words with the highest rankings. To do this, the first step is to create a graph. Let nouns and adjectives be vertices in the graph, noting intuitively that verbs, prepositions, and other parts of speech are not usually important when considering keywords. The nouns and adjectives in the graph were then connected by weighted edges based on co-occurrence, or closeness, scores between words. If two words appear within a certain number of words from each other in the passage, they will be connected in the graph with a higher edge weight the closer they are. Once the graph is constructed, the same process of recursively assigning scores to vertices used in PageRank is applied to the TextRank graph until each word's score converges. After convergence, each word in the graph will have a score and the top scores will become keywords.

THEORY:

TextRank is an algorithm based on PageRank, which often used in keyword extraction and text summarization.PageRank (PR) is an algorithm used to calculate the weight for web pages. We can take all web pages as a big directed graph. In this graph, a node is a web page. If webpage A has the link to web page B, it can be represented as a directed edge from A to B.

After we construct the whole graph, we can assign weights for web pages by the following formula:

$$S(V_i) = (1-d) + d * \sum_{j \in In(v_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

- S(Vi) - the weight of webpage i
- d - damping factor, in case of no outgoing links
- In(Vi) - inbound links of i, which is a set
- Out(Vj) - outgoing links of j, which is a set
- |Out(Vj)| - the number of outbound links

PageRank is for webpage ranking, and TextRank is for text ranking. The webpage in PageRank is the text in TextRank, so the basic idea is the same.

We split a document into several sentences, and we only store those words with specific POS tags. We use spaCy for POS tagging.

Because most of the words in a sentence are not useful to determine the importance, we only consider the words with NOUN, PROPN, VERB POS tags.

Each word is a *node* in PageRank. We set the window size as k.

$$w_1, w_2, w_3, w_4, w_5, \ldots, w_n$$

[w1, w2, …, w_k], [w2, w3, …, w_{k+1}], [w3, w4, …, w_{k+2}]are windows. Any two-word pairs in a window are considered have an undirected edge.

We take [time, Wandering, Earth, feels, throwback, eras, filmmaking] as the example, and set the window size k=4, so we get 4 windows, [time, Wandering, Earth, feels], [Wandering, Earth, feels, throwback], [Earth, feels, throwback, eras], [feels, throwback, eras, filmmaking].

For window [time, Wandering, Earth, feels], any two words pair has an undirected edge. So we get (time, Wandering), (time, Earth), (time, feels), (Wandering, Earth), (Wandering, feels), (Earth, feels).

Based on this graph, we can calculate the weight for each node(word). The most important words can be used as keywords.

RESULT:

Thus, the Algorithm returns the most important keywords in a text file along with their scores .

Sample Input:

"'The Wandering Earth, described as China's first big-budget science fiction thriller, quietly made it onto screens at AMC theaters in North America this weekend, and it shows a new side of Chinese filmmaking — one focused toward futuristic spectacles rather than China's traditionally grand, massive historical epics. At the same time, The Wandering Earth feels like a throwback to a few familiar eras of American filmmaking. While the film's cast, setting, and tone are all Chinese, longtime science fiction fans are going to see a lot on the screen that reminds them of other movies, for better or worse.'"

Sample Output:

science - 1.7276378287292111

fiction - 1.7123791481736559

filmmaking - 1.4388798751402918

China - 1.4172218029410266

Earth - 1.3088154732297723

tone - 1.0971675275482093

fans - 1.0971675275482093

Wandering - 1.0071059904601571

weekend - 1.002449354657688

America - 0.9976329264870932

budget - 0.9769693829073562

North - 0.9711240881032547

CONCLUSION:

Thus,we have studied the PageRank and TextRank algorithm & implemented the TextRank algorithm

REFERENCES:

[1] http://blogs.cornell.edu/info2040/2018/10/22/40068/

[2] https://towardsdatascience.com/textrank-for-keyword-extraction-by-python-c0bae21bcec0