International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 5 ǁ 2022 ǁ PP. 93-96

An Outline Study of Keyword Extraction

Maahi Singh, Muskan Mathur
Department of Electronics & Telecommunications, PICT, Pune (MH), India

Abstract
In this era of technology and the increasing impact of social media, the majority of unstructured data is generated in a variety of businesses today. Text analysis becomes a challenging task with unstructured data. After a lot of research, it was concluded that keyword extraction can be an incredible tool to summarize the data easily and facilitates deriving vital insights. It can be a beneficial tool for the researchers, students and academicians for extracting relevant information from a large amount of unstructured data available and thereby helps to automate indexing, summarizing and extracting vital keywords. Keyword extraction is scalable, reliable and based on real-time analysis. Various methodologies like Statistical, Linguistic, Graphical, ML and Hybrid can be used depending upon the requirement of the user. It deals with wide area of services from customer support to the monitoring of social media and also SEO. The present paper endeavours to present a brief outline of Keyword Extraction technique including approaches, steps and applications that caters to the need of present-day users.
Keywords*: frequencies, graphical, linguistic, n-gram, NLP, statistical, term score*

Date of Submission: 15-05-2022 Date of acceptance: 29-05-2022

I. Introduction

Natural Language Processing (NLP) is an artificial intelligence approach of communicating with an intelligent system using a natural language such as English. Machines use this technology to comprehend, analyse, manipulate, and interpret human speech. It helps developers organise activities such as translation, automatic summarization, Named Entity Recognition (NER), audio recognition, relationship extraction, and topic segmentation.

Speech and Text can be two forms of input and output of NLP system.

NLP is also used for social media data processing. It presents a challenge in developing powerful methods and algorithms that extract relevant information from a large volume of data gathered from multiple sources and languages.

1.1.1 Keyword Extraction

The technique that automatically extracts the most used and most important words and expressions from a text is Keyword Extraction. It facilitates summarizing the text content and topic recognition. It uses AI with NLP to break down human language that can be understood and analyzed by machines. Keyword Extraction technique is used to find keywords from regular documents and business reports, social media comments, online forums and reviews, news reports etc.

Keyword extraction from grammatically ambiguous text is not easy compared to structured text since it is hard to rely on the linguistic features in unstructured texts as in social media platforms like Twitter etc. The objective of extracting keywords is to find related news in a more effective manner. For this approach, a corpus that contains tweet texts from different domains is built to make this approach more generic instead of making it a domain-specific approach.

All keyword extraction algorithms include the following steps:

- *Candidate generation*. Candidate keywords are detected from the text.
- *Property calculation.* Computation of properties and statistics required for ranking.
- *Ranking.* The score of each candidate keyword is computed and sorted in decreasing order of their scores. The final n keywords representing the text are top n candidate keywords.

1.1.2 Keyword Extraction techniques:

Different techniques are used for keyword extraction. They are mainly categorized in two categories supervised and unsupervised, we will be focusing on the unsupervised approaches in this paper. From simple statistical approaches that detect keywords by counting word frequency, to more advanced machine learning approaches that create even more complex models by learning from previous examples.

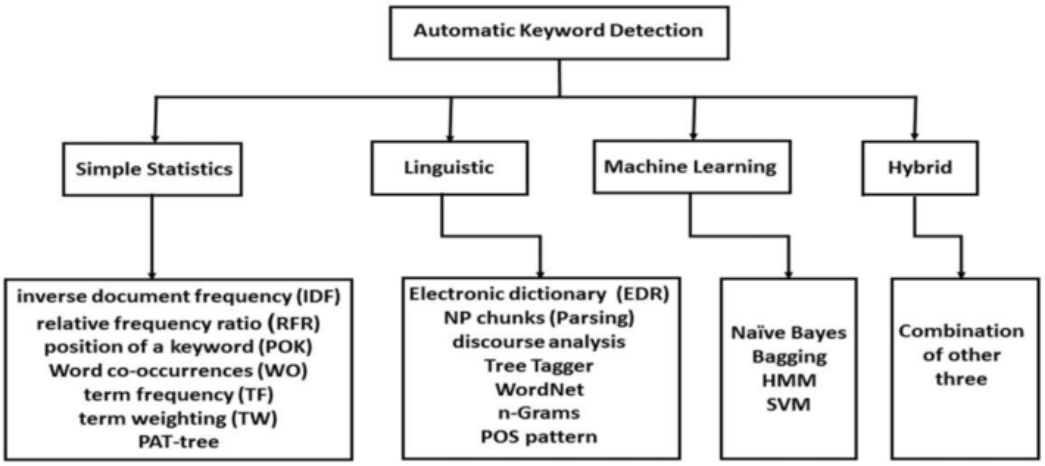

Fig.1 Different approaches to Keyword Extraction

Simple Statistical Approach:

The aim of statistical techniques is to use multiple statistics collected over a single text or across several documents to establish the individual term score in the document. The terms are ordered depending on their scores, and the top n terms are displayed as important keywords after the scores are calculated. There are several procedures to calculate n-gram scores, each with its own algorithm. Word frequency, word collocation, and co-occurrence are some of the most basic statistical procedures. Word frequency, word collocation, and co-occurrence are some of the statistical methodologies. TF-IDF and YAKE are the two most prevalent techniques.

1. TF-IDF:

It is the measure of the originality of a word by comparing the number of times a word appears in the document with the number of documents the word appears in. The equation of TF-IDF is given by:

TF-IDF = TF (t, d) * IDF (t)

$$\mathbf{tf}(t,d) = \frac{f_d(t)}{\max\limits_{w \in d} f_d(w)}$$

$$\mathbf{idf}(t,D) = \ln\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$$

$$\mathbf{tfidf}(t,d,D) = \mathbf{tf}(t,d) \cdot \mathbf{idf}(t,D)$$

$$\mathbf{tfidf}'(t,d,D) = \frac{\mathbf{idf}(t,D)}{|D|} + \mathbf{tfidf}(t,d,D)$$

$f_d(t) :=$ frequency of term t in document d

$D :=$ corpus of documents

In simple terms,

TF = number of times the term appears in a document/total number of words in the document
IDF = log(number of documents/number of documents the term appears)

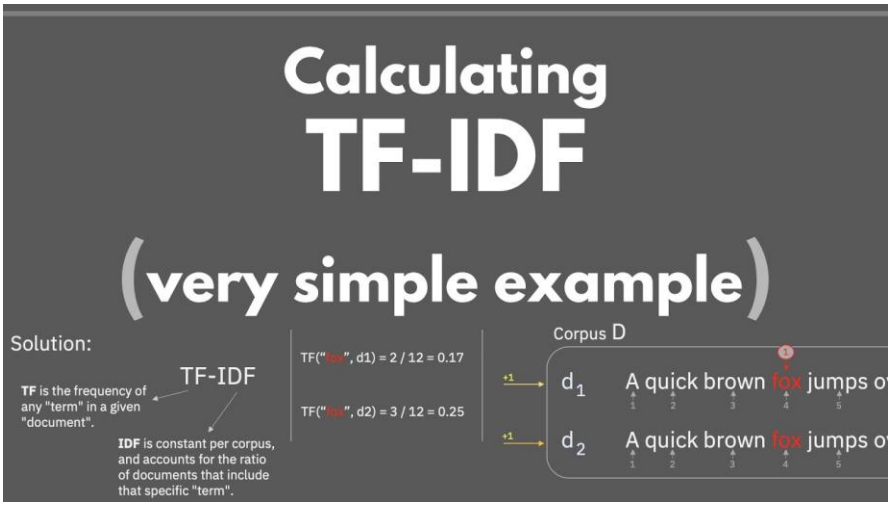

Fig.2 Calculation of TF-IDF with the help of an example

2. YAKE:

It's a simple unsupervised keyword extraction technique. It relies significantly on statistical text features that are chosen and computed from a single document. It distinguishes itself from the TF-IDF technique in that it does not require dictionaries or external corpora. As previously stated, the TF-IDF approach necessitates the use of a corpus in order to calculate IDF scores for each phrase in the corpus. The YAKE algorithm is domain-independent. It comprises mostly of five steps:

1. *Text Pre-processing and candidate term identification*
2. *Feature Extraction*
3. *Computing Term Score*
4. *N-gram generation and computing candidate keyword score*
5. *Data reduplication and Ranking*

Linguistic Approach:

Language information about texts and the words they include is frequently used in keyword extraction algorithms. To decide which keywords should be extracted, morphological or syntactic information (such as the part-of-speech of words or the relationships between words in a dependency grammar representation of sentences) is sometimes used. Certain PoS (e.g., nouns and noun phrases) are awarded greater scores in some circumstances because they typically include more information about texts than other categories.

Graph-based Approach:

In several methods, a text can be represented as a graph. Words can be thought of as vertices connected by a directed edge (i.e. a one-way connection between the vertices). In a dependency tree, those edges might be labelled as the relation between the terms. Undirected edges may be used in other document formats, such as when encoding word co-occurrences.

If words were represented by numbers, an undirected graph would look like this:

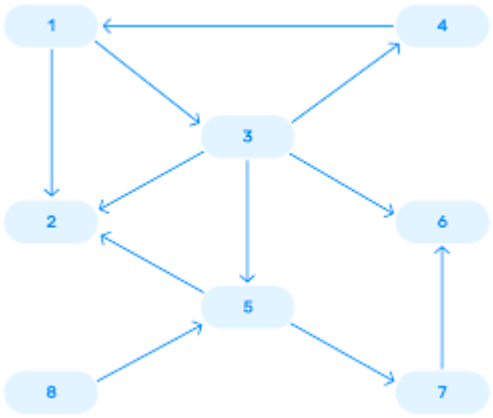

Fig.3 Undirected Graph

Once a graph has been created, the next step is to establish how to value the vertices. The number of edges or connections that land in the vertex (also known as the in-degree) plus the number of edges that start in the vertex (also known as the out-degree) divided by the maximum degree gives the degree of a vertex (which equals the number of vertices in the graph minus 1). The following is the formula for calculating a vertex's degree:

$$\mathbf{D_v = (D_v^{in} + D_v^{out}) / (N - 1)}$$

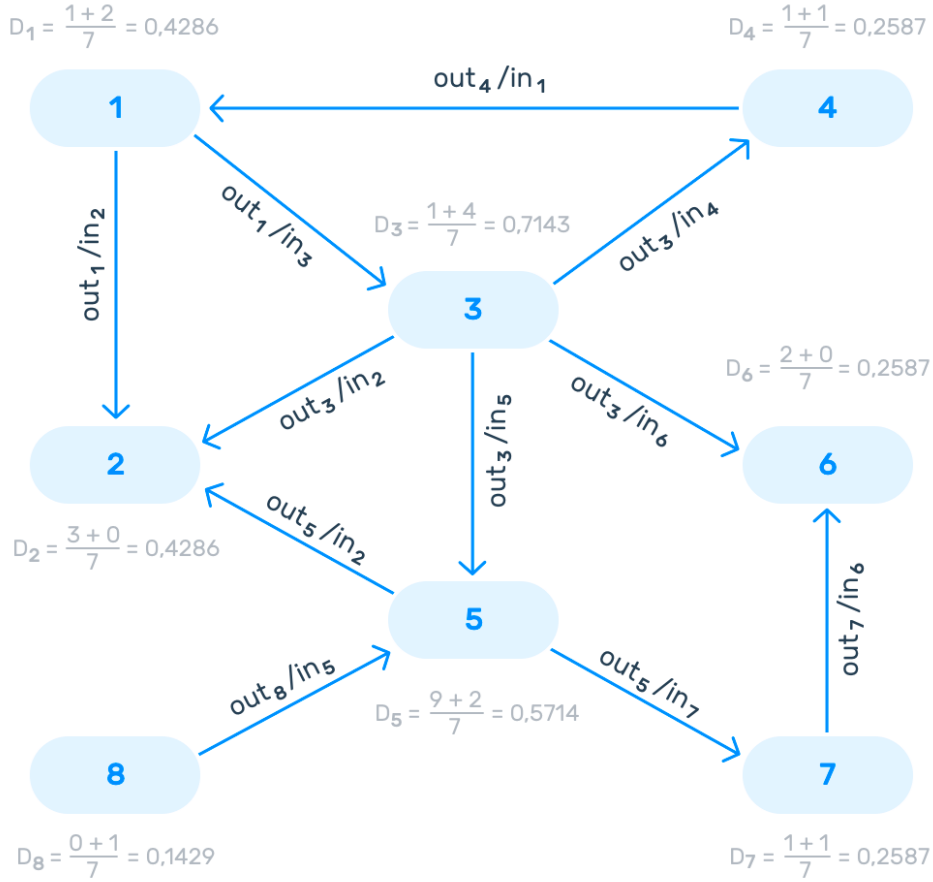

Fig.4 Calculating degree of Vertex

II. Conclusion:

There are many approaches to keyword extraction we can gather unstructured data and perform keyword extraction using any of the discussed approaches. Every approach has a different technique, unique formula and statistics to generate the right and most suitable keywords. This article was an overview of keyword extraction and its techniques and we will be working further on the practical implementation of some of the algorithms from every approach, comparing them and generating the evaluation metrics to see the most suitable algorithm.

References:

[1]. "Simple Guide to Keyword Extraction" – src. AnalyticsVidhya.com
[2]. Stuart Rose, Dave Engel, Nick Cramer (2010) "Automatic keyword extraction from individual documents" Text Mining: Applications and Theory (pp.1 - 20).
[3]. "Keyword Extraction: A guide to finding keywords in text" – src. Monkeylearn
[4]. Slobodan Beliga "Keyword extraction: a review of methods and approaches" University of Rijeka, Department of Informatics
[5]. Thushara Mg, Tadi Mownika, Ritika Magamaru "A Comparative Study on different Keyword Extraction Algorithms" presented at the 3rd INTERNATIONAL CONFERENCE ON COMPUTING METHODOLOGIES AND COMMUNICATION (ICCMC 2019)At: Surya Engineering College (SEC), Erode, India.
[6]. Er. Tanya Gupta "Keyword extraction – A Review" Published in International Journal of Engineering Applied Sciences and Technology, 2017 Vol. 2, Issue 4, ISSN No. 2455-2143, Pages 215-220