ISSN (Online): 2455-7838
SJIF Impact Factor : 6.093

EPRA International Journal of

Research & Development (IJRD)

Monthly Peer Reviewed & Indexed
International Online Journal

Volume: 4, Issue:1, January 2019

Published By
EPRA Publishing

SJIF Impact Factor: 6.093 Volume: 4 | Issue: 1 | January | 2019 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Peer Reviewed Journal

AN IMPROVED TEXT SUMMARIZATION USING FEATURE SELECTION AND OPTIMIZED NAIVE BAYES CLASSIFICATION COMPARED WITH LATENT DIRICHLET ALLOCATION

K.Gowri
Research Scholar, Department of Computer Science, NGM College, Pollachi, India

Dr. R.Manicka Chezian
Associate Professor, Department of Computer Science, NGM College, Pollachi, India

ABSTRACT

Perceptive the contents of a document via a text summarized version of the document needs a shorter time than reading the complete document, so the outline text becomes important. Report needs a great deal of your time and price once the documents square measure varied and long document. Therefore, automatic report needed to beat the matter of reading time and price. The propose options choice square measure the cornerstone within the generation method of the text outline. The outline quality is sensitive for those options in terms of however the sentences square measure scored supported the used options. The automated text categorization, a perfect task-specific outline will be narrowly outlined because the set of most-informative options selected specifically with the categorization performance in mind. The propose system have 3 part, initial pre-processing document supported porter and Lancaster methodology to get rid of the unwanted words from document. The second methodology feature choice supported completely different sort feature choice to weight every term. The Pruning techniques are propose victimization ignore the feature supported TF and DF to additional scale back the set of potential options words inside a document before applying a technique of feature choice. Finally classify the chosen feature supported optimize navie mathematician algorithmic program. The benchmark collections were chosen because the test beds: Reuters-21578. The experimental result show higher exactitude and recall compare with existing algorithms.

KEYWORDS: -*Text summarization, pre-processing, Feature Selection, Text Classification*

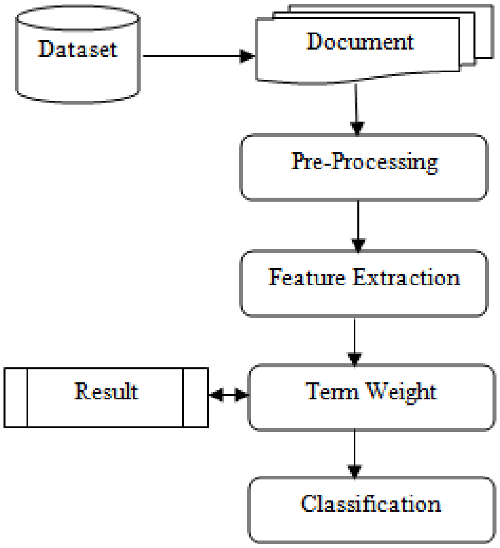
INTRODUCTION

Text account is that the drawback of making a brief, accurate, and fluent outline of an extended text document. Account may function a motivating reading comprehension take a look at for machines. To summarize well, machine learning models got to be able to comprehend documents and distill the necessary data, tasks that area unit extremely difficult for computers, particularly because the length of a document will increase. Text account is that the method of manufacturing shorter presentation of original content that covers no redundant and salient data extracted from one or multiple document. A outline will be outlined as a text that's created from one or additional texts, that contain a major portion of the knowledge within the original text(s), which is not any longer than 1/2 the first text(s).

Automatic text summarization involves.

- Elimination of redundancy: The sentences within the text that convey constant that means are afore said to be redundant and might be eliminated within the outline.
- Identification of serious Sentences: outline being a shorter illustration of text needs together with solely salient sentences from the first document.
- Generation of Coherent Summaries: Sentences hand-picked for summarisation must be ordered and classified so coherence and readability is maintained.
- Metrics for evaluating the mechanically generated Summaries: In most of the cases the standard of the outline is judged by humans and thence automatic analysis could be a fascinating feature.

There are 2 general approaches to automatic summarization: extraction and abstraction. Extractive ways work by choosing a set of existing words, phrases, or sentences within the original text to create the outline. In distinction, theoretic ways build an inside linguistics illustration and so use linguistic communication generation techniques to make an outline that's nearer to what a personality's may specific. Such an outline may embrace verbal innovations. Analysis so far has centered totally on extractive ways that are applicable for image assortment summarisation and video summarisation.



The above mentioned figure shows the text summarization over flow diagram, first load the dataset (corpus) than pre-process the document based on stop word or steaming. The feature selection techniques used to weight each term based on frequency, finally apply the text classification algorithm to get the result

RELATED WORK

Automatic summarizers usually determine the foremost necessary sentences from Associate in nursing input document. Major approaches for decisive the salient sentences within the text area unit term coefficient approach [1], symbolic techniques supported discourse structure [2], linguistics relations between words [3] and alternative specialized strategies [4]. Whereas most of the summarisation efforts have centered on single documents, many initial comes have shown promise within the summarisation of multiple documents. The techniques for automatic extraction may be classified into 2 basic approaches [5]. The primary approach is predicated on a collection of rules to pick out the necessary sentences, and therefore the second approach is predicated on an applied mathematics analysis to extract the sentences with higher weight.

Cluster primarily based strategies measures connectedness or similarity between every sentence in a very document therewith of sentences chosen for outline. Summaries address onto completely different "themes" showing within the documents that is incorporated through clump. Clump primarily based strategies become essential to get a significant outline. Documents area unit sometimes written such they address completely different topics one when the opposite in Associate in Nursing organized manner. Graph suppositious Approach illustration is Associate in nursing extractive summarisation model that provides a way to spot themes within the document.

Preprocessing steps, namely, stop word removal and stemming area unit done before, to get graphical read of the documents. Sentences within the documents kind nodes of Associate in nursing afloat graph.

Singular worth Decomposition (SVD) [9] could be a terribly powerful mathematical tool which will realize principal orthogonal dimensions of three-d knowledge. It applications in several areas and is thought by completely different names: Karhunen-Loeve reworks in image process, Principal Part Analysis (PCA) in signal processes and Latent linguistics Analysis (LSA) in text process. It gets this name LSA as a result of SVD applied to document word matrices, team's documents that area unit semantically associated with one another, even once they don't share common words. In automatic summarisation, similarity metrics area unit used for centrality-based context choice and in identification of redundant contexts. In general, similarity measures area unit either corpus-based or knowledge-based. Each of them is employed in extractive summarisation. Corpus-based measures use term frequencies determined in a very corpus to relate contexts to every alternative, whereas knowledge-based strategies utilize predefined linguistics relations between terms obtained from lexical resources.

The selection procedure is to spot a collection of sentences that contain necessary data. 3 criteria area unit optimized once choosing the sentences: connectedness, redundancy and length. Connectedness determines the importance of the data contained in a very outline with relevance the topics coated within the supply documents or a question just in case of query-focused summarisation. Redundancy measures the data overlap between the sentences chosen for the outline. Given a restricted outline length, summarisation systems try and maximize the connectedness whereas minimizing the redundancy. The task of content choice is to spot those sentences within the supply documents area unit value taking into an outline.

Redundancy could be a major issue in multi-document summarisation wherever many documents on identical topic might have a considerable data overlap. Then, the choice of the foremost relevant sentences can yield a collection of sentences with redundant data. Extract that consists of relevant however terribly similar sentences isn't smart. The joint optimization of each connection and redundancy could be an advanced task as a result of properties of individual sentences area unit keen about alternative sentences enclosed within the outline. A number of the sooner multi-document summarisation approaches handle these optimizations on an individual basis.

Traditional analysis studies usually have faith in human subjects, either for making the perfect summaries, or for judging the quality of various summaries. We tend to propose a hybrid approach specifically targeting analysis of the performance of a summarisation technique in automatic text categorization. Within the method, we tend to do outline a perfect outline, but rather than measurement a precise agreement of any given outline with the perfect, we tend to compare the categorization performance obtained with the particular and ideal summaries. Arguably, the planned analysis methodology is quite slender and ignores alternative necessary aspects of an outline.

Recently, several researches handle the difficulty of the options choice (FS) method. Thanks to its importance, FS affects the standard of applications performance [6]. FS aims in distinguishing that options area unit necessary and may represent the information. In [7] the authors incontestable that, embedding FS in a very system might facilitate effectively as follow. FS reduces the spatial property, take away unsuitable knowledge, and take away redundant options. Also, in hand of machine learning method, FS will cut back the number of knowledge that area unit required. Consequently, it improves the standard of system results.

Map Reduce framework is with success utilized for a numbers of text process tasks such as stemming [8], distribute the storage and computation hundreds in a very cluster [9],text clump [10], data extraction [11], storing and taking unstructured data[32], document similarity formula [12], tongue process [13] and pair wise document similarity [14]. Summarizing giant text assortment is a motivating and challenging downside in text analytics. Variety of approaches area unit steered for handling large text for automatic text summarisation [15, 16]. A Map Reduce primarily based distributed and parallel framework for summarizing giant text is additionally conferred.

EXISITING METHOD

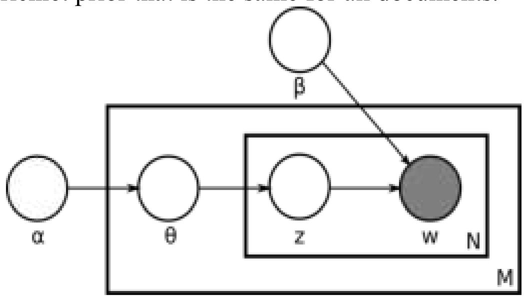
The existing technique is designed using semantic similarity-based clustering and topic modeling using Latent Dirichlet Allocation (LDA) for summarizing the large text collection over Map Reduce framework. The summarization task is performed in four stages and provides a modular implementation of multiple documents summarization.

- The first stage is the document clustering stage where text clustering technique is applied on the multi document text collection to create the text document clusters. The purpose of this stage is to group the similar text document for making it ready for summarization and ensures that all the similar set of documents participates as a group in summarization process.

- In the second stage Latent Dirichlet Allocation (LDA) topic modeling technique is applied on each individual text document cluster to generate the cluster topics and terms belonging to each cluster topic.
- In the third stage, global frequent terms are generated from the collection of multiple text documents.

Latent Dirichlet allocation

Latent Dirichlet Allocation (LDA) is a popular topic modeling technique which models text documents as mixtures of latent topics, which are key concepts presented in the text. A topic model is a probability distribution technique over the collection of text documents, where each document is modeled as a combination of topics, which represents groups of words that tend to occur together. Each topic is modeled as a probability distribution $\phi_k$over lexical terms. Each topic is presented as a vector of terms with the probability between 0 and 1. A document is modeled as a probability distribution over topics in LDA; the topic mixture is drawn from a conjugate Dirichlet prior that is the same for all documents.



Existing graphical representation of LDA model

LDA estimates the topic-term distribution and the document topic distribution from an unlabeled collection of documents using Dirichlet priors for the distributions over affixed number of topics.

$$\iint \prod_{t=1}^{K} P(\theta_t|\beta) \prod_{b=1}^{N} P(\theta_b|\alpha)\left(\prod_{t=1}^{Nb}\sum_{b=1}^{K} P(t_i|\theta)P(w_i|t,\emptyset)\right) \quad d\theta d\emptyset$$

The topic modeling for text collection using LDA is performed in four steps. In the first step a multinomial θ_t distribution for each topic t is selected from a Dirichlet distribution with parameter β. In second step for each document d, a multinomial distribution θ_b is selected from Dirichlet distribution with parameter α. In third step for each word w in documents a topic t from θ_b is selected. And finally, in fourth step a word w from $\theta_t$is selected to represent the topic for the text document.

K-means clustering algorithm

The k-means algorithm is a partitioning based clustering algorithm. It takes an input parameter, k i.e. the number of clusters to be formed, which partitions a set of n objects to generate the k clusters. The algorithm works in three steps. In the first step, k number of the objects is selected randomly, each of which represents the initial mean or center of the cluster. In the second step, the remaining objects are assigned to the cluster with minimum distance from cluster center or mean. In the third step, the new mean for each cluster is computed and the process iterates until the criterion function converges.

Drawbacks of Existing System

- Fixed K (the range of topics is fastened and should be identified before time)
- Uncorrelated topics (Dirichlet topic distribution cannot capture correlations)
- Non-hierarchical (in data-limited regimes stratified models permit sharing of data)
- Static (no evolution of topics over time)
- Bag of words (assumes words area unit exchangeable, syntax isn't modeled)
- Unsupervised (sometimes weak management is fascinating, e.g. in sentiment analysis)

PROPOSE METHODOLOGY

Pre-Processing

Pre-processing is structured illustration of the original inputted text. The importance of pre-processing is employed in almost each developed system connected with text process and linguistic communication processing. Pre-processing phase includes words identification, sentences identification, and stop words elimination, language stemmer for nouns and proper names, permitting input in correct format and elimination of duplicate sentences or words.

Stop Words Elimination

Stop words are a division of natural language. The motive that stop-words should be removed from a text is that they make the text look heavier and less important for analysts. Removing stop words reduces the dimensionality of term space. The most common words in text documents are articles, prepositions, and pro-nouns, etc. that does not give the meaning of the documents. These words are treated as stop words. Example for stop words: the, in, a, an, with, etc.

Word Stemming

Porters stemming algorithmic program is one among the foremost standard stemming several modifications and enhancements are created and prompt on the essential algorithm. It's supported the concept that the suffixes within the West Germanic area unit principally created from grouping of smaller and less complicated suffixes. It's 5 steps, and inside every step, rules area unit applied till one among them passes the conditions. If a rule is accepted, the suffix is removed consequently, and therefore the next step is performed. The resultant stem at the top of the fifth step is come back.

Removing suffixes by automatic suggests that is associate degree operation that is especially useful within the field of data retrieval. During a typical IR atmosphere, one encompasses an assortment of documents, every delineated by the words within the document title and probably by words in the document abstract. Ignoring the issue of exactly wherever the words originate, we will say that a document is represented by a vector of words, or terms.

Paice/Husk Stemmer: The Paice/Husk Stemmer could be a straightforward repetitious Stemmer – that is to mention, it removes the endings from a word in associate degree indefinite number of steps. The Stemmer uses a separate rule file, which is initial scan into associate degree array or list. This file is split into a series of sections, every section love a letter of the alphabet. The section for a given letter, say "e", contains the rules for all endings ending with "e", the sections being ordered alphabetically. Associate degree index will so be designed, leading from the last letter of the word to be stemmed to the primary rule for that letter.

Feature Selection

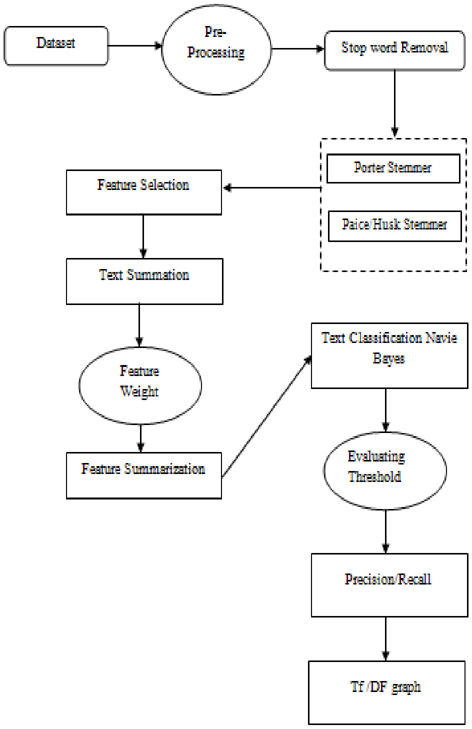
Feature choice plays a crucial role in text categorization. Automatic feature choice strategies like document frequency thresholding (DF), data gain (IG), mutual data (MI), and then on are applied in text summarisation. Feature choice mistreatment Mutual data Feature choice is associate degree particularly important step throughout classification, as a result of digressive and redundant options usually degrade the performance of classification algorithms each in speed and prediction accuracy. Feature choice strategies plan to notice reduced feature sets that minimize the likelihood of error. The estimation functions verify a particular set with discrimination between categories and might be divided into 2 main teams specifically, filter and wrapper. Initially, Filters live the importance of feature subsets that's on an individual basis given with classifier. Similarly, wrappers use the classifier's performance because the analysis operates. Filter is that the most vital method that's disturbed for feature choice than the wrapper method.

Optimized Navie Bayes Classification

The Bayesian classifier to determine if a sentence should be extracted or not. The system was able to learn from data. Some features used by their system include the presence of uppercase words, length of sentence, structure of phrase and position of words. The author assumed the following: s = a certain sentence, S = the sentences in the summary, and $F_1, F_2, \ldots . F_k$the features.

$$P(s \in S|F_1, F_2 \ldots . F_k) = \frac{\prod P(F_j|s \in S)P(s \in S)}{\prod P(F_j)}$$

Th naïve-bayes classifier which used term frequency (tf) which is the number of times that a word appears in a sentences and inverse document frequency (idf) which is the number of sentences in which a word occurs, to know words that hold point at the key concepts of a document.



Proposed architecture diagram for optimized text summarization

S.N O	LDA	KNN	Optimized Navie Bayes
1	linear classifiers	linear classifiers	linear and nonlinear classifiers
2	Gaussian class	Gaussian class	Gaussian with distributional
3	ignoring estimation error	ignoring estimation error	Identify the estimation error based classify
4	continuous-valued features	continuous-valued features	Continuous and non-continuous valued features

Difference between the Existing and Proposed Algorithms

EXPERIMENTAL RESULTS

Even though these numbers aren't adore different results since a set and not the entire Reuters 21578 split was used, they supply still attention-grabbing Insights. particularly the actual fact, that for an equivalent weight perform and therefore the same spatial property, it happens that, e.g., the breakeven worth is higher compared to a different perform however the eleven-point preciseness is lower, compared to an equivalent perform. It conjointly shows that" MSF" might be a stimulating various to chi-square and data gain, not just for feature choice in text classification, however conjointly to weight the importance of options in different classification tasks.

Performance Analysis

The main assessment metrics of co-selection measures are exactness, recall and F-score. Exactness (P) is computed as no. of sentences occurring in each candidate and reference summaries divided by the no. of sentences within the candidate outline. Recall (R) is that the no. of matched sentences in each candidate and reference summaries divided by the no. of sentences within the reference outline. F-score is combination of each exactness and recall. The F-score is nothing however a harmonic average of exactness and recall.

Precision

Precision is the number of True Positives divided by the number of True Positives and False Positives. Put another way, it is the number of positive predictions divided by the total number of positive class values predicted. It is also called the Positive Predictive Value (PPV).

$$Precision = \frac{TP}{TP + FP}$$

Recall

Recall is the number of True Positives divided by the number of True Positives and the number of False Negatives. Put another way it is the number of positive predictions divided by the number of positive class values in the test data. It is also called Sensitivity or the True Positive Rate.

$$Recall = \frac{TP}{TP + FN}$$

TP - True Positive, FN - False Negative, FP - False Positive.

F1 score

It is also called the F Score or the F Measure. Put another way, the F1 score conveys the balance between the precision and the recall.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

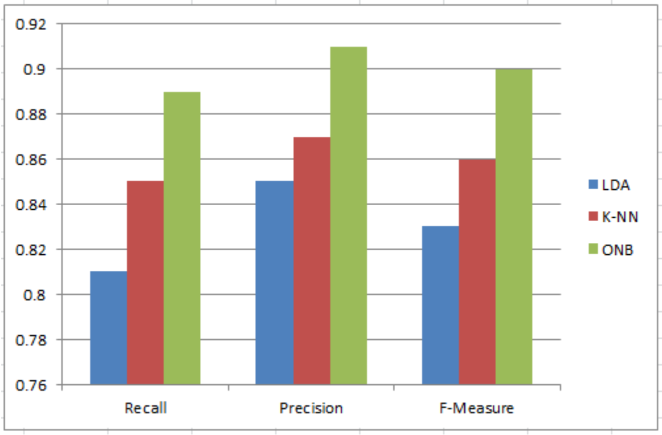
Comparison of Precision, Recall and F1 score

Optimized NB compared with existing algorithm in the context of Precision, Recall and F1 score calculated by the relevant formulas

Comparison table for P/R/F using existing with proposed system

Algorithm	Recall	Precision	F-Measure
LDA	0.81	0.85	0.83
K-NN	0.85	0.87	0.86
Optimized NB	0.89	0.91	0.90

Optimized NB compared with existing algorithm in the context of Precision, Recall and F1 score, The graph is plotted for the measured above context in which green bar shows Optimized NB and Red show KNN, and Blue shows the LDA model.



Comparison graph of Precision, Recall, F-measure for LDA, KNN and proposed optimized NB.

CONCLUSION

The propose options choice are the cornerstone within the generation method of the text outline. The outline quality is sensitive for those options in terms of however the sentences are scored supported the used options. The automated text categorization, a perfect task-specific outline will be narrowly outlined because the set of most-informative options elect specifically with the categorization performance in mind. The propose system have 3 section, 1st pre-processing document supported porter and Lancaster technique to get rid of the unwanted words from document. The second technique feature choice supported totally different kind feature choice to weight every term. The Pruning techniques are propose mistreatment ignore the feature supported TF and DF to any scale back the set of potential options words inside a document before applying a technique of feature choice. Finally classify the chosen feature supported optimize navie mathematician algorithmic rule. The benchmark collections were chosen because the test beds: Reuters-21578. The experimental result show higher exactness and recall compare with existing algorithms.

REFERENCES

1. *J. Salton and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval", Information Processing and Management, vol. 24, no.5, pp.513-323, 1988.*
2. *D. Marcu, "From Discourse Structures to Text Summaries", Proc. of the ACL 97/EACL-97 Workshop on intelligent scalable Text Summarization, pp.82-88, Madrid, Spain, 1997.*
3. *R. Barzilay and M. Elhadad, "Using Lexical Chains for Text Summarization", Proc. of the ACL Workshop on Intelligent Scalable Text summarization, pp. 10-17, Madrid, Spain, 1997.*
4. *D.R. Radev, H. Jing, and M. Budzikowska, "Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies", Proc. of ANLP-NAACL Workshop on Summarization, pp. 21-30, Seattle, Washington, April, 2000.*
5. *C.Y. Lin and E. H. Hovy, "The Automated Acquisition of Topic signatures for Text Summarization", Proc. of the Computational Linguistics Conference, pp. 495-501, Strasbourg, France, August, 2000*
6. *H. Xingshi, Qingqing, Zhang, Na, Sun, Yan, Dong, "Feature Selection with Discrete Binary Differential Evolution," in Artificial Intelligence and Computational Intelligence, 2009. AICI '09. International Conference on, 2009, pp. 327-330.*
7. *R. N. Khushaba, Al-Ani, A., Al-Jumaily, A., "Differential evolution-based feature subset selection," in Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, 2008, pp. 1-4.*
8. *Rajdho A, Biba M (2013) Plugging Text Processing and Mining in a Cloud Computing Framework. In Internet of Things and Inter-cooperative Computational Technologies for Collective Intelligence Springer, Berlin, Heidelberg, Germany, pp 369–390*
9. *Balkir AS, Foster I, Rzhetsky A (2011) A Distributed Look-up Architecture for Text Mining Applications using MapReduce. High Performance Computing, Networking, Storage and Analysis (SC), 2011 International Conference. Seattle, US, pp 1–11*
10. *Zongzhen H, Weina Z, Xiaojuan D (2013) A fuzzy approach to clustering of text documents based on MapReduce. In Computational and Information Sciences (ICCIS), 2013 Fifth International Conference on IEEE. Shiyang, China, pp 666–669*
11. *Chen F, Hsu M (2013) A Performance Comparison of Parallel DBMSs and MapReduce on Large-Scale Text Analytics. Proc. of the 16th International Conference on Extending Database Technology ACM. New York, USA, pp 613–624*
12. *Das TK, Kumar PM (2013) BIG Data Analytics: A Framework for Unstructured Data Analysis. International Journal of Engineering and Technology (IJET) 5(1):153–156*
13. *Momtaz A, Amreen S (2012) Detecting Document Similarity in Large Document Collection using MapReduce and the Hadoop Framework.BS Thesis. BRAC University, Dhaka, Bangladesh, pp 1–54*
14. *Lin J, Dyer C (2010) Data-Intensive Text Processing with MapReduce. Morgan & Claypool Publishers 3(1):1–177*
15. *Elsayed T, Lin J, Oard DW (2008) Pairwise Document Similarity in Large Collections with MapReduce. Proc. of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies. Stroudsburg, US, pp 265–268*
16. *Galgani F, Compton P, Hoffmann A (2012) Citation based summarization of legal texts. Proc. of 12th Pacific Rim International Conference on Artificial Intelligence. Kuching, Malaysia, pp 40–52.*