Modern Applied Science; Vol. 12, No. 9; 2018
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

An Enhanced Approach for Using Data Visualization for Sentiment Analysis and Auto Summarization Data

Osama Mohammad Rababah[1], Esra F. Alzaghoul[1] & Hussam N. Fakhouri[1]

[1] King Abdullah II School of Information Technology, The University of Jordan, Amman, Jordan

Correspondence: Osama Mohammad Rababah, King Abdullah II School of Information Technology, Amman, PC 11942, Amman, Jordan. Tel: 962-6-535-5000. E-mail: O.Rababah@ju.edu.jo

Received: June 29, 2018 Accepted: July 12, 2018 Online Published: August 27, 2018

doi:10.5539/mas.v12n9p190 URL: https://doi.org/10.5539/mas.v12n9p190

Abstract

With the rapid increase in the size of the data over the internet there is a need for new studies for text data summarization and representation; rather than storing the full text or reading the full text we can store and read a summary that represent the original text. Furthermore, there is a need also to represent the summarized text with visual representation; one picture worth ten thousandwords. In this paper we propose an approach for visual representation of the summarized text;visual resources give creative control over how message is perceived andprovide a faster way to know what where the text about.This approach were implemented and tested on a sample of two datasets one of 50 texts and the other dataset of 80 positive and negative movie comments, the evaluation has been done visually and the percent of success cases has been reported, the precision and recall has been calculated.

Keywords: automatic summarizer, data visualization, regular expression, natural language processing

1. Introduction

Auto summarization is the process of reducing a text document with a computer program in order to create a summary that preserves the most important points in the original document. Techniques that can make a coherent summary take into account variables such as length, style of writing and syntax (Alami et al., 2015). Automatic data summarization is part of the automated learning and data extraction.

The goal of automatic text summarization is presenting the source text into a shorter version with semantics (Das and Martins, 2007; Luhn, 1958). The most important advantage of using a summary isreducing reading time. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form (Edmundson, 1969). An Abstractive summarization is an understanding of the main concepts in a document and then expresses those concepts in clear natural language (Albanese, 2013).

There are two different groups of text summarization: Indicative and Informative. Inductive summarization only represents the main idea of the text to the user. The typical length of this type is 5 to 10 percent of the main text. On the other hand, the informative summarization system gives concise information of the main text. The length of informative summary is 20 to 30 percent of the main text (Freitas et al, 2001).

The automatic summarization means an automatically summarized output is given when an input is applied (Dipanjan, 2007). There are initially preprocesses for summarization such as Sentence Segmentation,Tokenization, Removing stop words and Word Stemming. Sentence Segmentation is separating document into sentences (Conroy and O'leary, 2001). Tokenization means separating sentences into words. Removing stop words means removing frequently occurring words such as a, an, the etc. And word stemming means removing suffixes and prefixes. After preprocessing each sentence is represented by attribute of vector of features (Conroy and O'leary, 2001).

Text summarization is the process of filtering the most important information from a set of sources to produce acondensed version for particular users and tasks (Maybury, 1995). Producing a summary that accurately reflects the meaning of the source text is a difficult task. One would not expect the summary to contain all of the information present in the original text, but enough information that conveys the most important concepts from the source. The sections below discuss the types of text summarization and how summarization may be evaluated for usefulness (Barzilay, 1999).

1.1 Types of Text Summarization

There are many factors involved in text summarization and numerous summarization methods. Texts may be summarized by a human as in news story headlines or movie previews, or automatically as done by search engines such as Google and AltaVista (Radev, Fan, & Zhang, 2001).

1.2 Human Summarization

Is currently the most preferred and reliable form of text summarization. News story headlines, movie previews and movie reviews are all examples of human summaries. They are usually considered to be of high quality, coherent and reflective of the source document. However, human summaries are often time consuming and labor intensive to produce (Zhang, Zincir-Heywood, & Milios, 2004).

1.3 Automatic Summarization

Is machine-generated output that presents the most important content from a source text to a user in a condensed form and in a manner sensitive to the user's or application's needs (Mani et al., 2002). Automatic summaries of text documents are faster and less expensive to generate in comparison to human summaries (Gupta et al,. 2012). However, automatic summaries have not achieved the level of acceptance achieved by human summaries, and it has previously been shown that human summaries provide at least 30% better information than automatic summaries (McKeown et al, 2002). Various methods for automatic summarization have been proposed, and large scale evaluations such as the Document Understanding Conference (DUC), (Harman and Over, 2004) and SUMMAC (Mani et al., 2002) have been conducted to judge systems and understand issues with summarization (Liu et al, 2004).

1.4 Single Document Summarization

Is the summarization of only one text document and can be thought of mostly as an aid to information retrieval. When users search for information online, they may require a single document to answer their question or to provide the information they need (Hovy et al, 1999).

1.5 Multi-Document Summarization

Is the summarizing of information from more than once source document. It is thought to be harder than single document in that more information has to be condensed into a single summary and the summary has to be reflective of more than one text source. Some summarizers rank the documents and the sentences within them using current information retrieval technologies (Marcu, 1999)

1.6 Extractive Summaries

Use information directly from the source document(s). Automatic summarizers are more likely to produce extractive summaries than their abstractive counterparts. Many of these rank the sentences contained within a single document or set of multiple documents and use the higher-ranking sentences as the summary. It has also been shown that using the lead sentence or leading characters (the first sentence or first few characters of a document) can provide a relatively good summary (Brandow et al., 1995; Erkan and Radev, 2004). Highly extractive summaries contain only words found in the source document. Less extractive summaries pull information from the source text but add in conjunctive or limited modifying information (Knight and Marcu, (2000).

Abstractive Summaries may contain material not present in the source text. These are more likely to be produced by humans where synonyms, or even entire rephrasing of words appearing in the document(s) may be used to condense the meanings of multiple words into one (Kupiec et al., 1995).

1.7 Indicative Summaries

Identify what topics are covered in source text, and alert the user to source content. These summaries generally provide a few sentences or even a few keywords related to just one information area, sometimes in relation to a topic-based query (Hingu et al., 2015).

1.8 Informative Summaries

Identify the central information about an event (who, what, where, etc.). They may be used as document "surrogates," i.e., they are used to stand in place of the source document(s) when the user has to find information quickly (usually for a question answering task) and does not have time to open the full text. Many tend to include the first sentence of the source document as part of the summary. In newswire text, the first sentence is sometimes introductory, giving a general overview of the contents of the document (Ryang & Abekawa, 2012).

1.9 Compression

Is also an important part of text summarization. Compression determines the size of the summary as a function of the document size. The summarization compression ratio is the ratio of the size of the compressed data to the size of the source data. This is usually set at a specific length for comparison and evaluation of summarization systems. The compression method may apply at the level of sentences, words or characters (Knight & Marcu, 2000)

2. Related Work

The history of automatic text summarization goes back to 1958 where researchers suggested that text summarization by computer was feasible though not trivial. These original algorithms were based on sentence position and word frequency count to select portions of the input text as extractive summaries (Harabagiu and Lacatusu, 2002). Many years later with the advent of the web and large set of online corpora the interest for automated text summarization renewed. New advancement on Natural Language Processing (NLP) and Information Retrieval (IR) techniques plus computers with higher speed and larger memories made more sophisticated algorithms feasible (Hirao et al, 2002). Years later machine learning techniques were applied on a set of natural language processing features to identify the important key part of the input text as a summary. The pioneers were, who developed a summarizer using a Bayesian classifier to combine features from a corpus of scientific articles and their abstracts and who experimented with other forms of machine learning and its effectiveness (Chuang and J. Yang, 2000).

Perhaps the most cited paper on summarization is that of (Luhn, 1958), that describes research done at IBM in the 1950s. In his work, Luhn proposed that the frequency of a particular word in an article provides an useful measure of its significance. this Earliest instances research on summarizing scientific documents proposed paradigms for extracting salient sentences from text using features like word and phrase frequency (Luhn, 1958), position in the text (Baxendale, 1958) and key phrases (Edmundson, 1969). Various work published since then has concentrated on other domains, mostly on newswire data. Many approaches addressed the problem by building systems depending of the type of the required text summary. (Alami, et al., 2015). The summary is defined by Hovy as the following; ”A summary is a text that is produced out of one or more (possibly multimedia) texts, that contains (some of) the same information of the original text(s), and that is no longer than half of the original text(s).” Some research uses the degree of lexical connectedness between potential summary and the remainder of the text. Connectedness is measured by the number of shared words or synonyms.

Another effective approach is to reward sections of the input text that include topic words, that have been determined to correlate well with the topic of the source text. Furthermore they have developed an open-source summarization environment, MEAD, that allows researchers to experiment with different features for an effective summarization. Some work has turned to the use of hidden Markov models (HMMs) and pivoted QR decomposition to reflect the fact that the probability of inclusion of a sentence in an extract depends on whether the previous sentence has been included as well. To automatically summarize user-contributed short text through a process of identifying and extracting key informative content we are inspired by recent efforts at automatic text summarization for creating a compact version of either a single document or a collection of documents.

In the short text ranking method we want to pick up the sentences that are playing the role of summary or abstract of all of the input short text data. Therefore, we look at the first few short text in our ranking as the summary of all input short text. There are studies to provide summaries in the format of the Tag cloud by considering the hidden relationships in the underlying content of the comments. With the use of tag cloud, some of those relationships have been discovered more efficiently. The criticism we make for it is that the tags alone are not reflecting the context. We need to have sentences to understand the main idea of a document. In a study the metrics for document summarization are developed. They first found the relevant sentences in a document and then applied novelty measures to filter the redundant sentences from the collection. There are works on using comments to summarize a document such as a blog. It was showed that the terms appearing in the comments are a good pivot of sentence importance in an article. To discover different aspects of the objects on the web, used user reviews. They have extracted different aspects of a product like a car such as mileage, engine, transmission, and extracted the crowd idea about each of such aspects. This works great with controlled vocabulary seen in e-commerce website such as amazon.com. Many studies apply graph based ranking for single or multi document summarization and they select the top-K sentences as the summaries of the input document(s). Example includes TextRank, LexRank. Similar to Google’s PageRank algorithm or Kleinberg’s HITS algorithm, these methods first build a graph based on the similarity relationships among the sentences in a document. MEAD and LexRank methods have shown good results for single or multi documents summarization giving some pages of concrete articles. The position of the sentences and the similarity of the sentences to the title of the resource are among features that are used. Three default features that come with the MEAD distribution are Centroid, Position, and Length. A centroid is a set of words that are statistically important to a group of documents

While extractive summarization is mainly concerned with what the summary content should be, usually relying solely on extraction of sentences, abstractive summarization puts strong emphasis on the form, aiming to produce a grammatical summary, which usually requires advanced language generation techniques. In a paradigm more tuned to information retrieval (IR), one can also consider topic-driven summarization, that assumes that the summary content depends on the preference of the user and can be accessed via a query, making the final summary focused on a particular topic (Das, and Martins,2007). (Radev et al., 2002), extraction is the procedure of identifying important sections of the text and producing them verbatim; abstraction aims to produce important material in a new way; fusion combines extracted parts coherently; and compression aims to throw out unimportant sections of the text

The semantic – pragmatic gap (Bach, 2004) is one of the main issues in computational linguistic research and query retrieval. The issue consists of being able to create an algorithm or a system that is able to understand the message that a certain phrase or text is trying to communicate. One of the ways to test if this is achieved by a given algorithm is to propose a query system and analyze the results in relation with the original query and decide if they are valid. A recommender system (Massimiliano Albanese, 2013) consists of a tool that is able to offer user personalized information, based on what their profiles are. This can consist of ranking systems or multimedia retrieval algorithms. In the current paper, the recommender system will be based on the auto generated summaries, while the user profiles will be represented by the raw text documents.

(Joel LaroccaNeto, Alex A. Freitas, Celso A. A. Kaestner) address the automatic summarization task. Present a summarization procedure based on the application of trainable Machine Learning algorithms which employs a set of features extracted directly from the original text. Also present some computational results obtained with the application of our summarizer to some well-known text databases, and we compare these results to some baseline summarization procedures. (Dipanjan Das Andr,e F.T. Martins 2007) investigate some of the most relevant approaches both in the areas of single-document and multiple document summarization, giving special emphasis to empirical methods and extractive techniques. emphasizes extractive approaches to summarization using statistical methods. A distinction has been made between single document and multidocument summarization.

3. Research Methodology

In order to achieve the objectives of this proposed approach, the following steps and procedures will be performed:

1- First step was to investigate and study previous methods and researches done in the field of analyzing text summarization and data visualization.

2- constructing the proposed approach steps and flow chart which are shown in figure 1

3- An Implementation for the proposed approach using Python toolkit NLTK, The NLTK (NLTK.org. (n.d.)) Python toolkit was developed as a tool for syntactic processing that contains multiple corpora and dictionaries that can be used for tagging and stemming.

4- Testing the implemented approach and recording the results.

4. Proposed Approach

The rapid and increase of online information and the growing of big data has recommend to start an intensive researches in the field of data visualization and automatic text summarization within the Natural Language Processing (NLP) area. While Summarization is the art of abstracting key content from one or more information sources, Visualization is the text transformation into a visual representation to make it more meaningful and helpful, in fact Visualizations make the textmore valuable when it transform the contents of the text into images or any other multimedia visualization (rather than just duplicate its message or structure it).

In this research wepropose an approach the visualize Large text information processing of the text with natural language processing techniques and after applying auto-summarization of the text to get the most relevant sentences that represent the text. The summarized text will be used to construct the search query that retrieve a visual representation of images that mostly express and represent and relevant to the summarized text. The visualization relevant visual images are retrieved using Google ranked images search engine that matches auto-text summary.

Proposed approach steps and Flowchart

The proposed approach of the consist of fourmain steps:

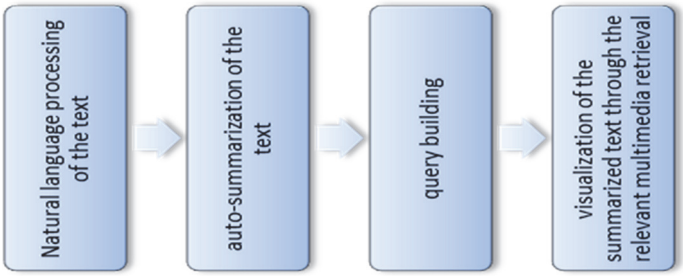

Figure 1. The four stepsare integrated steps that construct the approach.

The auto summarization algorithm consists of 10 steps as shown in figure 2. that start from the original text document and end with the final step of storing the retrieved images.

The second step is tokenization. Which is cutting string into still useful linguistic units. Tokenization has been done in order to reach more advanced text split, which allow controlling of tokenization process. tokenization is done using whitespaces and usage of regular expressions. Because It gives much better control over the process and it can be extended with the usage of text corpora's or even machine learning techniques such, as it can give better results for more complicated text data. Next steps is processing for the words in each sentence (like Stop word elimination). Then Compute term frequency in a text. Next step is Compute sentence relevancy which is mainly we use the calculation method based on similarity. Calculation method based on similarity is generally used of vector space model. like the calculation method of word similarity based on statistical. Where the sentences base are transformed into vectors of characteristic words space, and then the cosine angle between two vectors is used to describe the sentence relevance. Modeling and computing based on vector space are not an accurate reflection of semantic information of query sentence. Next step is we Summarize text by returning the most relevant sentences which is calculated through the computation of relevance between words of sentence and subject words, Next step is Construct the search query from the text summary to provide the user with a generic summary that highlights the most salient information in a text. The query string the used to search google website and retrieve the related image that represent the text summary.

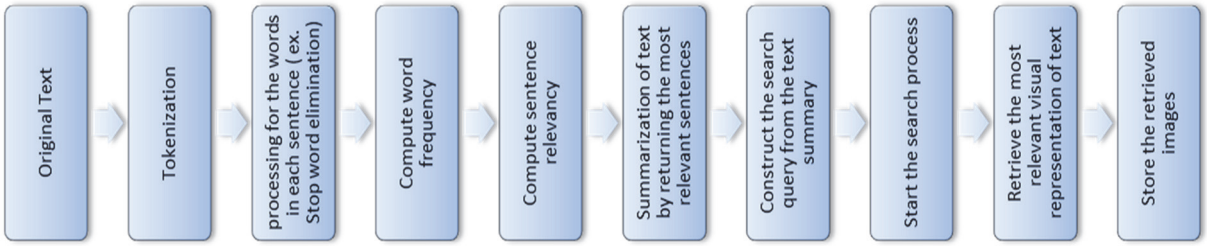

Figure 2. Proposed Approach Flowchart

5. Experimental Analysis

The implemented approach has been tested for two purposes and over towo datasets the first dataset is a group of 50 paragraphs that were collected manually and the second one is a amovie-review data for use in experiments obtained from Bo Pang and Lillian Lee dataset (cs.cornell.edu). the purpose of the first dataset is to examine the ability of the system to visualize a texts which has certain purpose and a clear frequent word such as Petra, sport, flower etc.. and the second one to test the ability of the system to visualize people opinion either positive or negative throw the movie review commentsthe description of the dataset as follows:

- First dataset: The data set represent a 50 text paragraph that has been collected from the internet for paragraphs related to different categories such as tourism places in Jordan such as Petra, sport, flowers, food types etc.., the text pargraph were used to text the implemented system and the visualize results has been obtained a sample of the test cases are described in table 1 :
- Second dataset: An 80 sample of movie-review data for use in experiments obtained from Bo Pang and Lillian Lee dataset (cs.cornell.edu). That are Available as a collections of movie-review documents labeled with respect to their overall sentiment polarity(positive or negative) or subjective rating (e.g.,

"two and a half stars") and sentences labeled with respect to their subjectivity status (subjective or objective) or polarity. Our goal after summarization is to represent the positive feeling and negative feeling of people in mages that represent their status of people classified as shown in figure 3 which represent a sample of people feeling regarding the movie

Figure 3. Sample of classification of people feeling regarding the movie

Test case	Text :	Visual representation
Petra	"Petra originally known to the Nabataeans is a historical and archaeological city in southern Jordan. Petra is famous for its rock-cut architecture and water conduit system. Another name for Petra is the Rose City due to the color of the stone out of which it is carved. Petra is one of the 7 wonders of the world."	
Flowers	A flower, sometimes known as a bloom or blossom, is the reproductive structure found in plants that are floral (plants of the division Magnoliophyta, also called angiosperms. the innermost whorl of a flower, consisting of one or more units called carpels. The four main parts of a flower are generally defined by their positions on the receptacle and not by their function. Many flowers lack some parts	
	Sport (British English) or sports (American English) includes all forms of competitive physical activity or games which, through casual or organised participation, aim to use, maintain or improve physical ability and skills while providing enjoyment to participants, and in some cases, entertainment for spectators. Usually the contest or game is between two sides, each attempting to exceed the other. Some sports allow a tie game; others provide tie-breaking methods, to ensure one winner and one loser.	

6. Evaluation and Relevant Measurement

The evaluation for the proposed approach has been dome visually whereas a human judgment was the reported for the 50 text paragraph. The implemented approach has successfully represented the text visually with 100 % of correct representation

For the dataset of the movie review the visual results has been reported as follows

Real comment	True Positive	False Positive	True Negative	False Negative
p	1	0	0	0
p	1	0	0	0
n	0	1	0	0
p	1	0	0	0
p	1	0	0	0
p	1	0	0	0
p	1	0	0	0
p	1	0	0	0
p	1	0	0	0
p	1	0	0	0
n	0	1	0	0
n	0	1	0	0
p	1	0	0	0
n	0	1	0	0
p	1	0	0	0
n	0	1	0	0
n	0	1	0	0
p	1	0	0	0
n	0	1	0	0
p	1	0	0	0
n	0	1	0	0
p	1	0	0	0
p	1	0	0	0
n	0	1	0	0
n	0	1	0	0
n	0	1	0	0
p	1	0	0	0
p	1	0	0	0
n	0	1	0	0
n	0	1	0	0
p	1	0	0	0
p	1	0	0	0
p	1	0	0	0
p	1	0	0	0
n	0	1	0	0
p	1	0	0	0
p	1	0	0	0
n	0	1	0	0
n	0	1	0	0
n	0	1	0	0
p	0	0	0	1
n	0	0	1	0
n	0	0	1	0
p	0	0	0	1
p	0	0	0	1
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
p	0	0	0	1
n	0	0	1	0

n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
p	0	0	0	1
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
p	0	0	0	1
n	0	0	1	0
p	0	0	0	1
n	0	0	1	0
n	0	0	1	0
n	0	0	1	0
p	0	0	0	1
n	0	0	1	0

The measurement that was used to measure the performance of the s algorithm for all test cases includes Recall, precision and Accuracy:

Recall is the true positivity rate of the classification, also referred to as sensitivity. Recall has been calculated as shown in Eq. 1:

$$\frac{True\ Positives}{TruePositives+FalseNegatives} \quad \text{Eq. 1}$$

The precision, calculated as the positive prediction value, as shown in Eq.2:

$$\frac{True\ Positives}{TruePositives+FalsePositives} \quad \text{Eq.2}$$

Accuracy is the rate with which items have been correctly classified and/or retrieved. Accuracy ahs been calculated as shown in Eq. 3:

$$\frac{True\ Positives+TrueNegatives}{TruePositives+FalseNegatives+TrueNegatives+FalsePositives} \quad \text{Eq.3}$$

The precision and recall values has been calculated for the 80 sample where the Precision is 0.784 and the Recall is 0.632 and the Accuracy 0.7125.

7. Conclusion

Based on the experimental results for this paper and the results obtained, some conclusions can be drawn that the proposed approach was able to successfully visualize and retrieve the images that were relevant to the submitted text; when the purpose of the text was clear such as place or field the results accuracy were 100% but when the text represented an opinion for people that has many frequent terms the accuracy was 0.7125. This can be explained by the fact that the information available on the Internet for people opinion, movies and movies reviews is usually wider and more directly related to the movie itself.

References

Alami, N., Meknassi, M., & Rais, N. (2015). Automatic texts summarization: Current state of the art. *Journal of*

Asian Scientific Research, *5*(1), 1-15.

Bach, K. (2004). Minding the gap. *The Semantics/Pragmatics Distinction*, *27*, 43.

Barzilay, R., McKeown, K., & Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In Proceedings of ACL '99.

Baxendale, P. (1958). Machine-made index for technical literature - an experiment. *IBM Journal of Research Development, 2*(4), 354-361.

Chuang, W. T., & Yang, J. (2000). *Extracting sentence segments for text summarization: a machine learning approach.* In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2000, 152-159.

Conroy, J. M., & O'leary, D. P. (2001). Text summarization via hidden markov models. In Proceedings of SIGIR '01, pages 406-407, New York, NY, USA.

Conroy, J. M., Schlesinger, J. D., Goldstein, J. (2009). CLASSY Query-Based Multi-Document Summarization. In the Document Understanding Workshop (presented at the HLT/EMNLP Annual Meeting.

Das, D., & Martins, A. F. (2007). A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, *4*, 192-195.

Dipanjan, D., Andre, F. T., Martins (2007). A Survey on Automatic Text Summarization, Language Technologies Institute, Carnegie Mellon University, November 21, 2007

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM, 16*(2), 264–285.

Harabagiu, S., & Lacatusu, F. (2002). Generating Single and Multi-Document Summaries with GISTEXTER. In Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization) hiladelphia, Pennsylvania, USA, 2002.

Hingu, D., Shah, D., & Udmale, S. S. (2015). *Automatic text summarization of Wikipedia articles.* In Communication, Information & Computing Technology (ICCICT), 2015 International Conference on. IEEE, 2015, 1–4.

Hirao, T., Sasaki, Y., Isozaki, H., et al. (2002). NTT's Text Summarization system for DUC-2002. In Workshop on Text Summarization (In conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization), Philadelphia, 2002.

Hovy, E., & Lin, C. Y. (1999). Automated text summarization in summarist. In Mani, I., & Maybury, M. T. (Ed.), *Advances in Automatic Text Summarization* (pp. 81).

Knight, K., & Marcu, D. (2000). Statistics-based summarization - step one: Sentence compression. In AAAI/IAAI.

Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In Proceedings SIGIR '95, New York, NY, USA. Retrieved from http://www.cs.cornell.edu/people/pabo/movie-review-data/

Liu, H., & Singh, P. (2004). ConceptNet —- APractical Commonsense Reasoning Tool-Kit. *BT Technology Journal, 22*(4), 211-226.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research Development, 2*(2), 159-165.

Mani, I., Maybury, M. T., Ed. (1999). Advances in Automatic Text Summarization. The MIT Press, 1999.

Marcu, D. (1999). Discourse Trees Are Good Indicators of Importance in Text. In Inderjeet Mani and Mark Marbury, editors, Advances in Automatic Text Summarization.MIT Press, 1999.

Massimiliano Albanese, A. D. (2013). A multimedia recommender system. ACM Transactions on Internet Technology (TOIT), 3.

McKeown, K., Barzilay, R., & Evans, D., et al. (2002). Tracking and summarizing news on a daily basis with the Columbia's Newsblaster. In Proceedings of the Human Language Technology (HLT) Conference. San Diego, CA, 2002.

McKeown, K., Evans, D., & Nenkova, A., et al. (2002). The Columbia Multi-Document Summarizer for DUC 2002. In Workshop on Text Summarization (In conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization), Philadelphia, 2002.

Neto, J. L., Freitas, A. A., & Kaestner, C. A. (2002, November). Automatic text summarization using a machine learning approach. In Brazilian Symposium on Artificial Intelligence (pp. 205-215). Springer, Berlin,

Heidelberg.

NLTK.org (n.d.). (n.d.). Retrieved January 4, 2017, from http://www.nltk.org/index.html

Radev, D. R., Fan, W., & Zhang, Z. (2001). Webinessence: A personalized web-based multi-document summarization and recommendation system. *Ann Arbor*, *1001*, 48103.

Ryang, S., & Abekawa, T. (2012). Framework of automatic text summarization using reinforcement learning. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.Association for Computational Linguistics, 2012, 256-265.

Vikrant, G., PriyaChauhan, S. G. Anita, B., & Shobha, K. (2012). A Statistical Tool for Multi-Document Summrization. *International Journal of Scientific and Research publication, 2*(5), may 2012 ISSN 2250-3153.

Zhang, Y., Zincir-Heywood, N., & Milios, E. (2004, May). Term-based clustering and summarization of web page collections. In *Conference of the Canadian Society for Computational Studies of Intelligence* (pp. 60-74). Springer, Berlin, Heidelberg.