Cape Peninsula
University of Technology

Online Content Clustering Using Variant K-Means Algorithms

by

Tawanda Chowuraya

Thesis submitted in fulfilment of the requirements for the degree

Master of Technology:

in the Faculty of Informatics and Design

at the Cape Peninsula University of Technology

Supervisor: Dr B. Kabaso

Cape Town
November 2019

DECLARATION

I am hereby declaring that this thesis is a compilation of my own work. It represents my own opinions and not necessarily those of the Cape Peninsula University of Technology. The material that has been used in compiling this thesis has been duly acknowledged and mentioned in reference. This thesis neither in whole or in part, has never been submitted to any other institute or university for academic examination towards any other qualification. The research was conducted under the capable supervision of Dr Boniface Kabaso.

Signed  **Date** 16/03/2020

ABSTRACT

We live at a time when so much information is created. Unfortunately, much of the information is redundant. There is a huge amount of online information in the form of news articles that discuss similar stories. The number of articles is projected to grow. The growth makes it difficult for a person to process all that information in order to update themselves on a subject matter. There is an overwhelming amount of similar information on the internet. There is need for a solution that can organize this similar information into specific themes. The solution is a branch of Artificial intelligence (AI) called machine learning (ML) using clustering algorithms. This refers to clustering groups of information that is similar into containers. When the information is clustered people can be presented with information on their subject of interest, grouped together. The information in a group can be further processed into a summary.

This research focuses on unsupervised learning. Literature has it that K-Means is one of the most widely used unsupervised clustering algorithm. K-Means is easy to learn, easy to implement and is also efficient. However, there is a horde of variations of K-Means. The research seeks to find a variant of K-Means that can be used with an acceptable performance, to cluster duplicate or similar news articles into correct semantic groups.

The research is an experiment. News articles were collected from the internet using *gocrawler*. g*ocrawler* is a program that takes Universal Resource Locators (URLs) as an argument and collects a story from a website pointed to by the URL. The URLs are read from a repository. The stories come riddled with adverts and images from the web page. This is referred to as a dirty text.

The dirty text is sanitized. Sanitization is basically cleaning the collected news articles. This includes removing adverts and images from the web page. The clean text is stored in a repository, it is the input for the algorithm. The other input is the K value. All K-Means based variants take K value that defines the number of clusters to be produced.

The stories are manually classified and labelled. The labelling is done to check the accuracy of machine clustering. Each story is labelled with a class to which it belongs. The data collection process itself was not unsupervised but the algorithms used to cluster are totally unsupervised. A total of 45 stories were collected and 9 manual clusters were identified. Under each manual cluster there are sub clusters of stories talking about one specific event.

The performance of all the variants is compared to see the one with the best clustering results. Performance was checked by comparing the manual classification and the clustering results from the algorithm.

Each K-Means variant is run on the same set of settings and same data set, that is 45 stories. The settings used are,

- Dimensionality of the feature vectors,
- Window size,
- Maximum distance between the current and predicted word in a sentence,
- Minimum word frequency,
- Specified range of words to ignore,
- Number of threads to train the model.
- The training algorithm either distributed memory (PV-DM) or distributed bag of words (PV-DBOW),
- The initial learning rate. The learning rate decreases to minimum alpha as training progresses,
- Number of iterations per cycle,

- Final learning rate,
- Number of clusters to form,
- The number of times the algorithm will be run,
- The method used for initialization.

The results obtained show that K-Means can perform better than K-Modes. The results are tabulated and presented in graphs in chapter six.

Clustering can be improved by incorporating Named Entity (NER) recognition into the K-Means algorithms. Results can also be improved by implementing multi-stage clustering technique. Where initial clustering is done then you take the cluster group and further cluster it to achieve finer clustering results.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to sincerely thank my supervisor Dr Boniface Kabaso for patiently guiding me and for his constant support towards this thesis. I gratefully thank my supervisor for his critical analysis and feedback throughout the course of this work. The feedback and critical analysis helped me to stay on course. I would also like to thank Sivaramalingam Kirushanth for literally holding my hand and walking me through this process. Sivaramalingam Kirushanth provided me with unconditional support, moral encouragement and help in many respects. I would also like to thank members of the Informatics and Design Faculty, Information Technology department at The Cape Peninsula University of Technology, for their direct and indirect contribution to my work. I also want to thank many other colleagues and acquaintances I have interacted with, who helped to shape and focus this research. I want to express sincere gratitude to the resources I found on the internet that helped me to understand machine learning.

DEDICATION

I would like to dedicate this work to all the people who have had a role to play in my life. The list includes my parents, siblings, acquaintances, friends and workmates for their support and creating an environment that made me become who I am today. Special mention goes to my past and present teachers and lecturers for nurturing my academic growth. I also mention my family for allowing me space to embark on this academic journey without hindrance.

GLOSSARY

AI Artificial Intelligence

BOW Bag of Words

CBOW Continuous BOW Bag of Words

CPUT Cape Peninsula University of Technology

FN False Negative

FP False Positive

ML Machine Learning

NER Named Entity Recognition

NLP Natural Language Processing

PV DBOW Paragraph Vector Distributed Bag of Words

PV DM Paragraph Vector Directory Memory

PV-DBOW Paragraph Vector Distributed Bag of Words

PV-DM Paragraph Vector Direct Memory

RI Rand Index

ROC Receiving Operating Curve

RSS Rich Site Summary

SG Skip Gram

SGNS Skip-gram Negative Sampling

TF.IDF Term Frequency Index Document Frequency

TN True Negative

TP True Positive

URL(s) Universal Resource Locator

CHAPTER ONE
Introduction

1.1. Introduction

The development and improvement of the internet and technology has seen an increase in the amount of information that is available to the internet users (Fitriyani & Murfi, 2016). There are millions of pages of information on each topic on the internet (Ross & Wolfram, 2000). The amount of information is projected to grow. In 2008, Google found about one trillion new links (Mulwad et al., 2010), which in 2001 were one billion (Li et al., 2002). A study by IBM and CISCO shows that we are generating about 2.5 Quintillion bytes of data every day and it is estimated to grow to 40 Yottabytes by year 2020 (Ranjan et al., 2016).

The websites on the internet have become the most common way to share information (Piskorski et al., 2011). The type of information most accessed on the internet is textual data. It is one of the main sources of information. Text information is a valuable resource and the analysis thereof is important and valuable (Gong et al., 2011).

Internet data is mainly in the form of blogs, social media and web news articles. Of these, web news articles are the most used means of sharing recent world events and it is an easily accessible media for keeping abreast with world events (Mahmud et al., 2018). Research has shown that there was a growth of online news articles by about 20% between 2012 and 2014. This growth results in large volumes of digital news stored in repositories. The huge volumes of news and text data are more than what an average human being can process (Forsati et al., 2008; Ong et al., 2005). Study has shown that 85% of information that is generated, is not utilized (IBM, 2018).

The growth of news articles can be attributed to ease of access to the internet, low cost and publishing freedom (Sun et al., 2002). The normal editing, proper checks and balances and journalistic ethics are not followed when it comes to internet publishing. Anyone who has access to a computer and internet or mobile devices that are data capable (Biscuitwala et al., 2013) can publish on the internet (Jacobson, 2000).

The publishing freedom of the internet has led to duplication of stories (Azzopardi & Staff, 2012). People just take stories from other sites, change the setting of the story then re-publish it (Pal & Gillam, 2013). Other people cover the same story from a different angle (Redden & Witschge, 2010). The resultant effect of internet news duplication is growth

of digital repository, and flooding of news articles (Messina & Montagnuolo, 2009). The internet is full of stories that discuss the same events.

As a result of the growth of online news articles and flooding of the internet, there is a need for an algorithm that can cluster news articles (Azzopardi & Staff, 2012). News article clustering refers to unsupervised assignment of news articles into groups. The groups are such that news articles in one group are related, while news articles in another group are not similar to the other groups (Xiong et al., 2009).

News articles can be clustered based on term frequency statistics, thus news articles with similar terms can be placed in the same cluster. Terms in news articles are compared, the similarity of terms will indicate if the news articles share the same topic. News articles are represented as term vectors. Distinct terms that appear in a news article space, are contained in a document vector (Jing et al, 2007). Each entry in the news article vector constitutes a measurement called term frequency. The term vectors of the news articles are then compared to each other to calculate the threshold of similarity. Cosine is an example of a similarity measure. The similarity threshold is the one that is used to place a news article in a cluster (Singh et al, 2011).

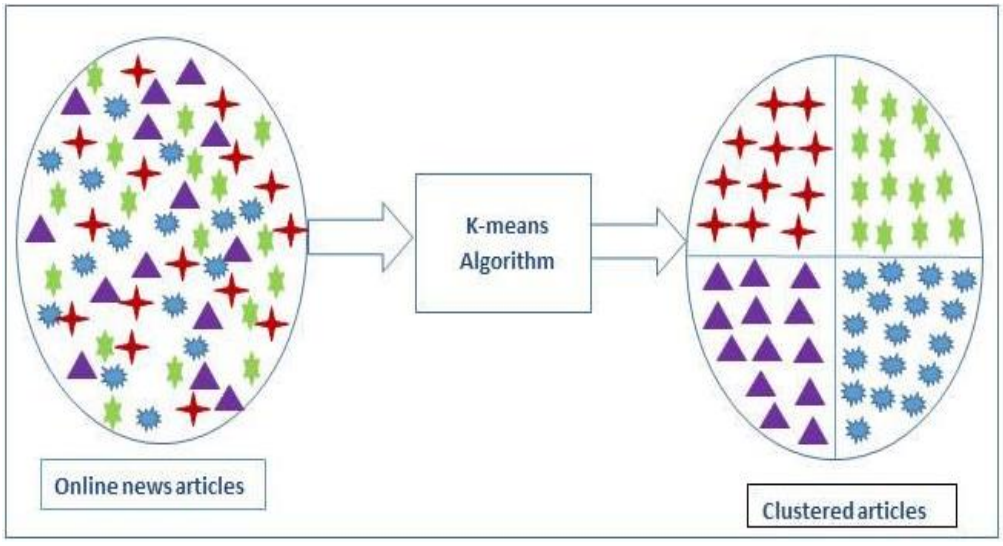

Figure 1.1: The clustering process

Figure 1.1 shows the clustering process. News articles are mixed on the internet. They are passed through an algorithm and separated into groups representing similar articles.

1.2. Background to the research problem

The internet has increased news production because of its speed and extended coverage, and it is an agent of diverse multiplicity news. It has also enabled public participation in news production because of its interactivity (Redden & Witschge, 2010). These factors have contributed to the increase in the amount of online news. However, much of the news articles are the same (Paterson, 2006). News duplication is rife on the internet.

The duplicated news is very annoying to the user (Henzinger, 2006). The duplication makes it difficult for a user to process information presented to them (Kumaran & Allan, 2005). A user has limited reading capacity, yet information continues to grow. This inability to process all the information creates knowledge gaps.

Duplication arises when the publisher issues follow up stories, or publishes a story and keeps updating it, or the same story is covered from different angles or uses different headings but referring to the same event (Redden & Witschge, 2010).

Duplication results in information overload of the internet. Information overload is another motivation for news clustering (Azzopardi & Staff, 2012). Clustering news articles will reduce the time wasted sifting through similar news articles and help readers to find what they are looking for easier. With the help of search engines, relevant news articles can be made to appear at the top of the search results. It improves efficiency because the news articles will be found in predefined clusters (Sahani et al., 2013).

Clustering is a ML technique that involves grouping of data. ML is a branch of AI (Buchanan, 2019), one of the concerns of AI is building algorithms that enable computers to learn on their own (Shabbir & Anwer, 2018). The algorithm is given data and learns to make models on its own without human intervention. Some of the categories that ML can be classified into are supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

1.3. Supervised learning

Supervised learning is a task in ML of deducing a meaning from training data. The training data is a set of training examples. Each example is made up of a vector of an input object and an output value (Al-rubaie & Chang, 2018). A supervised learning algorithm analyses training data to produce an inferred meaning, which is called a classifier or a regression meaning. The function or meaning that the algorithm produces

from the training data should be able to be generalized. This generalization is what is referred to as concept learning in human beings and animal psychology (François-lavet et al., 2018).

1.4. Unsupervised learning

In unsupervised learning there is no training data, the algorithm consumes input data with the aim of deriving a summarised version of the data (Hall et al., 2014). The data is then categorised into similar subsets (Xiong et al., 2009). Unsupervised learning is finding pattern in unstructured and noisy data. With unsupervised learning there is no external entity to perform the adjustments to the system weights. In most cases it is not known what result the system will generate. The system makes necessary adjustments according to the given data and decides what result to produce. Results of unsupervised learning are a new explanation or representation of the observation data, which will lead to improved responses or decisions (Oja, 2002).

Unsupervised learning performs the clustering task without prior knowledge of the structure of the data to be clustered. The fact that there is no prior knowledge of the structure of data, is what makes clustering an unsupervised learning task (Grira et al., 2004).

1.5. Semi-supervised learning

Semi-supervised learning uses mainly unlabelled data and labelled data. The small amount of labelled data is used to increase efficiency of the algorithm (Li & Liang, 2019).

1.6. Reinforcement learning

Reinforcement learning uses input data from the environment to inform the model how to react. Feedback is generated by punishment and rewards from the environment, and not through training like in supervised learning (Kaelbling et al., 1996).

1.7. Research focus

This research focuses on unsupervised learning. The advantage of unsupervised learning is that it can work in an environment where the researcher does not have prior knowledge, or structure of the news articles to be clustered (Sathya & Abraham, 2013). Supervised learning, semi-supervised learning and reinforcement learning require labelled data (Oliver et al., 2018). Labelled data is prior knowledge or information that informs the algorithm on how the clusters will be formed. They are not suited to work with

unlabelled data, where it is not possible to have prior knowledge of the news articles before clustering.

1.8. Unsupervised learning categories

The two broad categories that unsupervised learning can fall in are Hierarchical and Partitional algorithms.

1.9. Hierarchical clustering

Hierarchical clustering is represented in a tree like structure, often called a dendrogram. It shows nested groups in patterns and similarity levels at which groups change. Further clusters can be created by breaking the dendrogram at different levels (Zhao & Karypis, 2002). Most hierarchical algorithms stem from two popular methods, which are single-link and complete-link. In single-link the distance between two clusters is the minimum of the distances between all pairs of patterns drawn from the two clusters. In the complete-link algorithm, the distance between two clusters is the maximum of all pairwise distances between patterns in the two clusters (Saad et al., 2012).

1.10. Partitional clustering

Partitional clustering assigns news articles into unique clusters. A data set with a certain number of news articles will be divided such that the number of clusters formed may be equal or less than the number of news articles. A cluster will contain at least one news article. Each article will belong to only one cluster (Kutbay, 2018).

Partitional algorithms can produce a single partition, instead of a tree cluster that is obtained from the hierarchical technique. Partitional techniques perform better in applications with huge data sets (Jain et al., 2000).

K-Means is one of the algorithms that fall under the partitional algorithms group. It is one of the most popular, widely used and studied unsupervised clustering algorithms. Its popularity is due to its simplicity, ease of implementation and efficiency (Jain, 2010).

K-Means algorithm is used to cluster online news articles (Bradley & Fayyad, 1998; Owen & Owen, 2012). The K-Means algorithm has some weaknesses. Some of the weaknesses are the number of clusters has got to be defined beforehand. It is sensitive to outliers. The initial grouping tends to have a significant influence on the clusters, if there is little data (Teknomo, 2006). The weaknesses have led to the development of many variants to improve the algorithm.

1.11. Statement of the research problem

Duplication of online news articles increases the amount of information on the internet which leads to users being presented with monotonously similar results. Research has found out that K-means is the best algorithm to cluster online news articles (Jain, 2010). The challenge is finding the best variation of K-means that will produce better performance to cluster news articles, on a given set of news articles, number of clusters and number of iterations. Finding an algorithm with good performance to cluster news articles, using a specific variation for a given number of news articles and number of clusters is a challenging task.

News summarization can be performed after the successful implementation of an efficient clustering algorithm to cluster duplicate news articles scattered over the internet. Once the news articles are put in pre-defined containers, it will be easier to summarize them. Implementation of an algorithm with good performance to cluster online content into containers would manage the duplication of news articles on the internet.

1.12. Aim and objectives of the research problem

The aim of the research is to explore an algorithm that can be used to accurately cluster duplicate or similar online content into an accurate theme or topic using k-means unsupervised learning.

The objectives are as follows:

1. To compare the performance of variants of K-Means algorithms.
2. To filter the vast amounts of information by means of clustering.
3. To cluster articles that are similar.

1.13. Main research question

Given a set of news articles and K-means variations, how can we find the best variant with good performance to cluster news articles?

Formally: We are looking for a mapping to take a set of news articles (X), number of clusters (N), the K-Means variant algorithm (V) and the number of iterations (I), which can produce the best clustering results.

1.14. Methodology

The method used to answer the research questions was an experiment. There are independent and dependant variables that were manipulated, and results observed. The methodology is discussed in more detail in chapter three.

1.15. Research flow

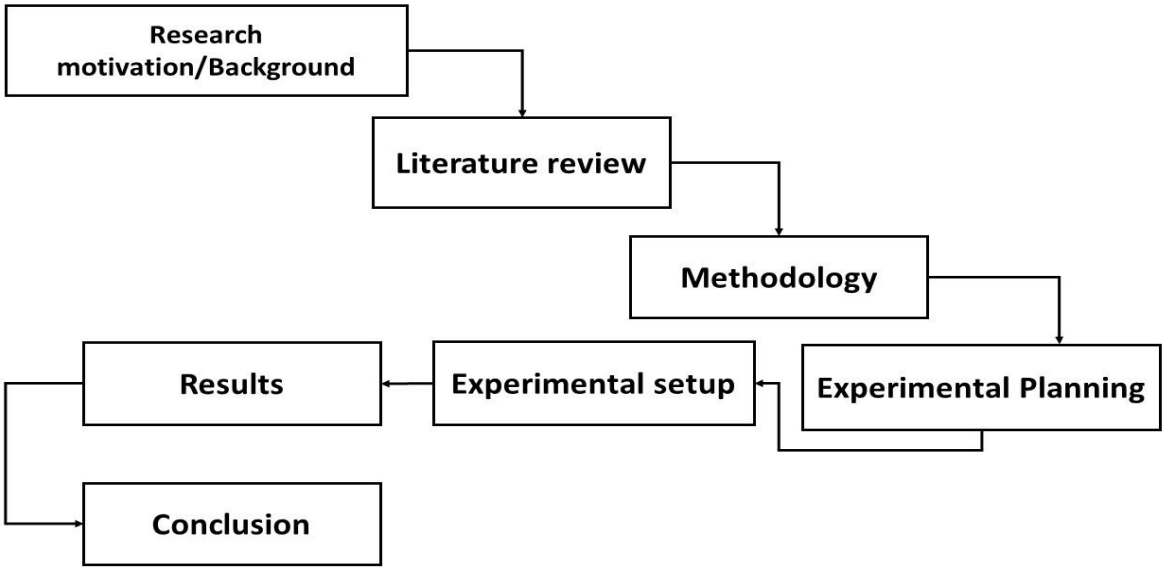

Figure 1.2: Research flow

The Figure 1.2 above gives an outline of how the research was carried out. Research motivation is informed by observation and experiences. The motivation leads to research questions. Research motivation and research questions are guided by literature.

1.16. Delineation of the research

The research looked at clustering online news articles using unsupervised learning and was delineated to K-means and variant algorithms. Clustering is a very broad and vast subject, with many algorithms and techniques. They would need a lot of time and resources to explore. However, the research was constrained in resources, hence it delineated its focus to look at a smaller scope of the subject so that it can be completed within a reasonable time.

1.17. Research contribution

The research is an input to the content summarization model, accurate summarization of content can happen if data is accurately clustered. The experiment will identify the variable set of K-Means variation (V), number of clusters (N) and number of iterations (I) that can be used to produce optimum performance. This knowledge can be used for the development of content summarization algorithms. When content has been correctly clustered it becomes easy to summarize it with an algorithm.

1.18. Ethical consideration

The research was carried out in an ethical and responsible manner. Ethics refers to norms for conduct that distinguish between acceptable and unacceptable behaviour. An ethically conducted research can be replicated and is generalizable (Shamoo & Resnik, 2009). Basic ethical and legal principles underlie all scholarly research and writing to ensure the accuracy of scientific knowledge, to protect the rights and welfare of research participants and to protect intellectual property rights (Roberts, 2010).

Researchers follow principles and updates established by their professional associations. These principles should also be observed in the design and implementation of research involving experimentation. Ethics is avoiding research misconduct. Misconduct is fabrication of research results, plagiarism and fabrication of data. Ethics ensures protection of the rights of participants such as anonymity, and the protection of vulnerable populations (Bornmann, 2013).

Before undertaking an experiment or any research for that matter, certain things must be observed such as, soliciting informed consent from the participants. The data collected should be anonymous. The participants should have the right to opt out of the research at any time they feel to do so. The integrity of the research should be safeguarded. The participants must be protected from emotional, physical and mental harm. Ensure that the wellbeing and privacy of participants is safeguarded.

Participants are safeguarded by using informed consent. Informed consent is an enrolment form that the participant signs to state that they have willingly agreed to participate in the research. The participant must be informed of factors such as the benefits and risks of the research. They must also be informed of the purpose of the research and the fact that they are free to withdraw participation at any time if they so wish (Shahnazarian et al., 2013).

This research complied with ethical principles and requirements of the Informatics and Design Faculty of the Cape Peninsula University of Technology (CPUT). It also complied with the general principles of experimental research. It does not manipulate the processes of data collection and analysis. The research used open source software and as such must comply with terms and conditions thereof. The news articles came from freely available public news sources. There was no ethical clearance needed before collecting the news articles.

Participation of social actors was not required, since the researcher was the sole participant in the research. There was no ethical clearance needed before commencing the experiment. There was no need to do informed consent enrolment since there is no external participants.

1.19. Thesis overview

Chapter one gives an introduction and background to the research problem. It discusses the statement and the aim of the research problem. It gives an outline of how the research was carried out. The delineation of the research, the research contribution, ethical consideration, voluntary participation and informed consent are discussed in chapter one.

Chapter two presents a systematic literature review. It discusses the review protocol to be followed in conducting the systematic literature review. It outlines an inclusion and exclusion policy which determines which studies will be selected. The chapter will also lay out a foundation for the research by explaining and defining the relevant terms used in this thesis.

Chapter three discusses the methodology and research process that was used in this investigation. It also discusses the motivation for the research, research questions, conceptual framework, data collection and generation.

Chapter four discusses the experimental planning. This includes the goals of the experiment, overview of how the goal was achieved, the participants and materials of the experiment, the specific tasks involved in the experiment and procedure of how the tasks are executed and deviation from the plan if any.

Chapter five discusses the experimental set up. Experimental set up is the outline of the experiment, how the data for the experiment is collected and passed to the clustering models and the parameters of the models used.

Chapter six discusses the results and findings of the research. The discussion precedes with the metrics that are used for evaluation. The results obtained are then discussed. Chapter six concludes by discussing the findings and future direction of the research.

Chapter seven is a summary of the thesis.

CHAPTER TWO

Literature review

2.1. Introduction

This chapter presents a systematic literature review. The systematic literature review was done in accordance to a planned protocol. A carefully laid out inclusion and exclusion policy was used to determine which studies would be selected. The purpose of the systematic literature review is to identify and evaluate available research that pertains to this research. Before discussing the systematic literature review, the chapter lays out the foundation of the research by explaining and defining some terms relevant to the research and thesis. The definitions will help the reader to have a better understanding of the research.

2.2. Artificial intelligence

AI is a term that is coined using two words which are artificial and intelligence (Russell & Norvig, 2003). Cambridge English dictionary defines artificial as something made by human beings to mimic something that exists naturally. It also defines intelligence as the ability to understand, learn and arrive at opinions and judgements based on reasoning. Coppin (2004) defines intelligence by the properties it exhibits. The properties include the ability to deal with new problems, new situations and the ability to come up with a plan related to a situation and to answer questions.

There are two definitions that this thesis will adopt. Shaikh & Fegad (2013) give the following definitions: “Artificial intelligence is the study of systems that act in a way that to any observer would appear to be intelligent.” and “Artificial Intelligence involves using methods based on the intelligent behaviour of humans and other animals to solve complex problems.”

The latter definition is a more fitting description of the attempt of this research. The work of AI started around 1950s. In 1950 Alan Turing published an article titled Computing Machinery & Intelligence. Shi (2011) discusses that since the inception of AI more than 60 years ago, its goal has been to build machines with human level intelligence. In other words, the development of intelligent systems and machines that can emulate, extend and expand human intelligence and exhibit intelligent behaviour. Shi further discusses that AI has had a lot of progress, especially in the fields of data mining, expert systems, natural language processing, robotics, and other applications related to ML applications (Nilsson, 2014). These have brought about some social and economic benefits (Robertson et al., 2018).

2.3. Machine learning

The last few decades have seen an increased amount of information available in digital form and online. The increase is due to advancement in computational hardware power and advancement in software, which enable us to generate, transmit and store large amounts of information (Witten et al., 2016).

This has necessitated the need for the branch of AI called ML to cluster and organise the information for easy access and retrieval (Sebastiani, 2002). ML is the ability of computational algorithms to learn from their environment and emulate human intelligence (El Naqa & Murphy, 2015).

The study of ML has grown from efforts of computer science and engineers, who were experimenting to see if computers can learn to play games, to learning algorithms that are used in speech recognition, computer vision and many other tasks. It has also seen a growth in the study of data mining, to discover hidden patterns in the ever-growing online data (Mitchell, 2006).

The high volumes of information and news articles on the world wide web has seen an increase in research focusing on development of algorithms that can process news. Progress in this respect will have benefits for applications that have to do with machine translation, speech recognition, information filtering, information retrieval, pattern detection of large datasets, knowledge extraction and online news clustering (Hofmann, 2001).

2.4. Clustering

A cluster or group is made up of objects that are similar. The objects in one group are different to objects in another group. A large amount of news articles is represented by a few clusters, this achieves simplification of the news articles (Rai & Singh, 2010). News article clustering is a search for underlying patterns in the set of news articles. The search for underlying groups in news articles is unsupervised learning. The result of the groupings is a new data model. Clustering is then described as unsupervised learning of an underlying data model (Berkhin, 2006).

2.5. Document vector

A document cannot be passed to a clustering algorithm in plain text. The document must be converted to a numerical value. The document vector model that was used for this experiment is Doc2vec. Doc2vec uses two models which are Paragraph Vector Direct

Memory (PV-DM) and Paragraph Vector Distributed Bag of Words (PV-DBOW). The document vectors can be passed to the clustering algorithms.

2.6. K-Means Variants

Clustering of news articles can be achieved using K-Means algorithm. There are many variants that were developed from the popular K-Means algorithm. Some of the variants are K-Median, k-Harmonic means, k-SVMeans, K-Modes and Weighted K-Means.

2.7. K-Median

K-Median minimizes the 1-norm distance between each point and the closest cluster centre (Whelan et al., 2015) in comparison to K-Means which uses squares of 2-norm distances to generate cluster centres (Bradley et al., 1997). The median is a statistic that is not easily affected by outliers, the median can only be affected by outliers, when about 50% of the data is tainted. k-Median algorithm places each point in the data set to its closest centre. The points put in the same centre will form a cluster. They are also put in a disjoint set. The new disjoint set is used to recalculate and update the cluster centre. A sum of distances from each point and respective cluster centres is calculated and forms a new epsilon. The iteration is carried out until improvement to epsilon is better than the one previously determined.

2.8. K-Harmonic Means

K-Harmonic Means is an algorithm that iterates and improves the clusters given by K centres, at each iteration. K-Harmonic Means approach is different from K-Means, in that K-Harmonic sums all data points of the harmonic average, of the squared distance from a data point to all the centres as its performance function, unlike K-Means which sums the with-in cluster variance (Zhang et al., 1999).

2.9. K-SVMeans

K-SVMeans groups datasets using heterogeneous similarity characteristics. The K-SVMeans will cluster one dimension of the data while at the same time it is learning a classifier in another dimension, this influences the intermediate cluster assignment decision on the original dimension. K-SVMeans is a hybrid algorithm that combines two clustering solutions, it is made up of K-Means and Support Vector Machines (SVM). Support Vector Machines is a popular supervised learning solution that is effective with text classification tasks (Bolelli et al., 2007). K-SVMeans is a hybrid algorithm, it combines an unsupervised algorithm with a supervised algorithm. This eliminates the need for labelled training examples for Support Vector Machine learning. The K-Means

cluster assignments are used to train an Online Support Vector Machine in the secondary data type, and the Support Vector Machine has effect on the clustering decisions of K-Means in the primary clustering space. This clustering style of heterogeneous datasets increases the clustering performance in comparison to clustering using a homogeneous data source (Deshmukh, 2014).

2.10. K-Modes

K-Modes algorithm was first published in 1997 and has become a common method for solving categorical data clustering problems in diverse application domains. The K-Modes algorithm modifies the K-Means algorithm to use a simple matching difference measure for categorical objects, it uses modes instead of means for clusters, and a frequency-based method to update modes in the clustering process to minimize the clustering cost function. The modification has removed the numeric-only restriction of the K-Means algorithm and enable the K-Means clustering procedure to be used to successfully cluster huge categorical data sets from real world databases (Ng et al., 2007).

2.11. Weighted K-Means

Weighted K-Means algorithm with distributed centroids was developed to cluster data sets ranging from numerical, categorical and mixed type data sets. The approach of this proposal allows given features, such as variables, to have different weights at different clusters. It supports the intuitive idea that features have different degrees of weight at different clusters. The idea is that feature weights become feature re-scaling factors for any considered exponent (de Amorim & Makarenkov, 2016).

Despite all the algorithms discussed above, clustering online news articles is still a challenge. The challenge is that some news articles may have the same title or theme but talk about different events. Example is one article may be talking about President Jacob Zuma opening parliament while another article talks about President Jacob Zuma visiting a prison facility. When clustering these two articles they can fall into one cluster, because the common term is President Jacob Zuma. There is a need for a solution that will cluster news articles on specific events.

The clustering solution is chosen after comparing the algorithms and evaluating their performance. An experiment is conducted, and a few identified metrics are used to evaluate the performance of the algorithms. There is a myriad of evaluation metrics, they

cannot be used all because of time constraint. The metrics used for this experiment are discussed below.

2.12. Evaluation parameters

The parameters that were used to measure the performance of the algorithm are Rand Index, F1 measure, Precision and Recall. These are common metrics used to compare the performance of ML algorithms. The metrics can be used for supervised learning, clustering and information retrieval tasks (Hanczar & Nadif, 2019).

Moerchen et al. (2007) did similar work, they developed a system for clustering high streams of news articles. The system analyses high streams of textual data and cluster similar articles into one cluster. They compare the performance of K-Means and other non-K-Means algorithms. The results are evaluated using F1 measure metric.

The confusion matrix below will help to understand how the different equations for the evaluation metrics were derived.

Table 2.1: Confusion matrix

	P (Predicted)	**N (Predicted)**
P (Actual)	True Positive	False Negative
N (Actual)	False Positive	True Negative

The confusion matrix in Table 2.1 was adapted from Sokolova & Lapalme (2009).

Cluster evaluation is made in reference to the manual cluster and the algorithm generated clustering results.

True positive (TP) is when a news article is correctly clustered into a correct cluster.
False positive (FP) is when a news article that does not belong to a cluster is put in the cluster.
False negative (FN) is when a news article is clustered elsewhere and not in the cluster to which it belongs to.
True negative (TN) is when a news article that does not belong to a cluster group, is assigned to the cluster (Kumar & Rathee, 2011).

2.12.1. Rand Index

Rand index (RI) measures pairwise similarity between the manual cluster and the clustering generated by the algorithm (Handl et al., 2003). Rand Index is a value between 0 and 1. A value of 1 denotes a perfect similarity (Rokach & Maimon, 2010). RI is computed as

$$\mathrm{RI} = \frac{TP+TN}{TP+FP+FN+TN}$$

Equation 2.1

2.12.2. Precision

Precision looks at the proportion of news articles that were assigned to one cluster, how many belong to that cluster. Precision is calculated according to the formula below.

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

Equation 2.2

2.12.3. Recall

Recall considers the proportion of the news articles that are known to belong to a certain cluster, how many have been correctly put in the cluster they belong to. Recall uses the formula below.

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

Equation 2.3

The Precision and Recall values range between 0 and 1. In an ideal situation it is desirable to have a score value of 1 for Precision and Recall.

2.12.4. F1 measure

F1 measure is a single figure that shows the effectiveness of an algorithm. The measure is computed from Precision and Recall. It is a harmonic mean of the two metrics (Forman, 2003). The Precision and Recall values can be weighted differently depending on what you want to measure. A 50% weight which is also called beta (β), means Precision and Recall are weighted equally. A higher beta (β) favours Precision.

$$F_\beta = \frac{1}{\beta \times \frac{1}{Precision} + (1-\beta) \times \frac{1}{Recall}}$$

Equation 2.4

2.13. Related work

In Ding & He's (2004) work in "Principal Component Analysis and Effective K-means Clustering" their approach applies K-means clustering on internet news articles. They use term frequency inverse document frequency for their story retrieval and weighting method. They explore the relationship between PCA and K-Means. This research is similar to the work above and will use the same approach taken by Ding & He. This research is different in that it compares performance of only K-Means based algorithms on clustering of news articles. While the work above was comparing K-Means and non-K-Means based algorithms.

Azzopardi and Staff (2012) present a design and evaluation of Incremental Clustering of News Reports. This is a system that reads news reports from RSS feeds. It clusters them as they come. The clustering is event specific. The news reports are presented using Bag of Words (BOW) and Term Frequency Index Document Frequency (TF.IDF) and uses a variation of K-Means that cluster in a single pass without need for cluster reorganisation. The system does not know the number of clusters before beginning. They conclude that the system is effective on clustering event specific news but performs poorly when doing general clustering.

Atefeh & Khreich (2015) provides a survey of techniques for event detection in twitter streams. The event detection in twitter is a challenging task. The challenge is that tweets are limited in length and are written by diverse people using informal language. These challenges have a negative effect on the performance of event detection algorithms. They note the potential that twitter has as a fast-growing microblog that can be used to extract user generated knowledge on real time world events. They mention the challenge of unavailability of testbeds for performance evaluation and comparison of different approaches.

In their paper, Sankaranarayanan et al. (2009) presented an algorithm using Naive Bayes classifier. The application called TwitterStand is used to extract the breaking news from twitter posts.

Phuvipadawat & Murata (2010) developed Hotstream, which is a web application that can track breaking news in twitter. The stories are collected by a streaming API, using predefined queries such as hash tags. Apache Lucene indexing is then used to group similar stories together. Similarity comparison of stories is done using TFIDF. Merge

Threshold technique is used to ensure that stories being assigned to a group are related to the story or stories already inside the group.

2.14. Other clustering techniques

2.14.1.1. Mean Shift

Mean-Shift clustering algorithm is referred to as a sliding window-based technique. It looks for dense area in data points. The difference between Mean-Shift and K-Means, is that Mean-Shift does not require user to give the K value or number of clusters, it can discover it automatically (Konstantinos, 2005). Its desirable attribute is cluster centres converge to the points of maximum density. The disadvantage is window size selection is sometimes non- trivial.

2.14.1.2. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is like mean shift because it is density based. DBSCAN's disadvantage is poor performance with clusters of varying density. The reason is threshold distance and min points that identify neighbourhood points vary with each cluster (Wang et al., 2015).

2.14.1.3. Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)

EMGMM is more flexible than K-Means. Gaussian Mixture Models assumes data points are gaussian distributed. This removes the assumption that data is circular when using the mean. Data is described using mean and standard deviation, clusters can take elliptical shape. GMMs are flexible on cluster variance compared to K-means. They can also assign a data point to multiple clusters (Ari & Aksoy, 2010).

2.15. Systematic literature review

A literature review is compiled to evaluate and organise literature on a subject. It serves to access the knowledge pool available on the topic. After completing a literature review you get a broader understanding of the subject area. Literature review can help you to develop the conceptual framework. Wide readership may help you to find similar research that has been conducted. Reading similar research helps you to know methods that have been used before.

2.16. Research questions

RQ1: What are the existing solutions for clustering online news articles using different algorithms?

RQ2: How does the different solutions for clustering online news articles compare to each other with respect to specific constraints, methods or approaches?

RQ3: What is the strength of the evidence in support of the different solutions?

RQ4: What implications will these findings have when creating an online news article clustering system?

2.17. Review Protocol

2.17.1.1. Databases

The systematic literature review was done by collecting papers from the following online Databases: ACM Digital Library, IEEE Xplore, Science Direct, Elsevier and google scholar.

2.17.1.2. Search Terms

Unsupervised AND (K-*) AND (online OR Web) AND (articles OR news OR content) AND (clustering OR grouping OR classification).

2.17.1.3. Inclusion and Exclusion Policy

The following inclusion and exclusion criteria were used to limit our search results:

- Inclusion Criteria: Primary studies
- Exclusion Criteria: Secondary studies (Reviews); Studies before 2000

2.17.1.4. Quality Assessment

This paper answers the systematic review questions by providing the evidence from the carefully selected literature after the screening process. Quality assessment questions are adopted from Malhotra (2015).

2.18. Results

Table 2.2 lists the results of studies obtained from the review protocol and a carefully selected search term, and a consideration for the inclusion and exclusion policy. The articles obtained answer literature review research questions.

Table 2.2:Results of literature review

Study	Reference	Type	Method(s)	Conclusion
S1	Salloum et al (2017)	Text Mining from social platform.	K-Means, Text Parsing Node	They used K-Means with different k value, k=4 was the reasonable value.
S2	Vishwakarma et al (2017)	Social media text mining.	K-means, K-medoid	K-Medoid perform better than K-Means on time and space complexity.
S3	Lo et al (2017)	Multilingual social media topic identification.	Peak Identification algorithm, TF- TFDF Clustering techniques: Means, Dirichlet Process Mixture Model, Latent Dirichlet Allocation.	Peak Identification found more relevant terms compared to TFIDF and TF. TFIDF found more hashtags, TF identified more generic terms. Selecting best performing algorithm is difficult each algorithm performs differently on different candidate dates.
S4	Li et al (2016)	Short text clustering from micro-blogs	Biterm Topic Model (BTM), Hierarchical Clustering (HC), BTM and K-means	K-Means performs better than HC.
S5	Hu et al (2017)	Event detection to discover news documents that report on the same event.	Word embeddings, then cluster words semantic classes via K-means algorithm. Adaptive clustering algorithm.	Adaptive online clustering method for online news event detection has improved precision and recall performance using time slicing and merging over traditional clustering algorithms.
S6	Makkonen et al (2004)	Spotting something previously unreported, tracing even development, grouping news on same event.	TDT approach using semantic classes, TFIDF, Connexor Functional Dependency Grammar parser for English (EN-FDG), Connexor's Term Extractor (ENBRACKETS)	
S7	Moerchen et al (2007)	Clustering high frequency news streams.	Geospace and media tool, Locality sensitive hashing and TFIDF.	Clustering documents with limited memory and processing time.

RQ1: There are several solutions for clustering online content. The studies listed in the table above present the solutions. S1, S2, S3, S4, S5, S6 and S7. S1 and S2 look at text mining on social media. S3 looks at multilingual social media topic identification. S4 addresses microblog short text clustering. S5 is about event detection to discover news documents that report on the same event. S6 is aimed at spotting something previously

unreported, tracing the development of an event, and grouping together news that discuss the same event. S7 looks at clustering high frequency news streams.

RQ2: The different clustering solutions work with one or two feature selection techniques. S1 uses Text Parsing Node, S3 uses Peak Identification Algorithm, Term Frequency and Term Frequency Document Frequency; S5 uses Word embeddings; S6 uses Term Frequency Document Frequency, Connexor Functional Dependency Grammar parser for English (EN-FDG), Connexor's Term Extractor (ENBRACKETS); and S7 uses Locality sensitive hashing and Term Frequency Document Frequency. Of all these studies the most used term feature is Term Frequency Document Frequency and the most used algorithm is K-Means.

RQ3: Of the seven studies listed above five of them use one or two of K-Means or variant of K-Means algorithms. This goes to show the popularity of K-Means algorithm.

RQ4: Most solutions are created using Term Frequency Document Frequency and K-Means techniques as evidenced by the studies listed in this review. Therefore, it makes sense to use them when creating an online news clustering system.

However, this research has found another model that is easy to implement and has readily available libraries to use for document vectorisation. The model is Doc2vec which is discussed in more detail in the subsequent sections of this thesis.

2.19. Discussion

In S1 they used K-Means with a different K value, k=4 was the reasonable value. On S2 they compared K-Means to K-Medoid. K-Medoid perform better than K-Means on time and space complexity. On S3 they compare K- Means, Dirichlet Process Mixture Model and Latent Dirichlet Allocation. The results are not conclusive as each algorithm performed better on a different parameter. S5 used Adaptive Clustering algorithm and complemented it with K-Means algorithm, the algorithms were not compared but Adaptive Clustering algorithm was observed to perform better on precision and recall by using time slicing and merging over traditional algorithms.

2.20. Conclusion

From the literature obtained it is evident that K-Means is the most popular among other algorithms due to its performance. Research on comparing several K-means variants on the same set of parameters and constraints still needs to be done.

CHAPTER THREE

Methodology

3.1. Introduction

This chapter discusses the methodology and research process that will be used in this experimental research. The literature review in chapter two identified a research gap that this investigation seeks to address. The methodology to be utilised in addressing the research gap is experimental research. The explanation of the experimental methodology will cover the research process. This chapter also discusses the research motivation, literature review analysis, research questions, conceptual framework, data collection and data generation.

3.2. Research process

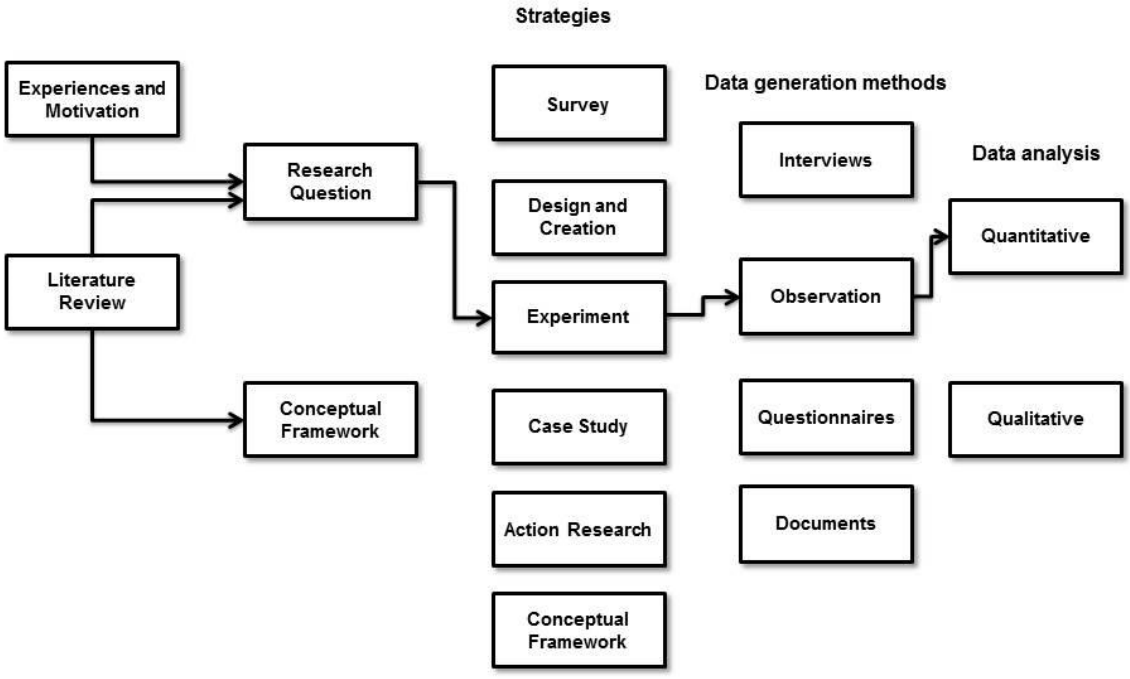

Figure 3.1: Research process

Figure 3.1 above shows the research strategy followed in the experimental research. The experience, motivation and literature review lead to the development of the research questions and the conceptual framework. The methodology used to answer the research questions is experimental research. A conceptual framework was used to design the

experimental setup. Data was gathered by observing the output of the experiment. The data is quantitative in nature.

3.3. Experiences and motivation

The motivation for this research was the need to cluster online news articles using ML algorithms, to manage duplication. Manual clustering of news articles using human effort is possible. However, the task is tedious and expensive for a human being. Human effort is prone to error and fatigue. Hence ML clustering is more desirable. Clustering is important because of overload of news articles. The news articles are polluted with noisy information. Clustering will reduce the noise and make provision for cleaner information.

Another motivation of this research is identifying an algorithm with a satisfactory performance. The algorithm should be able to cluster content that is similar. If there are three stories that are written by three different people, probably using different headings, but are discussing the same entities and time, they are similar. They should be put in the same cluster.

There are other solutions that have been developed to cluster news articles. These other solutions fall short when it comes to clustering stories that are the same. A story is considered the same if it discusses the same event and setting. The same story can be written by different people using different versions. Some solutions, for example recommender systems, will suggest stories based on titles of a story. Others will give results based on a search term or query.

3.4. Literature review analysis

Literature has it that K-Means is the most popular, efficient, studied and used clustering algorithm (Liberty et al., 2016). K-Means algorithm is popular because it is easy to understand and implement.

There are known weaknesses of K-Means. These weaknesses led to the development of many variants of the algorithm. In the process of improving the weaknesses new K-Means variants emerged. The many variants have opened a research gap of comparing the performance of these variants in order to identify the variant with acceptable performance. Other researchers have compared non-K-Means and K-Means algorithms. The problem now, is selecting a variant of K-Means to use from the horde of variants.

3.5. Research questions

Given a set of news articles and K-Means variations, how can we find the best variant with good performance to cluster news articles?

Formally: We are looking for a mapping to take a set of news articles (X), number of clusters (N), the K-Means variant algorithm (V) and the number of iterations (I), which can produce the best clustering results.

3.6. Research sub questions

Table 3.1: Research questions and sub-questions

Research Sub-Questions	Method(s)	Objectives
What K-Means algorithm variation can accurately cluster online content into semantic clusters?	Experiment	To evaluate the variation of K-means algorithm that can accurately cluster content into semantic clusters.
What is the effect of increasing the number of clusters on the accuracy of clustered content?	Experiment	To examine the effect of increasing the number of clusters on the accuracy of clustered content.

3.7. Research Design and Methodology

The aim of the research was to experiment and find the best K-Means variation, with a good performance that can be used to cluster online duplicate or similar content into themes or topic. The research is quantitative in nature. There are five variables that were observed in this research, which are size of the news set (X), K-Means algorithm variation (V), number of clusters (N), the number of iterations (I) and the clustering performance (e).

The X, V, N and I are independent variables and e is dependant variable. In order to ensure viability and reliability of the research the effects of the independent variables upon the dependent variable will be observed. This research methodology is closely aligned to the research approach used by (Easterbrook et al., 2008) where the effect of the independent variables shall be observed on the outcome of the number of dependent variables.

The performance was accessed using the following metrics, Rand Index, F1 measure, Precision and Recall. The metrics are discussed in section 3.8 under the evaluation parameters.

3.8. Conceptual framework

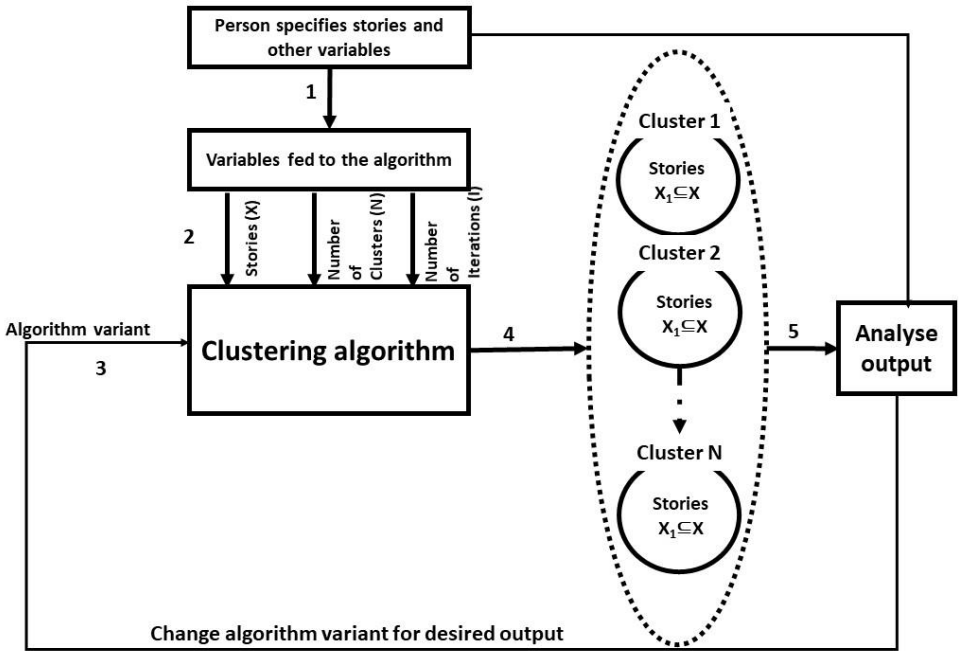

Figure 3.2: Conceptual framework

Figure 3.2 explains how the research was conducted. The experiment was implemented using Python programming language. In stage [1] the person does manual clustering of news articles and specifies the number of articles and clusters for the algorithm. In stage [2] the variables are put into the algorithm which then makes iterations and assign news articles into containers. The implemented algorithms took a set of news articles and number of clusters and a variation of K-means. The variables were run on all the variations and the performances were compared. The performance was obtained by comparing the result of the algorithm to manual analysis. The variables are then changed and run again. In stage [5] the person makes a manual comparison of the original clusters and the result of the algorithm.

3.9. Data collection

The data that was used in this experimental research was collected from websites in the form of stories. The stories are publicly available on the websites of news publishing houses. In that regard there was no need to request for permission to harvest the stories. However, care was taken to harvest the stories in an ethical manner. Ethical manner means that the collection of stories would not cause the websites from which the stories are collected, to crash or render them inaccessible. There would not be an injection of bots or programs that could harm the websites or servers that host them. The K-Means variants algorithms were obtained from literature.

The size of news article set X, which was obtained from online news sites, was manually clustered by the researcher. The manual clustering was done to validate the clustering performance.

The data collection was a two-prong approach. There was data that was fed into the system and data which was generated by the system itself. Algorithm variation V was selected from the literature, the number of clusters N and iterations I were carefully chosen during the experiment. The news articles X were obtained from the internet. The data generated from the system is discussed below.

3.10. Data generation

The experiment generated quantitative data. The data is the clustering output. The stories were allocated to the containers or labels. The labels are dependent on the defined K value, which is the number of clusters.

3.11. Data analysis

The stories obtained from different websites were manually clustered. Quantitative data was generated by the algorithm. The data generated by the algorithm and the manual clustering were evaluated to check the performance of the algorithms.

3.12. Experimental Research

3.12.1. Research definition

The research is an experimental type. The results obtained from the experiment are quantitative data. Research is a term that is made up by two syllables, “re” and “search” . Re is a prefix, denoting doing again, and search is a verb that describes a careful examination of a subject matter. These two syllables describe a process of establishing new knowledge by systematic and diligent inquiry (Mahmood, 2011). Research is defined as a systematic and diligent investigation of a subject matter to discover new facts or revise theories (Eneh, 2008). It is human nature to be inquisitive. When confronted by a phenomenon or the unknown, being inquisitive makes us quest for answers. The inquisitiveness is the route to seek knowledge, and the methods used to attain the knowledge is research. The goal of research is to report and communicate the newly discovered knowledge.

3.12.2. Experimental research

Experimental research has its roots in psychology and education. In the 19th century when psychology emerged as a discipline, its research methods were shaped around

the established method of physical sciences. Physical sciences relied on experimentation to establish principles and laws (Ross & Morrison, 2004).

3.12.3. Experimental research core

The research main question and sub questions are properly outlined in sections 3.5 and 3.6 respectively. The core of an experimental research is the research question. If the research question is not properly defined or operationalised, the experiment may lead to invalid results (Orero et al., 2018). Experimental research follows a strict design. The manipulation of variables in an experiment produces results that are used to validate the objective of the research (Harland, 2011).

Four characteristics exist in experimental research. These include control, manipulation, observation and replication. Variables that are not of direct interest need to be controlled. Controlling is basically minimizing the effect or influence of such variables by means of several methods. The methods include random assignment of subjects to groups, statistical techniques and standard deviation of groups. Manipulation is an operation on the independent variables. Manipulation influences the dependent variables. Observation is taking note of the resultant effect of the independent variables on the dependant variables. Replication basically means one can conduct subsequent experiments within the same experimental design. Several observations can be made on the experimental and control groups (Kirk, 2012).

Experimental research methods are skills that are needed to reduce errors, in the process of acquiring and communicating results. Communicating results is a very important aspect of experiments. The results can be communicated in the form of a report, journal or thesis that can be published. The way the results are communicated can also affect the way the experiment is conducted (Maxion, 2009). This calls for specification of a criteria that is relevant to the experiment and metrics that correspond with the measuring tools. If an experiment can be repeated, it means its results can be validated (Papadimitriou et al., 2012).

3.13. Experimental research quality

The quality of the research was guided by the instrument used to measure performance. The instrument and terminology thereof are discussed in chapter under Evaluation metrics. The quality of a research study is judged by considering threats to validity of a study and the results. A consensus by the research community has got to be reached on

how the reporting of validity will be done. A consensus must also be reached on the common terminology that must be used (Feldt & Magazinius, 2010).

The two forms of validity are internal and external validity (Jiménez-Buedo & Miller, 2010). Internal validity exists when the results obtained are a direct manipulation of the independent variable (Aziz, 2017). External validity means that your results can be generalized. It means your results can be applied to similar situations (Altermatt, 2009).

3.14. Conclusion

The chapter has discussed the research process. It discussed the motivation for undertaking the research, and literature review, and research questions. It discussed the research design, the research methodology, and the parameters that were used to evaluate performance. It outlined the conceptual framework and data collection, generation and analysis process. It concluded with definitions of research and experimental research.

CHAPTER FOUR

Experiment planning

4.1. Introduction

This chapter discusses the experimental goals. The experimental goals are the focus of this research, what the research is trying to answer or find out. It also discusses the participants that are involved in the research, the participants' roles including the researcher, computational hardware used, the clustering libraries and algorithms used are also described. It also discusses the tasks and procedures that were implemented to carry out the research. It discusses the deviation from the plan.

4.2. Experimental goals

The goals of this experimental research were:

- To compare the performance of several variants of K-Means algorithms. The idea is to use the same set of input data, the same set of constraints and the same k value. The input data is used on the K-Means variants and measure the performance. The variant with best clustering result will then be identified.

- To filter the vast amounts of information using clustering algorithms, information can be containerised into predefined containers. Such that information in the same container is the same. This will make it easy for users when they search for information to find it in one location.

- To cluster articles that are similar together. The title of the articles may be different, but if the theme of the story is the same in terms of the event being discussed, all the stories pertaining to the event should be able to be assigned to the same container by the algorithm.

4.3. Goal overview

The research questions below were the goal and focus of this experiment. They informed the experimental set up. The experiment was designed to answer these questions.

Main question: Given a set of news articles and K-means variations, how can we find the best variant with good performance to cluster news articles?

The main question is answered using experimental research methodology. The best algorithm can be found by setting up the experiment and feed the data to the algorithms then observe the results.

Formally: We are looking for a mapping to take a set of news articles (X), number of clusters (N), the K-Means variant algorithm (V) and the number of iterations (I), that can produce the best clustering performance. The clustering performance e was accessed using Rand Index, F1 measure, Precision and Recall.

Research sub questions:

I. What K-Means algorithm variant can cluster online content into semantic clusters with good performance?

The research sub question was answered by obtaining the variants from literature and iterate through the different algorithms, then observing the results.

II. What is the effect of increasing the number of clusters on the accuracy of clustered content?

The research sub question was answered by iterating through the different algorithms and increasing the number of clusters, then observing the results.

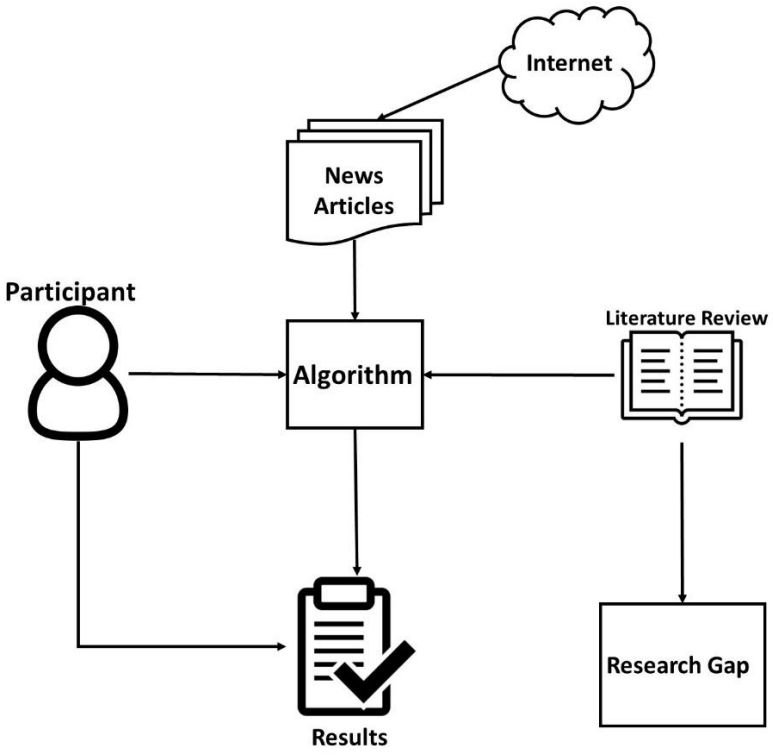

Figure 4.1: Schematic view

Figure 4.1 above shows the schematic view of this research. From the internet we get curated data in the form of news articles. The news articles are fed into the algorithm. The algorithm produces results of clustered news. The participant manipulates the variables of the algorithm and checks the accuracy of the results. The literature review identifies the research gap.

4.4. Participants

The experiment had no other participants besides the researcher. The researcher was the sole participant in this research. In that regard there was no need for ethical clearance or consideration before undertaking the research.

The participant interacted with websites on the internet from which the data was collected. The literature review component provided input in terms of algorithm variants to be used in the experiment. The Literature review also provided the research gap which was pursued. An algorithm manipulated the data and other variables. After the manipulation it produced output data. Figure 4.2 shows the participants of the research.

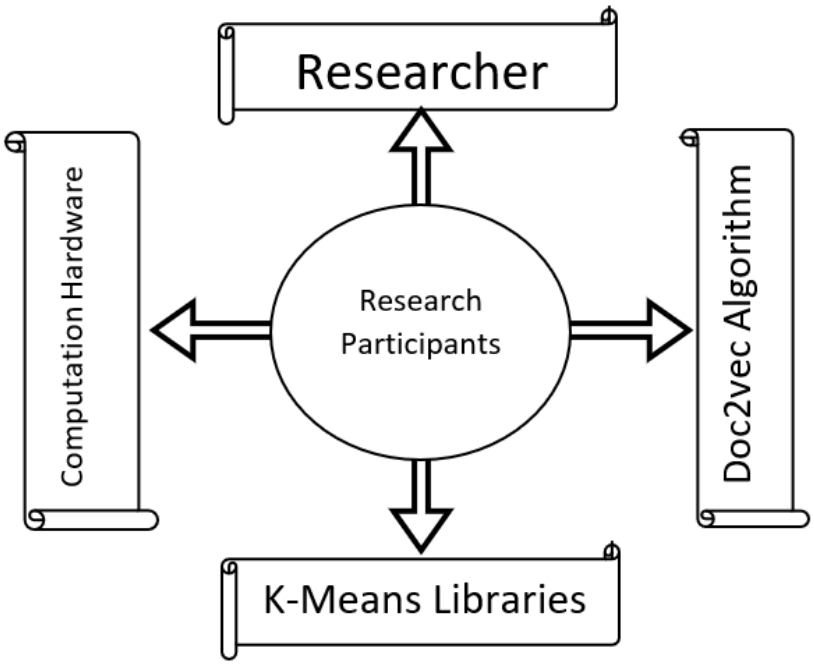

Figure 4.2: Interaction Between Participants and other Research Components

4.5. Research components

The researcher carried out the experiment, using a laptop and open source software. The open source software used was Python. The laptop used to carry out the experiment was supplied courtesy of CPUT. The university also supplied a venue to conduct the research. The K-Means algorithms came from the literature and libraries were open source github projects obtained from the internet. The libraries used were done in Python. The Doc2vec algorithm was an open source Python library obtained from the internet. A gocrawler was used to harvest stories from several websites. The stories were stored in a database file. Python is the programming language that was used together with an IntelliJ IDEA 2019.1 x64 IDE.

The stories were collected from internet and manually clustered or assigned labels. The stories were then stored in a Cassandra database. The stories were converted to vector representation using Doc2vec algorithm. An algorithm was run on the story vectors (X), K-Means variant (V) and number of clusters (K). The results obtained are compared to the manual clustering. The results are analysed to determine the variant that produces the best results.

4.6. Procedure

A gocrawler collects articles from the internet. The stories came contaminated with adverts and pictures. They were cleansed so that only text or image of the article remains. The cleansing will make the clustering task easier.

The articles were stored in a database where they were retrieved for clustering. It is not possible to pass the stories directly into the algorithm from the sanitization process. Doc2vec had to convert the stories into a vector representation first. The vector representation of the stories was fed into the K* model together with other settings.

4.7. Deviation from plan

In any experiment or set up there is a risk of deviating from the set-out plan, because of unforeseen eventualities. A contingent plan should be put in place to accommodate such eventualities. There is a risk of using alternatives to the tasks and procedures explained earlier. An example is, instead of using gocrawler to collect stories, an alternative is simply copying a story from a website. The story can be copied and pasted on word or any other text capable application then be stored in a database. Another programming language other than Python can be used. A different database design and development program other than Cassandra can be used. The above-mentioned options would be a deviation from the plan.

This research will be carried out strictly using gocrawler, Python and a database as per experiment plan. Any deviation from plan will delay completion of the research. The researcher does not have skills for other languages, it would mean learning another language other the planned one and looking for algorithm libraries in other language, which may prolong the research.

4.8. Conclusion

Chapter four expanded on the methodology discussed in chapter three. Chapter four discussed the experimental goals. It talked about the participants that are involved in the research, the participants' roles. The participants are the researcher, computational hardware, the clustering libraries and algorithms used. It also talked about the tasks and procedures that were implemented to carry out the research. It discussed how the deviation from plan can be controlled.

CHAPTER FIVE

Experimental setup

5.1. Introduction

Chapter five discusses in detail how the experiment was carried out. It talks about how the news articles were converted into a suitable form for the algorithms to consume and the algorithms involved. It talked about the data, which is the news articles. It also talked about the parameters or settings for the algorithms. The chapter uses a diagram to explain the experimental setup. The diagram gives details of how the experiment was conducted. The diagram is a step by step outline of the process. The diagram was developed from the research motivation. The chapter answers the research questions. The results of the experiment were evaluated for performance.

5.2. Experimental outline diagram

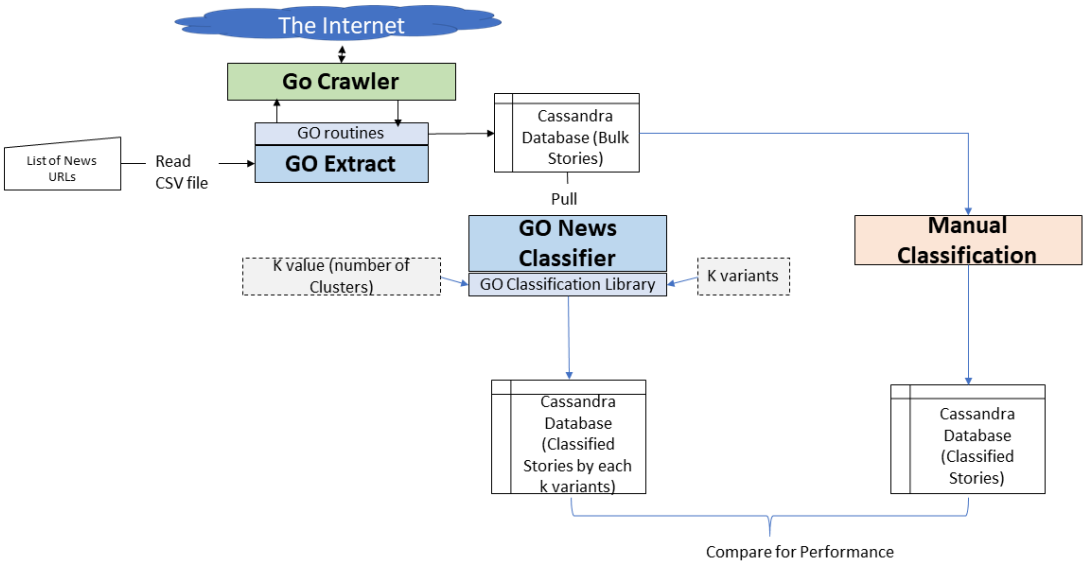

Figure 5.1: Experimental outline

Figure 5.1 above explains the experimental outline.

5.3. Experiment outline

The experiment was run on a windows machine, using Microsoft windows 10 pro. The processor used is an Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz, 1992 Mhz, 4 Core(s), 8 Logical Processor(s). The RAM size was 8GB.

The input to the experiment is stories that came from the internet. The stories were published between April 2015 and April 2019. A gocrawler was used to collect the stories from the internet. A gocrawler is a program that takes a Universal Resource Locator (URL) as an argument and collects a story from a website pointed to by the URL. The stories on the websites are contaminated with adverts and images. The gocrawler will remove the adverts and images and return a clean text.

The stories that are collected will be manually classified and labelled. The labelling is done (see Table 5.1) to check the accuracy of machine clustering. Each story is labelled with a class to which it belongs. Below is a table showing the stories and manual labels to which they belong.

Table 5.1: Shows labels and indexes of stories under the label

Story category	Story Label Number									
Accident	7	11								
Car Breaking	43									
Election	5	6	20							
Floods	31	32	33	34	35	40	44			
Inflation	36	37	38	39						
Land	1	10	16	25	26	27				
Murder	2	3	21	22	23	24				
Rape	8	9	12	13	14	15				
Terrorism	4	17	18	19	28	29	30	41	42	45

The gocrawler reads a list of URLs from the database. It outputs an array of articles. The output is stored in a Cassandra database. This is the raw data that was fed into the algorithm. The algorithms were varied on the same data set.

5.4. Parsing Documents to The Algorithm

The collection of the news articles was not a totally unsupervised process. The algorithms that were used to cluster the stories are the ones that are totally unsupervised. The database was manually created by copying and pasting the URL into the file. The columns used to save the links in the database are Number, URL, Cluster category and Cluster key. The Cluster category is used to check the accuracy of the algorithm.

The stories which were collected from the internet were stored as single documents in plain text. The documents cannot be parsed to the clustering algorithm in plain text. The document needs to be converted into a numerical value. Numerical value or document

vector is the form appropriate for the algorithm to process. There are many techniques available to make numerical representation of documents, but most of them do not offer good performance. The document vector representation technique for this experiment is discussed below.

5.5. Numerical Document Representation Techniques

5.5.1. Doc2vec

The research used Doc2vec algorithm to convert documents into a vector representation. Doc2vec is known for producing good results. It is a simple technique and very easy to use. Doc2vec evolved from word2vec. To understand how Doc2vec works it is important to discuss word2vec first.

5.5.2. Word2vec

Word2vec is a useful model that can transform words into vector representation. It captures the semantic relationship of words. The word and word's context relationship are modelled. The modelling will bring out relationships such as synonyms, analogies and antonyms of the word. A word vector can be several hundred dimensions. Every unique word in a corpus is assigned a vector in the space. The words are converted to vector representation so that algorithms can perform some operation on the vector, which is a numerical value, rather than on the text (Landthaler et al., 2017).

Word2vec is a combination of two algorithms. The algorithms are Continuous Bag-Of-Words (CBOW) and Skip-gram (SG).

5.5.3. Continuous Bag of Words

Continuous Bag of Words (CBW) is a very simple task. It is also a very common technique. The output of CBW is poor. This is because CBW has some disadvantages such as it does not make consideration for word ordering (Meyer, 2016).

5.5.4. Skip Gram

Skip Gram Model is a simple neural network model. The neural network is trained on a single hidden layer to perform a certain task. The training is usually done on a huge vocabulary corpus. The weights of the hidden layer are the vectors of the words. If the neural network is given an input word from a sentence, it should look at the nearby words and pick one at random. It then computes the probability of the random word being in the vocabulary. There are several steps involved in the algorithm. These include:

Data Preparation - Text data from internet is dirty. The process of cleaning text data, removing punctuations, stop words, replacing digits and converting text to lowercase. After cleaning the text, the corpus is then tokenized on white space. This will create a list of words.

Hyperparameters - Parameters used are window size, this is the number of words that are considered context neighbours of the target words. The window does slide along the sentences and each word becomes the target word.

n is the size of the hidden layer. It is the size of the word embedding. A good n value normally has the range of 100 to 300 dimensions.

Epochs is the number of iterations. Each iteration goes through the entire training set. Learning rate controls the amount of adjustment made to the weights with respect to the loss gradient.

Generate training data - at this juncture the corpus is turned into a one-hot encoding representation for the word2vec to train on. To generate the one-hot training data, word2vec() object is initialised first then using the object w2v call the function generate_training_data and pass settings and corpus as arguments. Generate_training_data performed the following sub functions:
v_count—Length of vocabulary that is the number of unique words in the corpus.
words_list—List of words in vocabulary
word_index—Dictionary with each key as word in vocabulary and value as index
index_word—Dictionary with each key as index and value as word in vocabulary
for loop to append one-hot representation for each target and its context words to training_data using word2onehot function.

Model training - training_data function will train the model. We run the function w2v.train(training_data) and pass in the training data which then calls the function train. The word2vec model is made up of 2 matrices w1 and w2. For demonstration purposes w1 is initialised to 9x10 matrice and w2 initialised to 10x9. This allows the back-propagation error to be calculated. In the actual training the weights should be randomly initialized with function np.random.uniform().

Forward pass - The training of the first epoch is done using first training example by passing in w_t which represents one-hot vector for target word to the function

forward_pass. In the forward_pass function produces h by doing a dot product between w1 and w_t. Another dot product is done between w2 and h to which produces output layer u. A run u through softmax to force each element to the range of 0 and 1 to give us the probabilities for prediction before returning the vector for predictiony_pred, hidden layer h and output layer u.

Error - The error for a certain set of target and context words can be calculated using y_pred, h and u. The error is calculated by the sum difference between y_pred and each of the context words inw_c.

Backpropagation - backprop function is used to calculate the amount of adjustment needed to alter the weights. This done by giving arguments error EI, hidden layer h and vector for target word w_t. The weights are updated by multiplying, weights to be adjusted (dl_dw1 and dl_dw2) with learning rate and then subtracting it from the current weights (w1 and w2).

Loss - the loss function is used to calculate the total loss. It is divided into two parts as: One part takes the negative of the sum for all the elements in the output layer. Another part takes the number of the context words and multiplies the log of sum for all elements, after exponential, in the output layer.

Inference - from the trained weights, search the word vector for a word in the vocabulary. This can be achieved by searching the index of the word against a trained weight(w1). To find similar words, implement function vec_sim compute the cosine similarity between words.

Further improvements - The backpropagation step requires the adjustment of the weights for the words that were not in the training sample. The process can long for a large vocabulary. To speed up the process you can implement Skip-gram Negative Sampling (SGNS) to improve the training speed and quality of the output word vectors. This is done by adjusting the training to modify a smaller percentage of the weights and not all of them.

5.5.5. Overview of Doc2vec

Doc2vec was developed from word2vec. It has two models embedded in it. The models are Paragraph Vector Distributed Memory (PV-DM) and Paragraph Vector Distributed

Bag of Words (PV-DBOW). They work almost similarly to the Skip-gram and CBOW models.

5.5.6. Paragraph Vector Distributed Memory

PV-DM is a technique for generating vector of words for a document. The technique was developed from Continuous Bag of Words model. PV-DM uses a target word to predict a context. A sliding window is used to create a vector of the whole paragraph. A SoftMax is used to predict context for all the words in the sentence as the window slides to create word embedding. The embeddings are averaged or concatenated. A vector is created for each paragraph and another vector is created for each word. The vectors are then concatenated or averaged. The paragraph and word vectors are trained using stochastic gradient descent. Backpropagation is used to obtain the gradient. A random fixed length context is used to calculate the gradient error which is used to update the parameters in the model (Le & Mikolov, 2014).

5.5.7. Paragraph Vector Distributed Bag of Words

PV-DBOW tries to use a word to predict a context. The method of updating the parameters is the same as PV-DM. PV-DBOW works in the same way as Skip-Gram. Doc2vec generates two vectors. One produced by PV-DM and another by PV-DBOW (Le & Mikolov, 2014). In this experiment one could make a choice on which vector model technique to use.

5.6. The clustering models

There are two methods for clustering that are being used to analyse the data. The methods are K-Means and K-Modes. Though there are several variants of K-Means, the experiment can-not explore more, because of time constraint. The experiment can be replicated on a larger scale later. The documents that were collected from the internet, were converted to numerical representation using Doc2vec. Word and document vector training run parallel, while word vectors are being trained, the document vector is also trained. The output will be the numerical value of a document. The numerical value is the one that is then passed to the clustering algorithm.

5.7. Performance

The two algorithms used the same data set and run on the same settings. Settings mean same number of iterations, word size, cluster size etc. The output of the algorithms is cluster labels. Each news article is assigned to a cluster label. The two algorithms' performance is measured in terms of run time. The algorithms' output is checked against

the manual cluster to evaluate performance. The performance is basically to compare the input data against the output data. The metrics used to access performance are Rand Index, F1 measure, Precision and Recall.

5.8. Data

Input data are news articles that will be consumed by the algorithm. The amount of input data collected is forty-five stories. More data can be used when the experiment is run on a larger scale. They are manually clustered into groups, the groups are further clustered into particular events by use of keys. The stories are indexed 1 to 45 in the database file. For example, there are stories that discuss terrorism. They marked with key “T”.

Output data is generated by the algorithm. The output data are the labels, the stories pertaining to the label and the run time. Performance will be measured by comparing the input and output data. The stories discussing a particular event should be grouped together.

5.9. Parameters

The parameters are the settings that are adjusted on the algorithms. They were used to train the Doc2vec model. The Doc2vec model is the one that converted the documents into vectors. Once the document was converted into a vector it was then passed to a clustering model. The parameters are defined as below.

1) w_size – is dimensionality of the feature vectors.
2) w_window - The maximum distance between the current and predicted word within a sentence.
3) w_min_count – ignores words with frequency less than specified range.
4) w_workers - is the threads to train the model, training will run faster on a multicore machine.
5) w_dm - Defines the training algorithm. If dm=1, ‘distributed memory’ (PV-DM) is used. Otherwise, distributed bag of words (PV-DBOW) is employed.
6) w_alpha - is the initial learning rate.
7) w_min_alpha – learning rate will decrease to w_min_alpha as the training progresses.
8) w_epochs – is the number of iterations per cycle.
9) w_start_alpha - Initial learning rate, if given it replaces starting alpha from constructor, for one call to train. This can be used when making multiple calls to train, it is not recommended to manage alpha learning-rate yourself.

10) w_end_alpha - Final learning rate, this has a linear drop from w_start_alpha. If it is given it replaces the final w_min_alpha from constructor, for one call to train (). It can be used if you are making multiple calls to train (), it is not recommended to manage alpha learning-rate yourself.
11) nclusters - is number of clusters to form as well as the number of centroids to generate.
12) n_init - is the number of times the algorithm will be run, with the best output selected from those independent runs.
13) init - is the method used for initialisation.

5.10. Cluster algorithm

Figure 5.2: Clustering process stages

Figure 5.2 above shows the clustering process.

Stage 1)

Input: Target (table, column)

Output: List of stories (Plain text)

Procedure:

Def read news articles

Return stories

Stage 2)

Input: Corpus C, Settings S

Output: Vector of Words V

Procedures:

Generate_Training_Data(C,S)

v_count(C)

return the number of unique words in C

words_list(C)

list each v_count(C) number of words in C

word_index(words_list(C))

return a dictionary i as the index of the word of each word in words_list(C)

index_ word(words_list(C))

return a dictionary word as value to index of words in words_list(C)

Stage 3)

Input: (story vector, story id)

Output: story id cluster label

Procedure:

Pick random points as cluster centres called centroids.

Assign each story to nearest cluster by calculating its distance to each centroid.

Find new cluster centre by taking the average of the assigned points.

Repeat Step II & III until none of the cluster assignments change.

Return story id cluster label.

5.11. Conclusion

Chapter five discussed in detail how the experiment was conducted. It discussed how the news articles were converted into a form suitable, for the algorithms to consume. It also discussed the algorithms involved. It discussed the data, which is the news articles that came from the internet. It also discussed the parameters or settings for the algorithms. The chapter uses a diagram to explain the experimental setup. The diagram gives details of how the experiment was run. The diagram is a step by step outline of the process. The diagram was developed from the research motivation.

CHAPTER SIX

Findings and discussions

6.1. Introduction

This chapter discusses the results and findings of the experiment. The results are a comparison of K-Means and K-Modes algorithms. They are presented in the form of graphs. They show the different scores obtained for the different evaluation metrics.

6.2. Experiment results

6.3. Doc2vec Run time in seconds

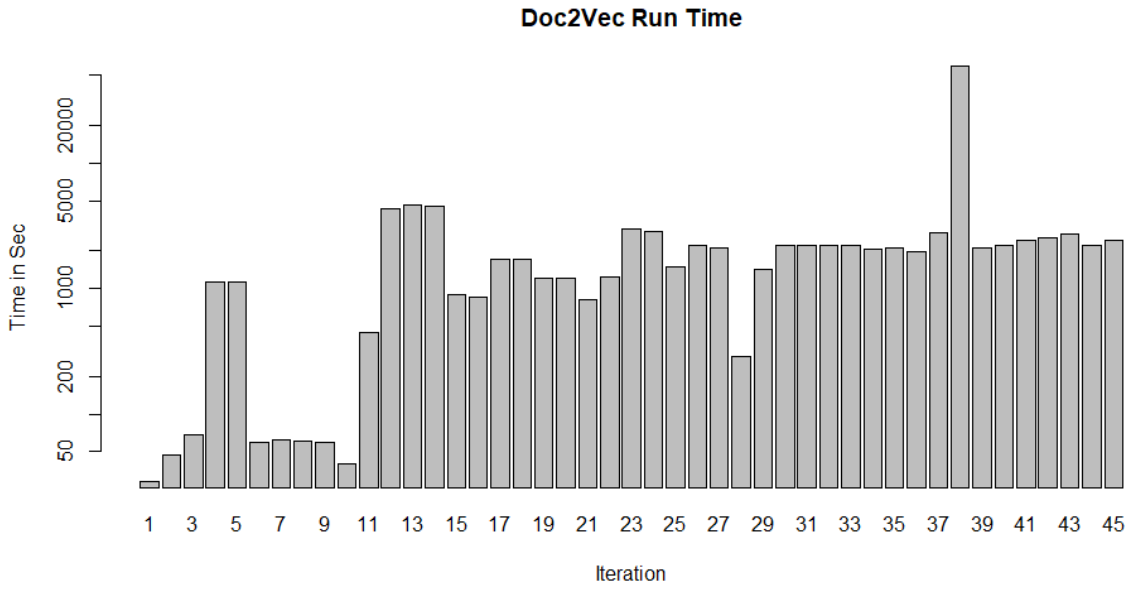

Figure 6.1:Graph shows run time of Doc2vec

The graph in Figure 6.1 shows Doc2vec run time in seconds. The duration taken in seconds on the Y axis against the different settings on the X axis. The graph shows the time it took the Doc2vec algorithm to convert the news articles into vectors. The highest time taken was on point 38 with a time run of 58832.1582 seconds. The lowest run took 28.96024585. The average run time was 2947.31316 seconds. The settings for the maximum and minimum iterations are shown below.

6.3.1. Doc2vec settings

Table 6.1: Doc2vec settings

w_size	w_window	w_min_count	w_workers	w_dm	w_alpha	w_min_alpha	w_epochs	w_start_alpha	w_end_alpha	nclusters	n_init	init	
150	3	7	20	1	0.005	0.001	50000	0.001	-0.006	10	20000	10	**Maximum**
100	5	500	8	1	0.025	0.001	1000	0.001	-0.006	8	1000	10	**Minimum**

Table 6.1 shows Doc2vec settings or parameters that obtained the maximum and minimum time to convert the news articles into vector representations. The parameters were explained in Chapter 5 section 5.9 Parameters.

6.3.2. K-Means model run time in seconds

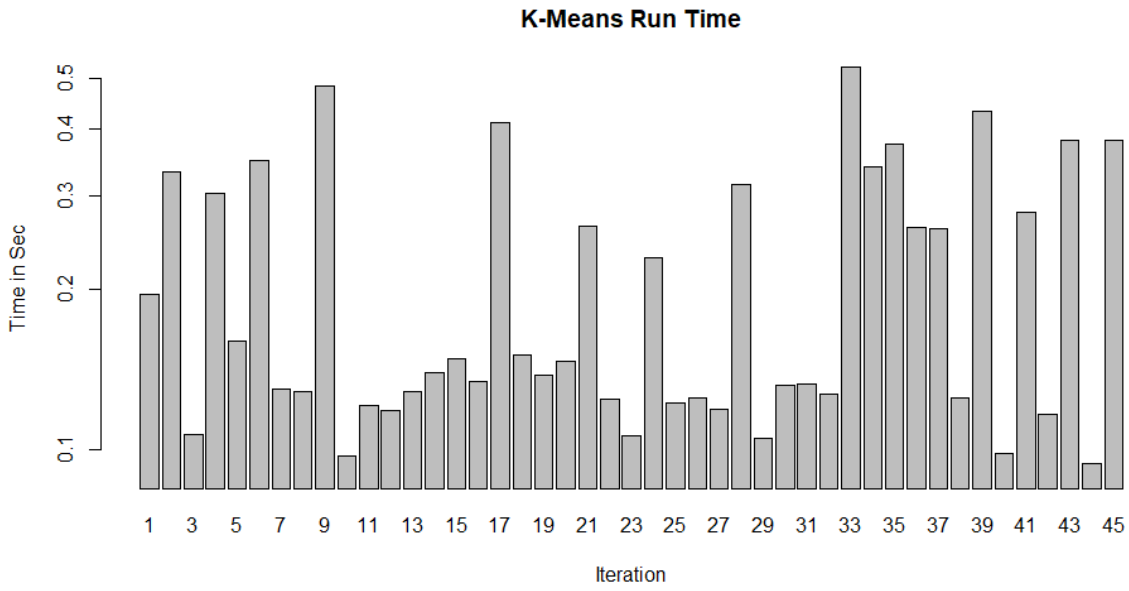

Figure 6.2: Graph shows K-Means run time

Figure 6.2 shows the runs made by the K-Means model. After the Doc2vec model has generated vectors, the output was then sent to the K-Means model to produce clusters. The runs show how long it took for the K-Means model to produce clusters for the different settings. The longest run was for run 33, took 0.523695946 seconds for k=8, N=45 and 20000 iterations. The shortest run was for run 44 for K=10, N=45 and 20000 iterations.

6.3.3. Combined Doc2vec and K-Means run time in seconds

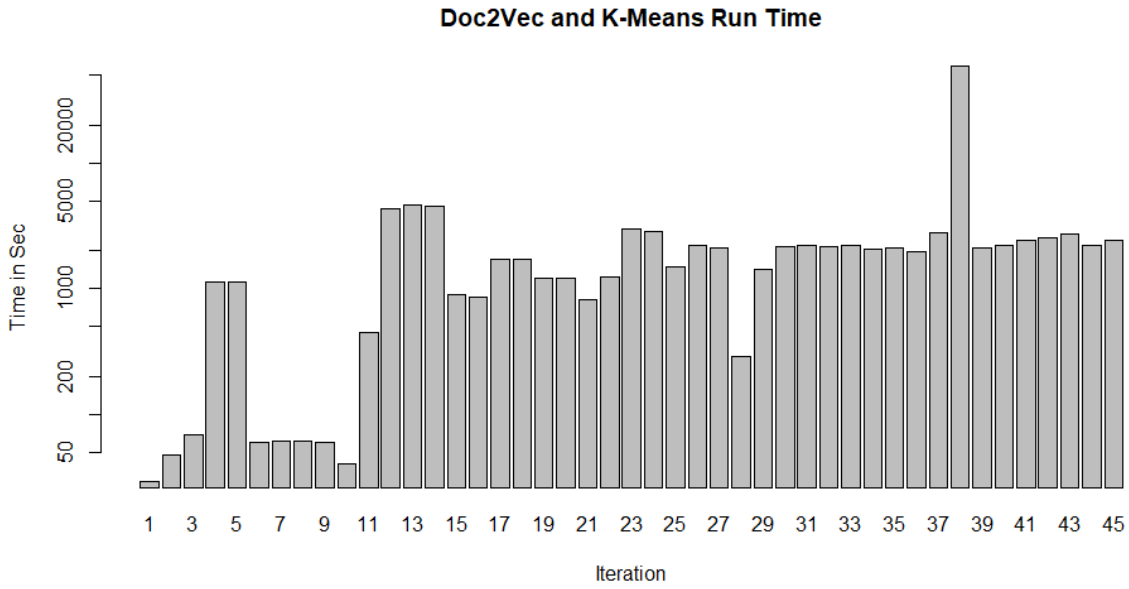

Figure 6.3: Graph shows Doc2vec and K-Means combined run time

Figure 6.3 is a graph with combined run time for the two models. Doc2vec converted news articles into vectors, then the vectors are passed to K-Means to do the clustering. The graph illustrates the run time for converting the news articles into vectors and the clustering model, combined.

6.3.4. K-Modes model run time in seconds

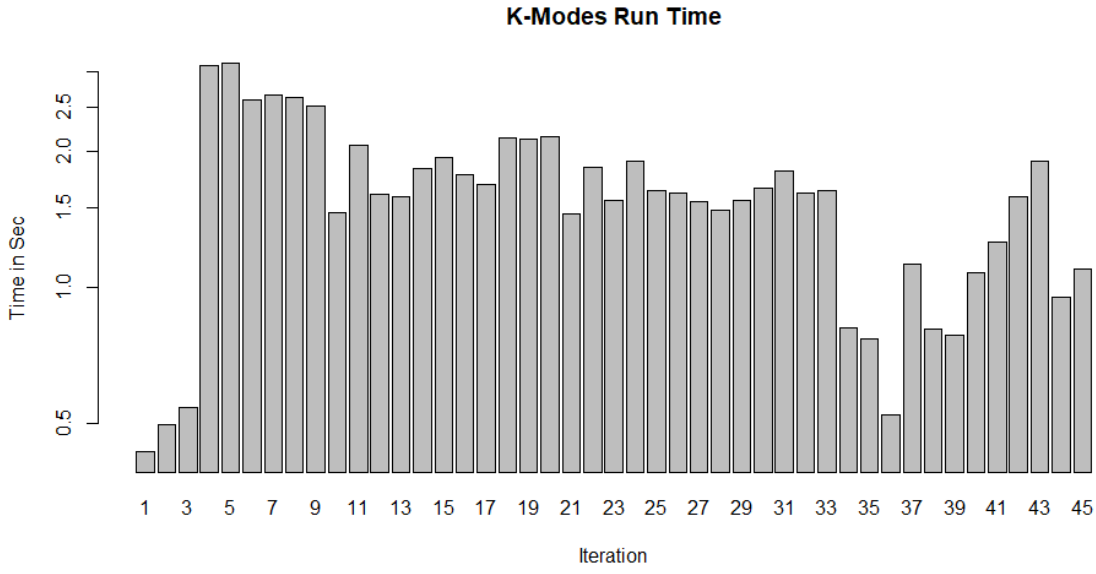

Figure 6.4: Shows K-Modes run time

The graph above in Figure 6.4 shows the time taken by the K-Modes to produce clusters. K-Modes took the output vectors from the Doc2vec model and produced clusters. The longest run was for run 5 which took 3.120784998 seconds for K=10, N=45 and 1000 iterations. The shortest run was for run 1which took 0.43283534 for K=8, N=45 and 1000 iterations. The average run was 1.606172689 seconds.

6.3.5. Combined Doc2vec and K-Modes run time in seconds

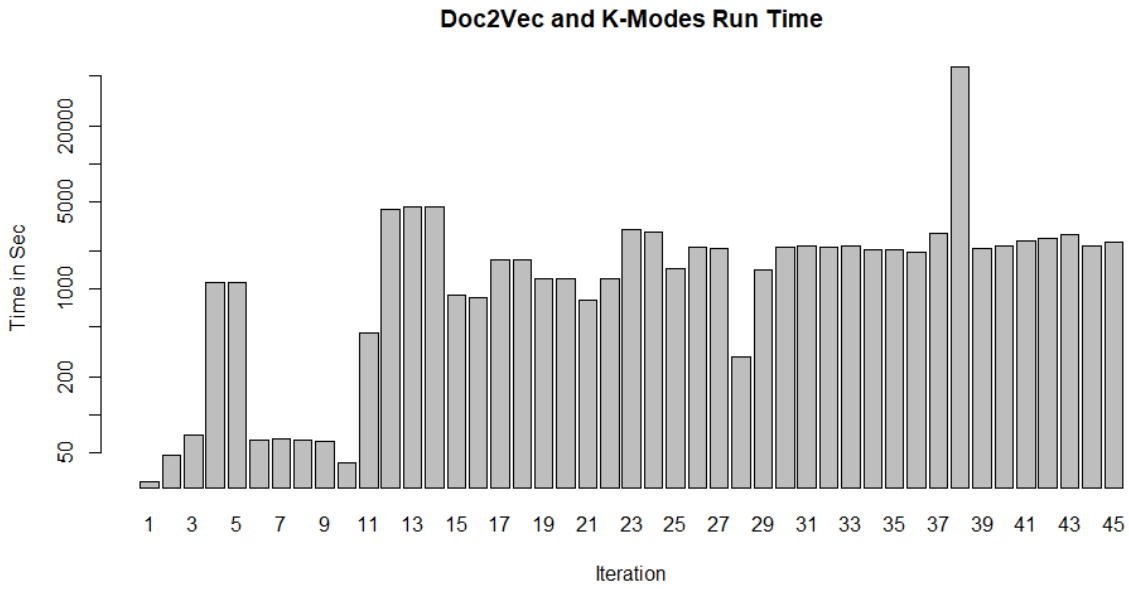

Figure 6.5: Shows Doc2vec and K-Modes combined run time

Figure 6.5 is a graph with combined run time for the two models. Doc2vec converted news articles into vectors, then the vectors are passed to K-Modes to do the clustering. The graph illustrates the run time for converting the news articles into vectors and the clustering model, combined.

6.3.6. K-Means results for all metrics

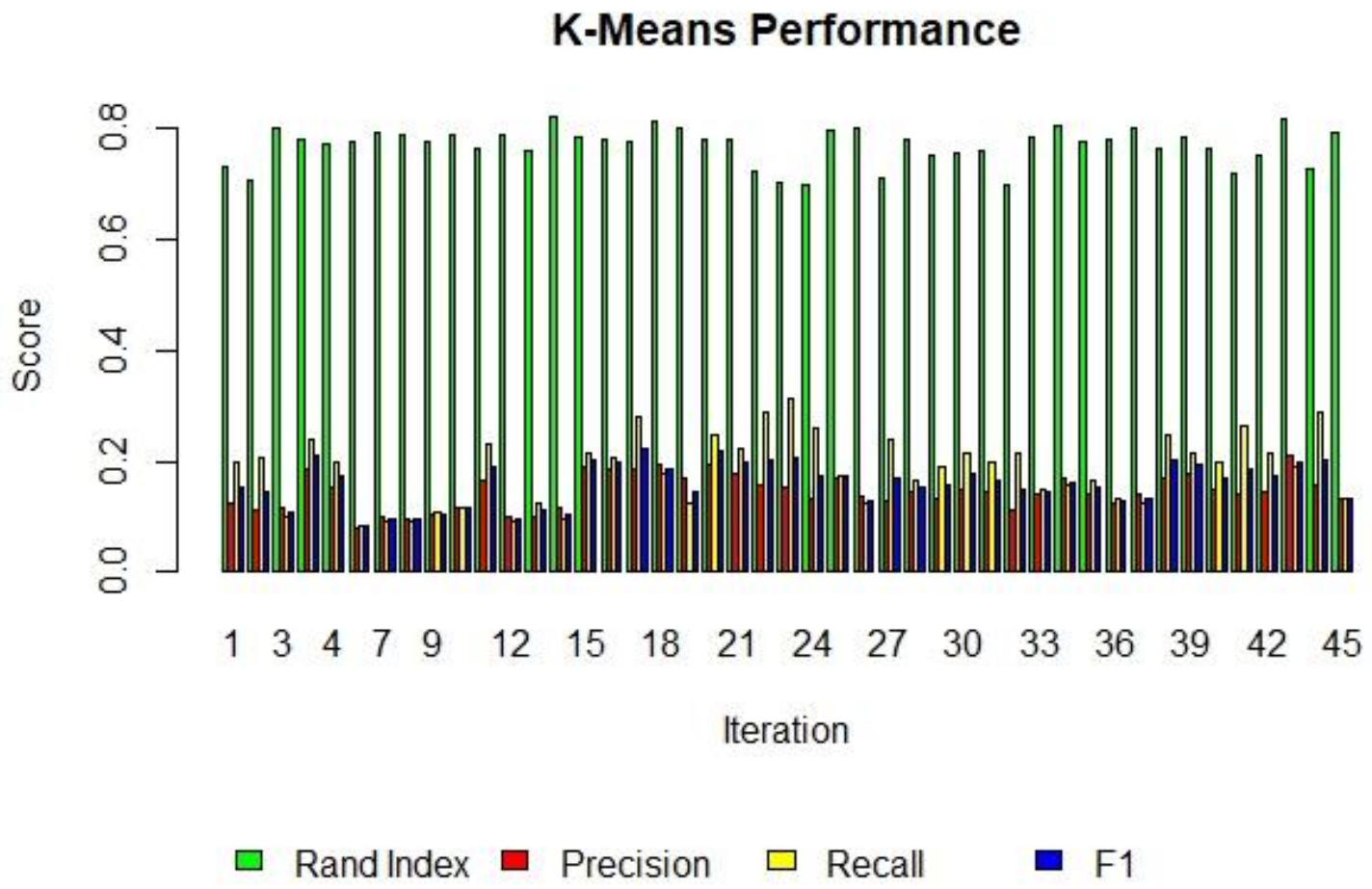

Figure 6.6: K-Means results

Figure 6.6 shows the highest and shortest values obtained for all the metrics on K-Means.

- **K-Means Rand Index**
    - Run 14 had the highest Rand Index measure with a value of 0.818318318. The settings were K=12, N=45 and 10000 iterations.
    - Run 24 had the shortest Rand Index measure with a value of 0.695459579. The settings were K=12, N=45 and 5000 iterations.

- **K-Means Precision**
    - Run 43 had the highest Precision of value 0.209090909. The settings were K=10, N=45 and 20000 iterations.
    - Run 6 had lowest Precision of value 0.080645161. The settings were K=10, N=45 and 1000 iterations.

- **K-Means Recall**
    - Run 23 had the highest Recall of value 0.314049587. The settings were K=10, N=45 and 5000 iterations.
    - Run 6 had the lowest Recall of value 0.082644628. The settings were K=10, N=45 and 1000 iterations.

- **K-Means F1 Measure**
    - Run 17 had the highest F1 Measure of value 0.222222222. The settings were K=12, N=45 and 5000 iterations.
    - Run 6 had the lowest F1 Measure of value 0.081632653. The settings were K=10, N=45 and 1000 iterations.

6.3.7. K-Modes results for all metrics

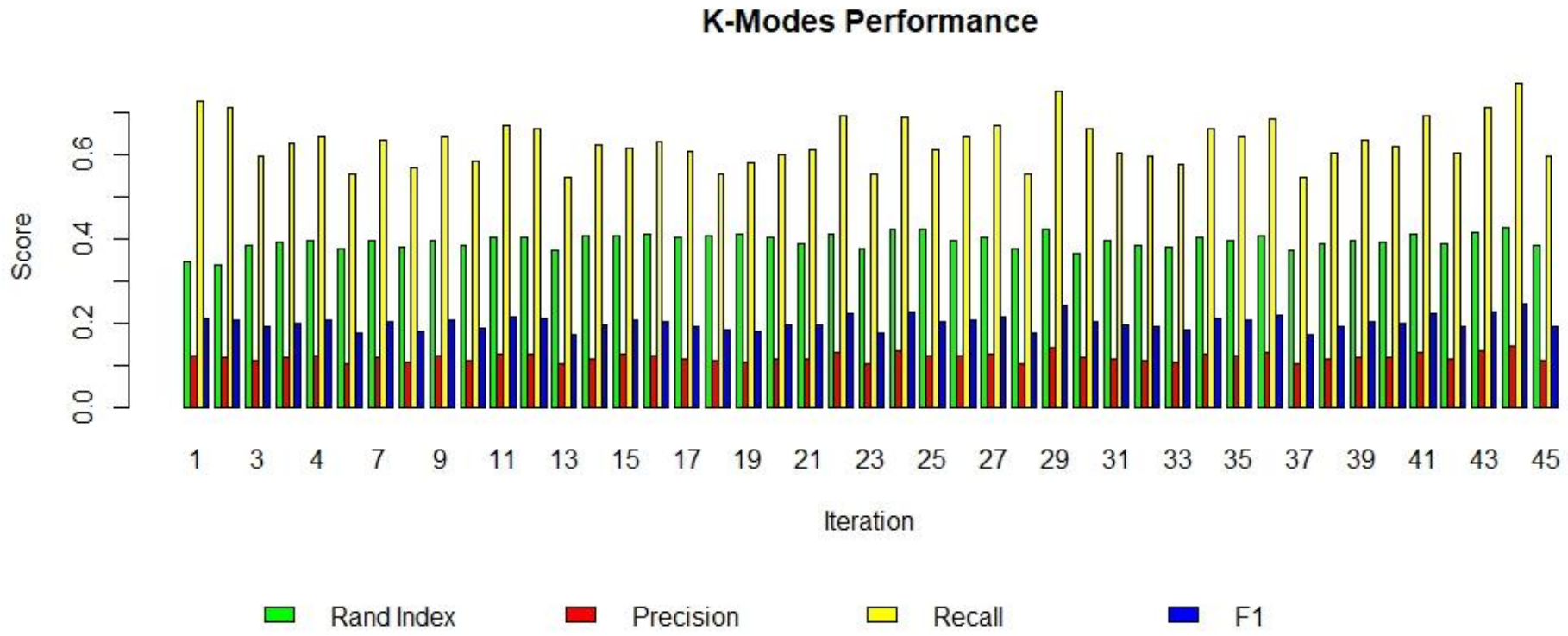

Figure 6.7: K-Modes results

Figure 6.7 shows the highest and shortest values obtained for all the metrics on K-Modes.

- **K-Modes Rand Index**
    - Run 44 had the highest Rand Index measure with a value of 0.429292929. The settings were K=10, N=45 and 20000 iterations.
    - Run 2 had the shortest Rand Index measure with a value of 0.341414141. The settings were K=8, N=45 and 1000 iterations.

- **K-Modes Precision**
    - Run 44 had the highest Precision of value 0.147619048. The settings were K=10, N=45 and 20000 iterations.
    - Run 13 had lowest Precision of value 0.104761905. The settings were K=10, N=45 and 10000 iterations.

- **K-Modes Recall**
    - Run 44 had the highest Recall of value 0.768595041. The settings were K=10, N=45 and 20000 iterations.
    - Run 13 had the lowest Recall of value 0.545454545. The settings were K=10, N=45 and 10000 iterations.

- **K-Modes F1 Measure**
    - Run 44 had the highest F1 Measure of value 0.247669774. The settings were K=10, N=45 and 20000 iterations.
    - Run 13 had the lowest F1 Measure of value 0.175765646. The settings were K=10, N=45 and 10000 iterations Means and K-Modes Rand Index comparison

6.3.8. K-Means and K-Modes Rand Index comparison

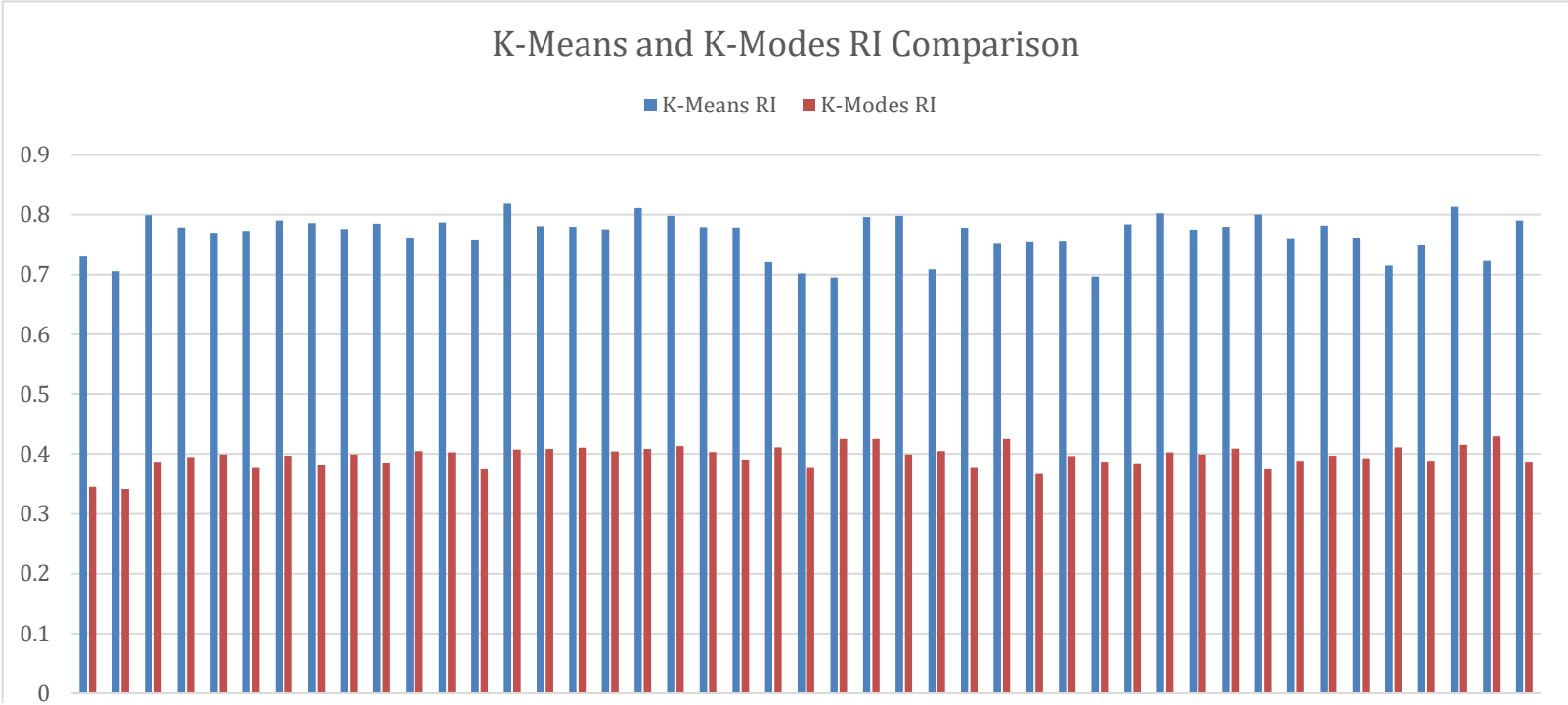

Figure 6.8: K-Means and K-Modes Rand Index comparison

Figure 6.8 is a graph showing comparison of Rand Index scores obtained for K-Means and K-Modes.

6.3.9. K-Means and K-Modes Precision comparison

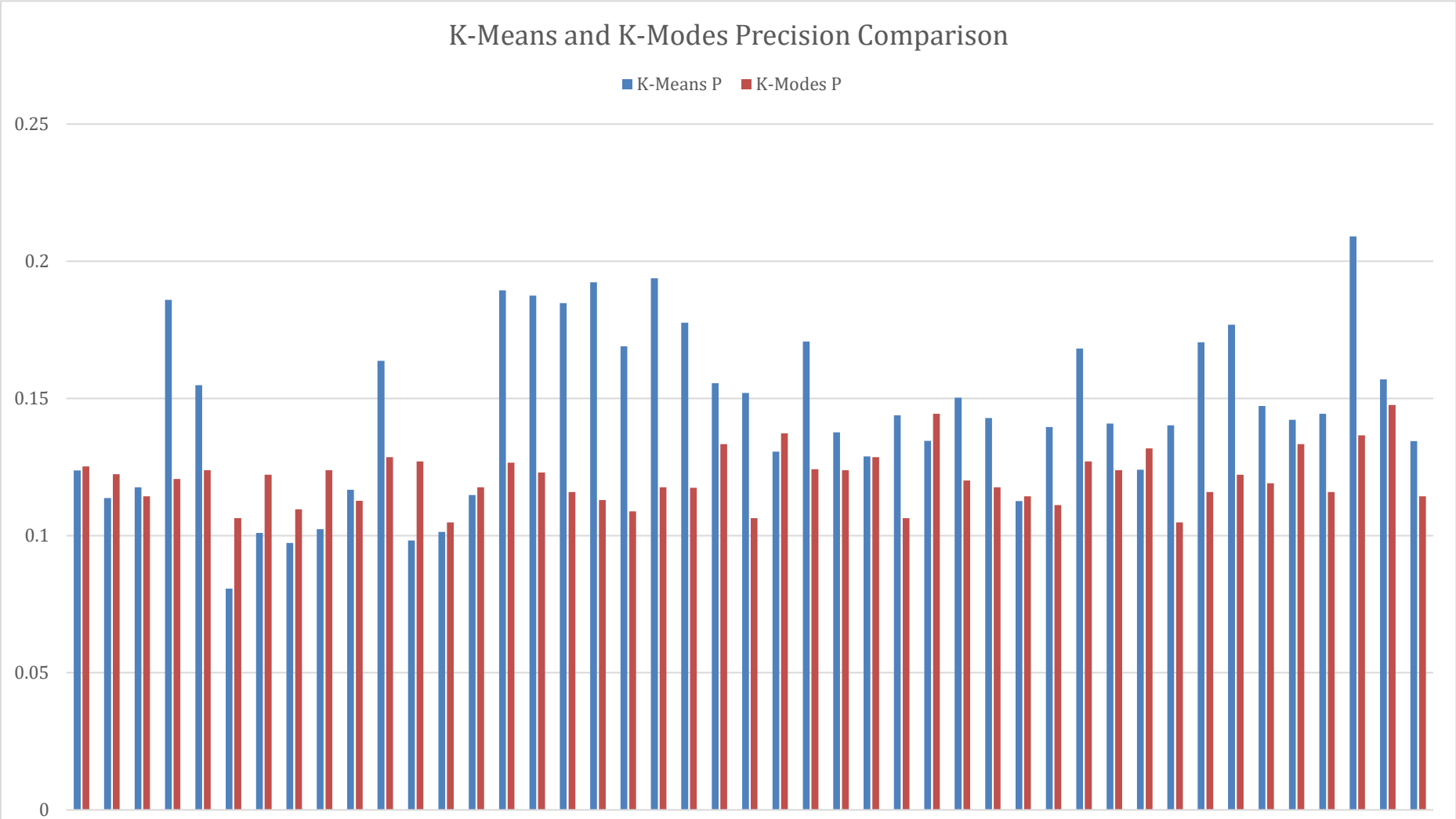

Figure 6.9: K-Means and K-Modes Precision comparison

6.3.10. K-Means and K-Modes Recall comparison

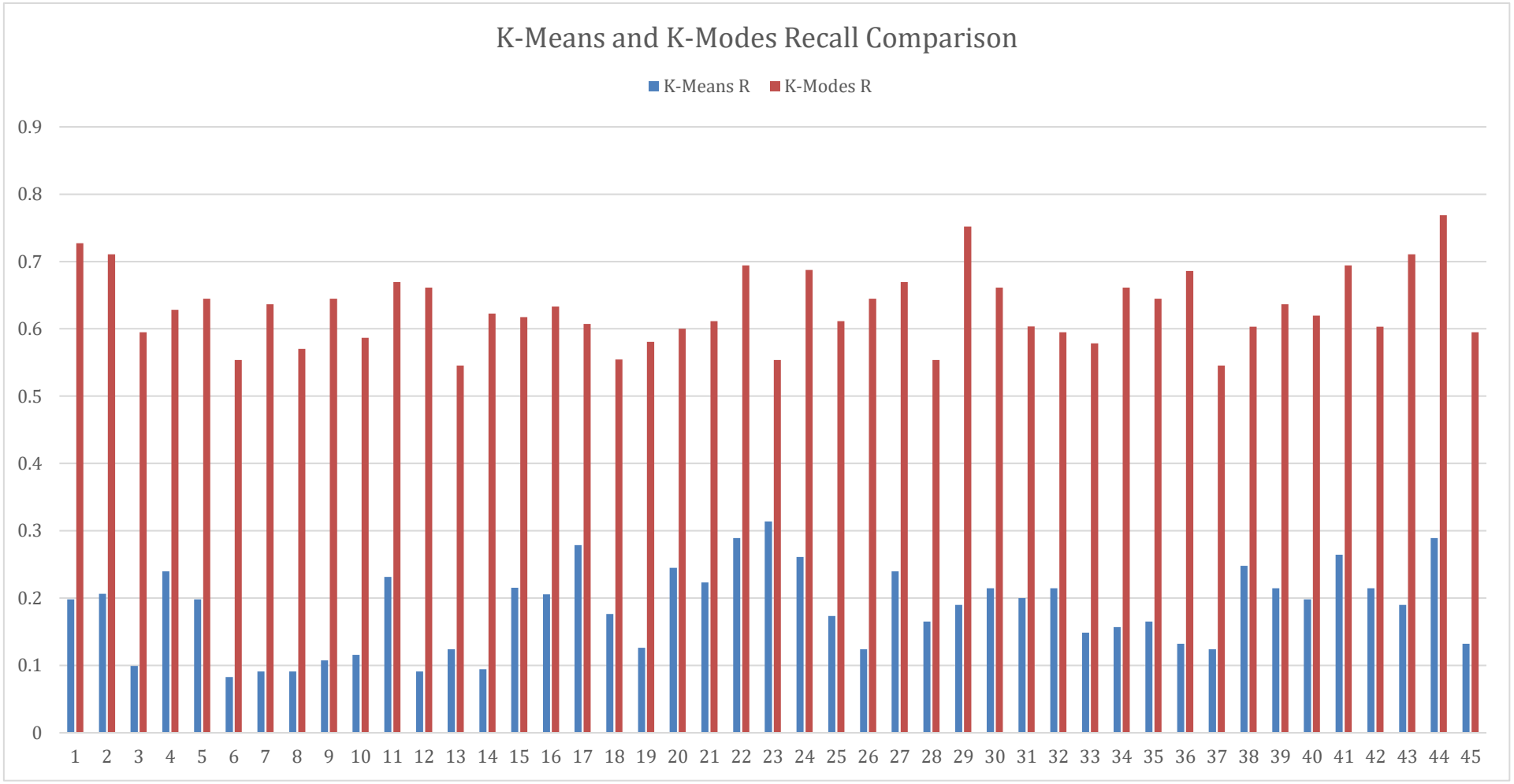

Figure 6.10:K-Means and K-Modes Recall comparison

6.3.11. K-Means and K-Modes F1 measure comparison

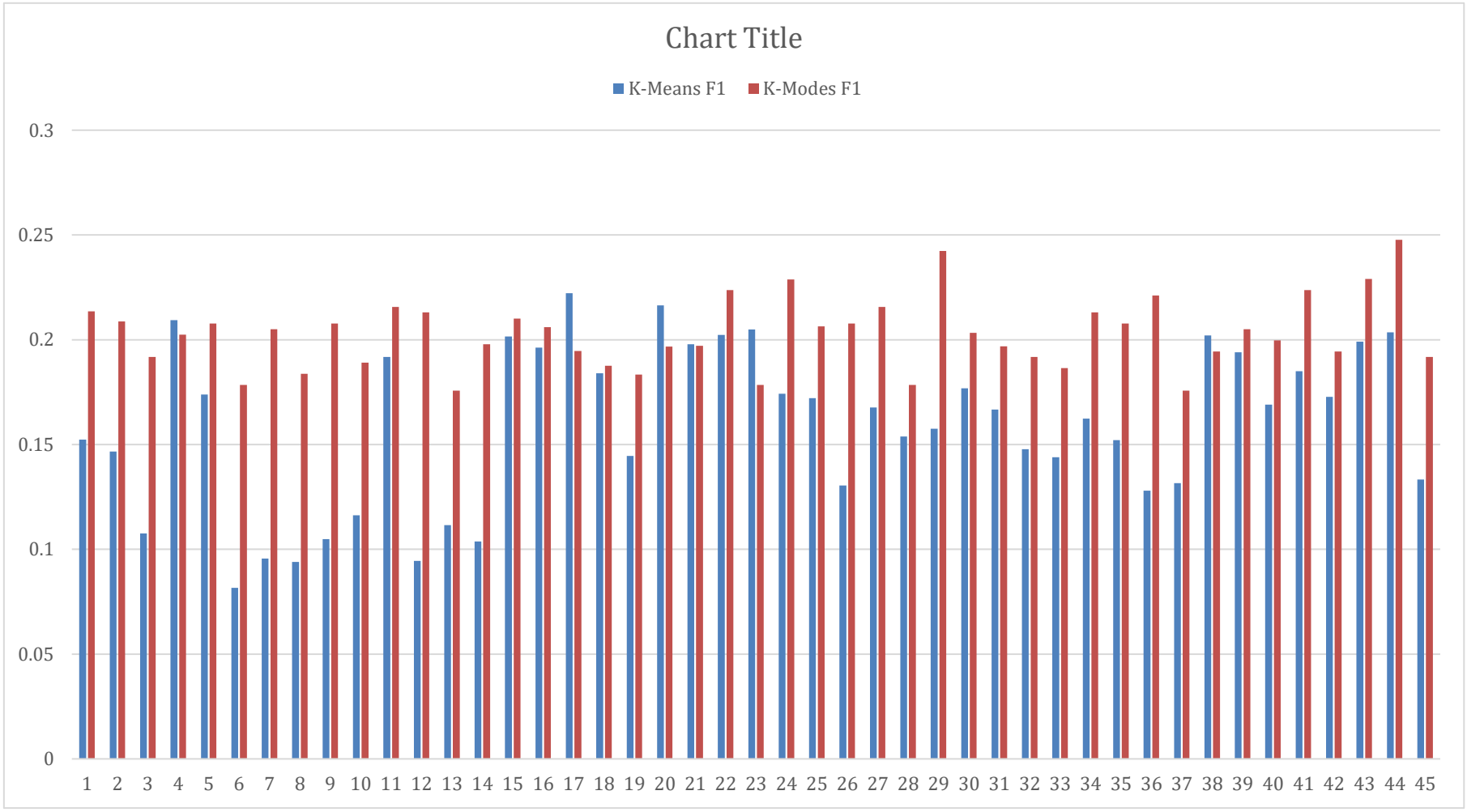

Figure 6.11: K-Means and K-Modes F1 measure comparison

6.3.12. Performance and time comparison

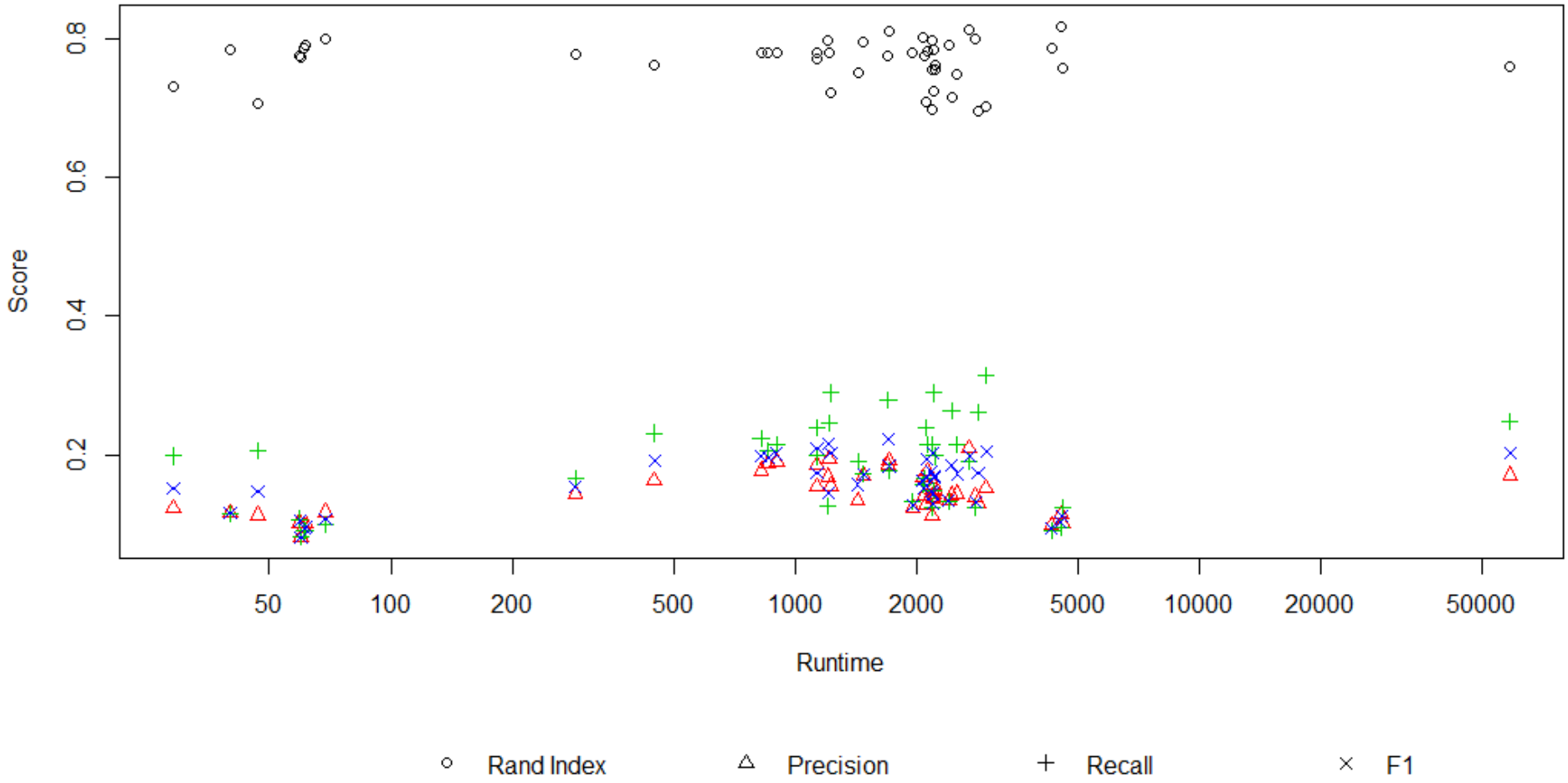

Figure 6.12: Performance and Time

Figure 6.12 shows the score of the algorithms against time. It demonstrates that there is no significant difference in score in relation to the amount taken. The least time taken does not translate to low score and the highest time taken does not translate to a high score either.

6.4. K-Means results

K-Means had a high Rand Index measure than K-Modes. It also had a relatively low Precision, Recall and F1 measures as shown in ***Figure 6.8: K-Means and K-Modes Rand Index comparison***.

6.5. K-Modes results

K-Modes has a higher recall than K-Means, this shows poor performance of K-Modes. There are stories that were correctly clustered, but there is also a higher number of stories which were incorrectly clustered. A cluster with known stories that belong to it, had some stories correctly assigned to it, and some stories that should have been assigned to it put elsewhere. Hence the number of high recall and low precision as shown by the graph. The average recall rate was 0.63054121, the highest and lowest Recall rates are shown above. K-Modes has a low Rand index measure.

$$F1 = \frac{2(Precision * Recall)}{Precision + Recall}$$

Equation 6.1

Equation 6.1 was derived from Equation 2.4 . F1 score is low if both precision and recall are low, or when 1 of them is low. In the case of K-Modes we have a high recall and low precision. Equation 6.1 was used to calculate the F1 score.

6.6. Results and findings

The results answered the three research questions, one main question and two sub questions.

6.7. Main research question

The experiment was set up to find an algorithm with decent performance that can be used to cluster news articles. The experiment manipulated different variables as shown below on the research question. There were Doc2vec parameters which were also manipulated.

The main research question was, given a set of news articles and K-means variations, how can we find the best variant with good performance to cluster news articles?
Formally: We are looking for a mapping to take a set of news articles (X), number of clusters (N), the K-Means variant algorithm (V) and the number of iterations (I), that can produce the best clustering results.

6.8. Sub question one

The first sub question was: What K-Means algorithm variation can accurately cluster online content into semantic clusters?

K-Means performs better than K-Modes. The settings and the best Rand Index value for the performance has been shown above in section 6.7. Those are the settings that can be used to achieve the best clustering results. The clustering algorithm is also supported by a vector representation technique, which is Doc2vec. It also has its own settings that are separate from the clustering algorithms. The settings for Doc2vec that achieved this good result are tabulated below.

Table 6.2: Best performing Doc2vec settings

w_size	w_window	w_min_count	w_workers	w_dm	w_alpha	w_min_alpha	w_epochs	w_start_alpha	w_end_alpha	nclusters
300	4	7	8	1	0.005	0.001	100000	0.001	-0.006	12

Doc2vec is an algorithm that is used to form vectors of documents. There are parameters that were varied to favour good performance. Some of the parameters are, w_size, which is dimensionality of the feature vectors. This affects the number of features that are formed. Increasing w_size increases the number of features, like wise decreasing w_size decreases the number of features, and it has an effect on the cluster results. w_window is the maximum distance between the current and predicted word within a sentence. This affects how the document vector is made. Increasing or decreasing w_window has an effect on the size of the context that is used to create features the word vectors. w_epochs is the number of iterations per cycle that the algorithm does to create the vectors. Table 6.2 shows the parameters that obtained a good result. Appendix A is a table with different Doc2vec settings.

6.9. Sub question two

The second sub question was: What is the effect of increasing the number of clusters on the accuracy of clustered content?

There was no significant difference in accuracy from increasing the number of clusters. The highest Rand Index obtained for K-Means for K value 12 was a score of 0.818318318. After incrementally changing the K value from 9 to 20 the highest Rand index obtained was 0.84. The detailed results are shown below.

6.10. Results for varying the K Value

6.10.1. K-Means results for varied K Value

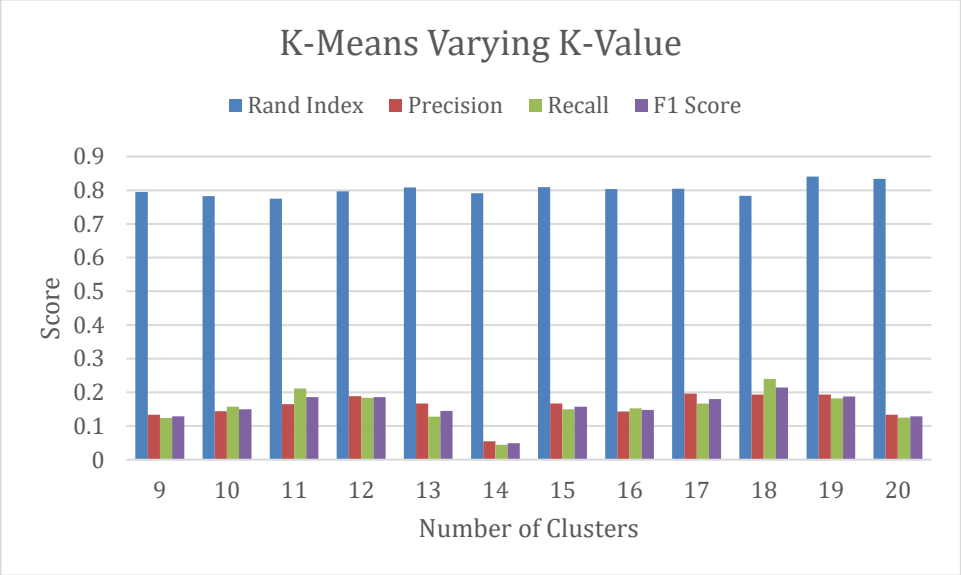

Figure 6.13: K-Means results for varied K value

Figure 6.13 shows results obtained on running the algorithm with different K values. The results above are for the K-Means algorithm, comparing the number of clusters and the score obtained. The K value was varied to evaluate the effect of increasing the number of clusters on the clustering performance. The clusters are varied from 9 to 20. The results for varying the K value for the K-Means algorithm are shown in the table below.

Table 6.3: K-Means results for varied K value

	1			2			3			4	
	Rand Index	**K-Value**	**Iterations**	**Precision**	**K-Value**	**Iterations**	**Recall**	**K-Value**	**Iterations**	**F1 Score**	**K-Value**	**Iterations**
Highest	0.84	12	10000	0.196078	10	20000	0.24	10	5000	0.214286	12	5000
Lowest	0.7746289	12	5000	0.054545	10	1000	0.044118	10	1000	0.04878	10	1000

Table 6.3: K-Means results for varied K value with maximum and minimum score obtained for each parameter.

6.10.2. K-Modes results for varied K Value

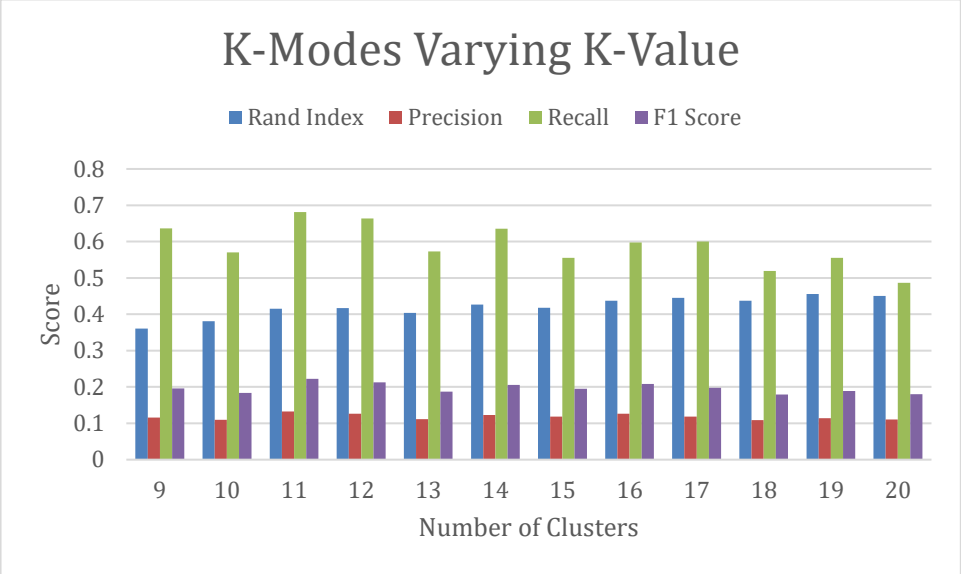

Figure 6.14: K-Modes results for varied K value

Figure 6.14 is a graph showing results obtained on running the algorithm with different K values. The results above are for the K-Modes algorithm, comparing the number of clusters and the score obtained. The K value was varied to evaluate the effect of increasing the number of clusters on the clustering performance. The clusters are varied from 9 to 20. The results for varying the K value for the K-Modes algorithm are shown in the table below.

Table 6.4: K-Modes results for varied K value

	1			2			3			4		
	Rand Index	**K-Value**	**Iterations**	**Precision**	**K-Value**	**Iterations**	**Recall**	**K-Value**	**Iterations**	**F1 Score**	**K-Value**	**Iterations**
Highest	0.4555556	10	20000	0.132773	10	20000	0.681034	10	20000	0.222222	10	20000
Lowest	0.3606061	8	1000	0.108466	10	10000	0.486486	10	1000	0.179431	10	10000

Table 6.4: K-Modes results for varied K value with the maximum and minimum score obtained for each parameter.

6.11. Discussion of results

The experiment's performance was done using evaluation metrics discussed in previous sections. The experiment was run with limited computing power, and small amount of data set. The research had a constraint of time. The study program had to be done and completed within a short space of time. The results of the comparison of two algorithms namely K-Means and K-Modes are presented in figures and tables in the previous sections. The results obtained and presented show that K-Means has a better performance over K-Modes.

6.12. General observation

The experiment ran very successfully. The research was done on a smaller data set. The experiment was run on a laptop with less computing power, this limited the number of runs. Not all possible settings and runs could be done as each run took a long time to complete, with the highest run going for as long as about ten hours. This affected the generalizability of the results. The experiment can be replicated on a server with higher resources and more data can be used for generalizability.

Clustering can be improved by incorporating Named Entity (NER) recognition into the K-Means algorithms. NER happens where an algorithm takes a news article and identifies relevant information. The relevant information can be people, time, product organisation, place or any relevant entity (Tkachenko & Simanovsky, 2012). NER can be used to scan an entire corpus of news articles and identify major entity tags discussed in them. The tags can help to quickly and efficiently cluster news articles (Foley et al., 2018).

NER is used in Natural Language Processing (NLP) tasks. The tasks include machine translation, text summarization and question answering applications. NER can also be used as a stand-alone tool for text search (Konkol, 2012).

Results can also be improved by implementing a multi-stage clustering technique. This is where initial clustering is done and then you take the cluster group and further cluster it to achieve finer clustering results. Multi-stage clustering is discussed in the work of Chakraborti & Dey (2016), where they propose a two-stage clustering technique. The first stage clusters phrases at sentence level. Similar phrases in a sentence discuss similar activities. The second stage is to apply a divisive clustering technique that

identifies subgroups in the sentences. They conclude that the two-stage system performs better than K-Means, Fuzzy C-Means and cosine similarity technique.

6.13. Conclusion

This chapter presented the results and findings of the experiment. The results were a comparison of K-Means and K-Modes algorithms. The results are presented in the form of graphs. They show the different scores obtained for the different evaluation metrics.

CHAPTER SEVEN
Conclusion

The research was motivated by the need to manage the vast amount of news articles on the internet. Users are presented with a huge amount of news articles, most of which are similar. Much of the news articles are redundant, as there is a lot of duplication on the internet. The huge amount of information and duplication makes it difficult for users to get information they want easily. The reading rate of a user has not changed, yet information continues to grow every day. The duplicated information is annoying to the user, because the user has got to sift through hundreds of pages, in order to update themselves on a subject matter. There is a need to cluster the duplicated news articles into containers.

Literature has shown that K-Means is widely used for the task of clustering. The challenge is that there are hordes of K-Means variants that have been developed. It is therefore difficult to pick the variant with good performance. Comparing the variants to identify the one with better performance was the focus of this research. However, the research was constrained in terms of time, as such not many variants were able to be compared.

A literature review was conducted to align this research in view of other past studies that have been conducted. The results of the literature review showed that K-Means is the most popular among other algorithms due to its performance. Research on comparing several K-means variants on the same set of parameters and constraints still needs to be done.

An experiment was conducted to answer the research questions and the goals. The goals were, to find the best variant of K-Means that can cluster news articles with a good performance and observe the effect of increasing the number of clusters on the performance of the algorithms. The experiment compared two algorithms namely K-Means and K-Modes. Forty-five news articles were collected from the internet. The collected news articles were converted to a numerical value using Doc2vec algorithm. The vector representation of news articles from Doc2vec was fed into the two clustering algorithms that were compared. The number of clusters were varied on each iteration, then the effect was observed. The iterations were run with different settings of Doc2vec. The settings have been discussed in chapter 5. The clustering performance was accessed using Rand Index, Precision, Recall and F1 score metrics. The results

obtained from the experiment show that K-Means performs better than K-Modes. The experiment results also show that there is no significant difference in performance that can be realised by increasing the number of clusters.

The experiment was run on a laptop with limited computing power. An iteration of Doc2vec took an average run time of about 1 hour, with the highest recorded run time of 10 hours. It is desirable to replicate the experiment on more powerful computing power like a server or cluster and to use more data. Results can also be improved by using NER, where clustering is done based on entities in the news articles. Entities like names of people, organisations, places and date and time can help to cluster stories discussing the same event. A future direction to pursue is multi-stage clustering, where clustering is done to put stories into containers. After which, the containers can be further clustered to increase the granularity of the clusters.

References

Al-rubaie, M. & Chang, J.M. 2018. Privacy Preserving Machine Learning : Applications , Threats , and Solutions. *IEEE security and privacy*: 1–16.

Altermatt, B. 2009. External Validity. : 1–3.

de Amorim, R.C. & Makarenkov, V. 2016. Applying subclustering and Lp distance in Weighted K-Means with distributed centroids. *Neurocomputing*, 173: 700–707.

Ari, Ç. & Aksoy, S. 2010. UNSUPERVISED CLASSIFICATION OF REMOTELY SENSED IMAGES USING GAUSSIAN MIXTURE MODELS AND PARTICLE SWARM OPTIMIZATION Department of Electrical and Electronics Engineering Bilkent University Selim Aksoy Department of Computer Engineering Bilkent University. *Computer Engineering*: 1859–1862.

Atefeh, F. & Khreich, W. 2015. A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence*, 31(1): 132–164. http://doi.wiley.com/10.1111/coin.12017.

Aziz, H.A. 2017. Comparison between Field Research and Controlled Laboratory Research. *Archives of Clinical and Biomedical Research*, 01(02): 101–104.

Azzopardi, J. & Staff, C. 2012. Incremental Clustering of News Reports. *Algorithms*, 5(3): 364–378. http://www.mdpi.com/1999-4893/5/3/364.

Berkhin, P. 2006. Survey of Clustering Data Mining Techniques. *Grouping Multidimensional Data: Recent Advances in Clustering*: 25–71.

Biscuitwala, K., Bult, W., Lécuyer, M., … T.P.-A.S. & 2013, U. 2013. Dispatch: Secure, resilient mobile reporting. *Citeseer*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.490.7565&rep=rep1&type=pdf 27 September 2019.

Bolelli, L., Ertekin, S., Zhou, D. & Giles, C.L.. L. 2007. A clustering method for web data with multi-type interrelated components. In *Proceedings of the 16th international conference on World Wide Web - WWW '07*. New York, New York, USA: ACM Press: 1121. https://www.scopus.com/inward/record.uri?eid=2-s2.0-35348916642&partnerID=40&md5=022b0849a7300ec253907af3e904856a.

Bornmann, L. 2013. Research Misconduct—Definitions, Manifestations and Extent. *Publications*, 1(3): 87–98. http://www.mdpi.com/2304-6775/1/3/87.

Bradley, P.S. & Fayyad, U.M. 1998. Anchoring strength of dual rubbed alignment layers in liquid crystal cells. In *15th International Conference on Machine Learning (ICML98)*. 91–99.

Bradley, P.S., Mangasarian, O.L. & Street, W.N. 1997. Clustering via Concave Minimization. *Advances in neural information processing systems*: 368–374.

Buchanan, B.G. 2019. Artificial intelligence in finance.

Chakraborti, S. & Dey, S. 2016. Multi-level K-means text clustering technique for topic identification for competitor intelligence. In *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*. IEEE: 1–10. http://ieeexplore.ieee.org/document/7549332/ 6 October 2019.

Coppin, B. 2004. *Artificial intelligence illuminated*.

https://books.google.com/books?hl=en&lr=&id=LcOLqodW28EC&oi=fnd&pg=PR5&dq=Coppin,+B.,+2004.+Artificial+intelligence+illuminated.+Jones+and+Bartlett+Publishers.&ots=sXxndEMCJZ&sig=ogKRI7UhSR0UmiNevgpQfcjgoyk 27 September 2019.

Deshmukh, M. 2014. Intrusion Detection System Using Clustering. , 5(5): 6656–6658.

Ding, C. & He, X. 2004. K-means clustering via principal component analysis. In *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*.

Easterbrook, S., Singer, J., Storey, M.-A. & Damian, D. 2008. Selecting Empirical Methods for Software Engineering Research. *Guide to Advanced Empirical Software Engineering*: 285–311. http://www.springerlink.com/index/n815725515063p2m.pdf.

Eneh, O. 2008. Research and computer applications in developing countries–A review. *researchgate.net*. https://www.researchgate.net/profile/Onyenekenwa_Eneh/publication/281275728_RESEARCH_AND_COMPUTER_APPLICATIONS_IN_DEVELOPING_COUNTRIES_-A_REVIEW/links/55de2b5208ae79830bb58625.pdf 29 September 2019.

Feldt, R. & Magazinius, A. 2010. Validity threats in empirical software engineering research - An initial survey. In *SEKE 2010 - Proceedings of the 22nd International Conference on Software Engineering and Knowledge Engineering*.

Fitriyani, S.R. & Murfi, H. 2016. The K-means with mini batch algorithm for topics detection on online news. In *2016 4th International Conference on Information and Communication Technology, ICoICT 2016*.

Foley, J., Sarwar, S. & Allan, J. 2018. Named Entity Recognition with Extremely Limited Data. *arxiv.org*. https://arxiv.org/abs/1806.04411 29 September 2019.

Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*.

Forsati, R., Meybodi, M., Mahdavi, M. & Neiat, A. 2008. Hybridization of K-Means and Harmony Search Methods for Web Page Clustering. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE: 329–335. http://ieeexplore.ieee.org/document/4740468/ 8 November 2018.

François-lavet, V., Henderson, P., Islam, R., Bellemare, M.G., François-lavet, V., Pineau, J. & Bellemare, M.G. 2018. An Introduction to Deep Reinforcement Learning. (arXiv:1811.12560v1 [cs.LG]) http://arxiv.org/abs/1811.12560. *Foundations and trends in machine learning*, II(3–4): 1–140.

Gong, L., Zeng, J. & Zhang, S. 2011. Text stream clustering algorithm based on adaptive feature selection. *Expert Systems with Applications*, 38(3): 1393–1399. https://linkinghub.elsevier.com/retrieve/pii/S095741741000669X.

Grira, N., Crucianu, M. & Boujemaa, N. 2004. Unsupervised and Semi-supervised Clustering: A Brief Survey. *A Review of Machine Learning Techniques for Processing Multimedia Content, Report of the MUSCLE European Network of Excellence (6th Framework Programme)*.

Hall, P., Dean, J., Kabul, I., Inc, J.S.-S.I. & 2014, U. 2014. An overview of machine learning with SAS® enterprise miner™. *Citeseer*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.648.3176&rep=rep1&type=pdf 3 October 2019.

Hanczar, B. & Nadif, M. 2019. Controlling and Visualizing the Precision-Recall Tradeoff for External Performance Indices. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 687–702. http://link.springer.com/10.1007/978-3-030-10925-7_42.

Handl, J., Knowles, J. & Dorigo, M. 2003. Ant-based clustering: a comparative study of its relative performance with respect to k-means, average link and 1d-som. *Design and application of hybrid …*.

Harland, D. 2011. *STEM student research handbook*. https://books.google.com/books?hl=en&lr=&id=WWJCuWuJC5oC&oi=fnd&pg=PR5&dq=Harland,+D.J.,+2011.+STEM+student+research+handbook.+NSTA+Press.&ots=gwpmqV5uBo&sig=ZecGzMFyEJ44HcQySUn8x-voJ1M 29 September 2019.

Henzinger, M. 2006. Finding Near-duplicate Web Pages: A Large-scale Evaluation of Algorithms. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*: 284–291. http://doi.acm.org/10.1145/1148170.1148222.

Hofmann, T. 2001. Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning*.

Hu, L., Zhang, B., Hou, L. & Li, J. 2017. Adaptive online event detection in news streams. *Elsevier*. https://www.sciencedirect.com/science/article/pii/S0950705117304550 8 October 2019.

IBM. 2018. The Power of One : IBM + Hortonworks The Challenge : Driving Analytics Benefit from Today ' s Data. , (May).

Jacobson, M.W. 2000. Biomedical publishing and the Internet: Evolution or revolution? *Journal of the American Medical Informatics Association*.

Jain, A.K. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*.

Jain, A.K., Duin, P.W. & Jianchang Mao. 2000. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1): 4–37. http://ieeexplore.ieee.org/document/824819/.

Jiménez-Buedo, M. & Miller, L.M. 2010. Why a Trade-Off? The Relationship between the External and Internal Validity of Experiments. *Theoria. Revista de Teoría, Historia y Fundamentos de la Ciencia*, 25(3): 301–321. https://www.pdcnet.org/theoria/content/theoria_2010_0025_0003_0301_0321 2 October 2019.

Kaelbling, L., Littman, M. & Moore, A. 1996. Reinforcemen t Learning? A Surv ey. *Journal of Artificial Intelligence Research*, 4(1): 237–285.

Kirk, R. 2012. Experimental design. *Wiley Online Library*. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118133880.hop202001 29 September 2019.

Konkol, M. 2012. Named entity recognition: technical report no. DCSE/TR-2012-04. https://dspace5.zcu.cz/bitstream/11025/21550/1/Konkol.pdf 29 September 2019.

Konstantinos, D.G. 2005. Mean Shift Clustering. *Computer*.

Kumar, V. & Rathee, N. 2011. Knowledge discovery from database using an integration of

clustering and classification. *International Journal of Advanced Computer Science and Applications*, 2(3). http://thesai.org/Publications/ViewPaper?Volume=2&Issue=3&Code=IJACSA&SerialNo=6.

Kutbay, U. 2018. *Partitional clustering*. https://books.google.com/books?hl=en&lr=&id=jnuQDwAAQBAJ&oi=fnd&pg=PA19&dq=Kutbay,+U.,+2018.+Partitional+clustering.+Recent+Applications+in+Data+Clustering&ots=z75DTvqzSi&sig=zlA5gt2x0GtoPxIQXwEv3pCNNrg 2 October 2019.

Landthaler, J., Waltl, B., Huth, D., Braun, D., Matthes, F., Stocker, C. & Geiger, T. 2017. Extending thesauri using word embeddings and the intersection method. In *CEUR Workshop Proceedings*.

Larsen, B. & Aone, C. 1999. Fast and effective text mining using linear-time document clustering.

Le, Q. & Mikolov, T. 2014. Distributed representations of sentences and documents. *jmlr.org*. http://www.jmlr.org/proceedings/papers/v32/le14.pdf 29 September 2019.

Li, L., Shang, Y. & Zhang, W. 2002. Improvement of HITS-based algorithms on web documents. In *Proceedings of the eleventh international conference on World Wide Web - WWW '02*. 527. http://portal.acm.org/citation.cfm?doid=511446.511514.

Li, W., Feng, Y., Li, D. & Yu, Z. 2016. Micro-blog topic detection method based on BTM topic model and K-means clustering algorithm. *Automatic Control and Computer Sciences*, 50(4): 271–277. http://link.springer.com/10.3103/S0146411616040040 8 October 2019.

Li, Y. & Liang, D.-M. 2019. Safe semi-supervised learning : a brief introduction. , 13(4): 669–676.

Liberty, E., Sriharsha, R. & Sviridenko, M. 2016. An Algorithm for Online K-Means Clustering. In *2016 Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*. Philadelphia, PA: Society for Industrial and Applied Mathematics: 81–89. http://epubs.siam.org/doi/10.1137/1.9781611974317.7 19 January 2019.

Lo, S.L., Chiong, R. & Cornforth, D. 2017. An unsupervised multilingual approach for online social media topic identification. *Elsevier*. https://www.sciencedirect.com/science/article/pii/S0957417417301847 8 October 2019.

Mahmood, S.T. 2011. Factors Affecting the Quality of Research in Education : Student ' s Perceptions. , 2(11): 34–40.

Mahmud, A., Ahmed, T., Hakim, A. & Sultana, T. 2018. Text mining of news articles to detect violation of human rights. http://dspace.bracu.ac.bd/xmlui/handle/10361/11419 27 September 2019.

Makkonen, J., Ahonen-Myka, H. & Salmenkivi, M. 2004. Simple semantics in topic detection and tracking. *Springer*. https://link.springer.com/article/10.1023/B:INRT.0000011210.12953.86 8 October 2019.

Malhotra, R. 2015. A systematic review of machine learning techniques for software fault prediction. *Applied Soft Computing Journal*.

Maxion, R.A. 2009. Experimental Methods for Computer Science Research. *Fourth Latin-American Symposium on Dependable Computing*.

Messina, A. & Montagnuolo, M. 2009. A generalised cross-modal clustering method applied to multimedia news semantic indexing and retrieval. In *Proceedings of the 18th international conference on World wide web - WWW '09*. New York, New York, USA: ACM Press: 321. http://portal.acm.org/citation.cfm?doid=1526709.1526753.

Meyer, D. 2016. How exactly does word2vec work ? *Uoregon.Edu,Brocade.Com*.

Mitchell, T.M. 2006. *The Discipline of Machine Learning*. http://www-cgi.cs.cmu.edu/~tom/pubs/MachineLearningTR.pdf.

Moerchen, F., Brinker, K. & Neubauer, C. 2007. Any-time clustering of high frequency news streams. *Citeseer*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.539.3521&rep=rep1&type=pdf 8 October 2019.

Moerchen, Fabian, Brinker, K. & Neubauer, C. 2007. Any-time clustering of high frequency news streams. *Proc. Data Mining Case Studies …*.

Mulwad, V., Finin, T., Syed, Z. & Joshi, A. 2010. Using linked data to interpret tables. In *CEUR Workshop Proceedings*.

El Naqa, I. & Murphy, M.J. 2015. What Is Machine Learning? In *Machine Learning in Radiation Oncology*.

Ng, M., Li, M., Huang, J. & He, Z. 2007. On the Impact of Dissimilarity Measure in k-Modes Clustering Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3): 503–507. http://ieeexplore.ieee.org/document/4069266/.

Nilsson, N. 2014. *Principles of artificial intelligence*. https://books.google.com/books?hl=en&lr=&id=mT-jBQAAQBAJ&oi=fnd&pg=PP1&dq=Nilsson,+N.J.,+2014.+Principles+of+artificial+intelligence.+Morgan+Kaufmann.&ots=hL-oeN1D9q&sig=sYGAGwFmNRrPNY-M4zQWftEjIbY 27 September 2019.

Oja, E. 2002. Unsupervised learning in neural computation. *Theoretical Computer Science*, 287: 187–207.

Oliver, A., Odena, A., Raffel, C., Cubuk, E.D. & Goodfellow, I.J. 2018. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*.

Ong, T., Chen, H., Sung, W. & Zhu, B. 2005. Newsmap : a knowledge map for online news. , 39: 583–597.

Orero, P., Doherty, S., Kruger, J.L., Matamala, A., Pedersen, J., Perego, E., Romero-Fresco, P., Rovira-Esteva, S., Soler-Vilageliu, O. & Szarkowska, A. 2018. Conducting experimental research in audiovisual translation (AVT): A position paper. *Journal of Specialised Translation*.

Owen, S. & Owen, S. 2012. *Mahout in action*. https://sisis.rz.htw-berlin.de/inh2011/12399459.pdf 27 September 2019.

Pal, A. & Gillam, L. 2013. Set-based Similarity Measurement and Ranking Model to Identify Cases of Journalistic Text Reuse . Introduction :

Papadimitriou, D., Fàbrega, L., Vilà, P. & … D.C. 2012. Measurement-based research: methodology, experiments, and tools. *dl.acm.org*.

https://dl.acm.org/citation.cfm?id=2378968 29 September 2019.

Paterson, C. 2006. News Agency Dominance in International News of the Internet. *Papers in International and Global Communications, Centre for International Communications Research, Leeds University*, (01/06): 1–24. http://ics-www.leeds.ac.uk/papers/cicr/exhibits/42/cicrpaterson.pdf%5Cnpapers2://publication/uuid/66C2F54C-434F-42FF-855A-CE59DFE31A8C.

Phuvipadawat, S. & Murata, T. 2010. Breaking news detection and tracking in Twitter. In *Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2010*.

Piskorski, J., Tanev, H., Atkinson, M., van der Goot, E. & Zavarella, V. 2011. Online News Event Extraction for Global Crisis Surveillance. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 182–212. http://link.springer.com/10.1007/978-3-642-24016-4_10.

Rai, P. & Singh, S. 2010. A Survey of Clustering Techniques. *International Journal of Computer Applications*.

Ranjan, R., Georgakopoulos, D. & Wang, L. 2016. A note on software tools and technologies for delivering smart media-optimized big data applications in the cloud. *Computing*, 98(1–2): 1–5. http://link.springer.com/10.1007/s00607-015-0471-8.

Redden, J. & Witschge, T. 2010. *A new news order? Online news content examined*. https://books.google.com/books?hl=en&lr=&id=f7VKBbuplgYC&oi=fnd&pg=PA171&dq=Redden,+J.+and+Witschge,+T.,+2010.+A+new+news+order%3F+Online+news+content+examined.+New+media,+old+news:+Journalism+and+democracy+in+the+digital+age&ots=gNDn4HYns-&sig=N4OHLi5QW5 3 October 2019.

Roberts, C. 2010. *The dissertation journey: A practical and comprehensive guide to planning, writing, and defending your dissertation*. https://books.google.com/books?hl=en&lr=&id=YYFuxoeoiWMC&oi=fnd&pg=PR1&dq=Roberts,+C.M.,+2010.+The+dissertation+journey:+A+practical+and+comprehensive+guide+to+planning,+writing,+and+defending+your+dissertation.+Corwin+Press.&ots=DnFFAo075e&sig=Z9kpn4n0hi 27 September 2019.

Robertson, S., Azizpour, H., Smith, K. & Hartman, J. 2018. Digital image analysis in breast pathology—from image processing techniques to artificial intelligence. *Translational Research*.

Rokach, L. & Maimon, O. 2010. Chapter 15— Clustering methods. *The Data Mining and Knowledge Discovery Handbook*.

Ross, N.C.M. & Wolfram, D. 2000. End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine. *Journal of the American Society for Information Science and Technology*.

Ross, S. & Morrison, G. 2004. Experimental research methods. *researchgate.net*. https://www.researchgate.net/profile/Gary_Morrison/publication/201382131_Experimental_Research_Methods/links/004635266ef06ed3e6000000.pdf 29 September 2019.

Russell, S.J. & Norvig, P. 2003. *Artificial Intelligence A Modern Approach; PearsonEducation*.

Saad, F.H., Mohamed, O.I.E. & Al-Qutaish, R.E. 2012. Comparison of Hierarchical Agglomerative Algorithms for Clustering Medical Documents. *International Journal of Software Engineering & Applications*, 3(3): 1–15.

http://www.airccse.org/journal/ijsea/papers/3312ijsea01.pdf.

Sahani, A., Sarang, K., Umredkar, S. & Patil, M. 2013. Automatic text categorization of marathi documents using clustering technique. In *2013 15th International Conference on Advanced Computing Technologies (ICACT)*. IEEE: 1–5. http://ieeexplore.ieee.org/document/6710543/.

Salloum, S., Al-Emran, M. & Shaalan, K. 2017. Mining text in news channels: a case study from Facebook. *researchgate.net*. https://www.researchgate.net/profile/Said_Salloum/publication/318946704_Mining_Text_in_News_Channels_A_Case_Study_from_Facebook/links/59873e71aca27266ada22568/Mining-Text-in-News-Channels-A-Case-Study-from-Facebook.pdf 8 October 2019.

Sankaranarayanan, J., Samet, H., … B.T.-… geographic information & 2009, U. 2009. Twitterstand: news in tweets. *dl.acm.org*. https://dl.acm.org/citation.cfm?id=1653781 29 September 2019.

Sathya, R. & Abraham, A. 2013. Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *International Journal of Advanced Research in Artificial Intelligence*, 2(2): 34–38.

Scaiella, U., Ferragina, P., Marino, A. & Ciaramita, M. 2012. Topical clustering of search results. In *WSDM 2012 - Proceedings of the 5th ACM International Conference on Web Search and Data Mining*.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1): 1–47. http://portal.acm.org/citation.cfm?doid=505282.505283.

Shabbir, J. & Anwer, T. 2018. Artificial Intelligence and its Role in Near Future. , 14(8): 1–11. http://arxiv.org/abs/1804.01396.

Shahnazarian, D., Hagemann, J., Aburto, M. & Rose, S. 2013. Informed-Consent-Booklet-4.4.13.pdf.

Shaikh, M.S. & Fegad, V. 2013. Analysis and Modelling of Strong-AI to Engineer BIONIC Brain for Humanoid Robotics Application. *American Journal of Embedded Systems and Applications*.

Shameem, M.U.S. & Ferdous, R. 2009. An efficient K-means algorithm integrated with Jaccard distance measure for document clustering. In *1st South Central Asian Himalayas Regional IEEE/IFIP International Conference on Internet, AH-ICI 2009*.

Shamoo, A. & Resnik, D. 2009. *Responsible conduct of research*. https://books.google.com/books?hl=en&lr=&id=dP7oKntCUUUC&oi=fnd&pg=PR9&dq=Shamoo,+A.E.+and+Resnik,+D.B.,+2009.+Responsible+conduct+of+research.+Oxford+University+Press.&ots=PF55O9Qpjt&sig=oi8zwzNtSMunWG8fTrYJVqup_cQ 27 September 2019.

Shi, Z. 2011. *Advanced Artificial Intelligence*.

Sokolova, M. & Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*.

Sun, A., Lim, E.-P. & Ng, W.-K. 2002. Web classification using support vector machine. In *Proceedings of the fourth international workshop on Web information and data management - WIDM '02*. New York, New York, USA: ACM Press: 96. http://portal.acm.org/citation.cfm?doid=584931.584952.

Teknomo, K. 2006. K-means clustering tutorial. *sigitwidiyanto.staff.gunadarma.ac.id*. http://sigitwidiyanto.staff.gunadarma.ac.id/Downloads/files/38034/M8-Note-kMeans.pdf 27 September 2019.

Tkachenko, M. & Simanovsky, A. 2012. Named entity recognition: Exploring features. In *11th Conference on Natural Language Processing, KONVENS 2012: Empirical Methods in Natural Language Processing - Proceedings of the Conference on Natural Language Processing 2012*.

Vishwakarma, S., Nair, P.S. & Rao, D.. 2017. A Comparative Study of K-means and K-medoid Clustering for Social Media Text Mining. *ijasret.com*. http://ijasret.com/VolumeArticles/FullTextPDF/95_IJASRET-A_Comparetive_Study_of_K-means_and_K-medoid_Clustering_for_Social_Media_Text_Mining.pdf 8 October 2019.

Wang, W., Wu, Y., … C.T.-… C. on M. & 2015, U. 2015. Adaptive density-based spatial clustering of applications with noise (DBSCAN) according to data. *ieeexplore.ieee.org*. https://ieeexplore.ieee.org/abstract/document/7340962/ 29 September 2019.

Whelan, C., Harrell, G. & Wang, J. 2015. Understanding the K-Medians Problem. : 219–222.

Witten, I.H., Frank, E., Hall, M.A. & Pal, C.J. 2016. *Data Mining: Practical Machine Learning Tools and Techniques*.

Xiong, H., Wu, J. & Chen, J. 2009. K-Means Clustering Versus Validation Measures: A Data-Distribution Perspective. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2): 318–331. http://ieeexplore.ieee.org/document/4711107/.

Zhang, B., Hsu, M. & Dayal, U. 1999. K-Harmonic means - A data clustering algorithm. *HP Laboratories Technical Report*.

Zhao, Y. & Karypis, G. 2002. Evaluation of hierarchical clustering algorithms for document datasets. In *International Conference on Information and Knowledge Management, Proceedings*.

Appendix A

Table below shows the different settings applied to Doc2vec.

Setting No.	w_size	w_window	w_min_count	w_workers	w_dm	w_alpha	w_min_alpha	w_epochs	w_start_alpha	w_end_alpha	nclusters	Run Time in secs
1	100	5	500	8	1	0.025	0.001	1000	0.001	-0.006	8	28.96
2	100	5	500	8	1	0.025	0.001	1000	0.001	-0.006	8	46.87
3	100	5	5	8	1	0.025	0.001	1000	0.001	-0.006	10	68.84
4	500	5	5	8	1	0.005	0.001	10000	0.001	-0.006	10	1133.25
5	500	5	5	8	1	0.005	0.001	10000	0.001	-0.006	10	1133.46
6	500	5	5	8	1	0.005	0.001	1000	0.001	-0.006	10	59.87
7	500	5	5	8	1	0.005	0.001	1000	0.001	-0.006	10	61.59
8	500	5	5	8	1	0.005	0.001	1000	0.001	-0.006	10	61.14
9	500	5	5	8	1	0.005	0.001	1000	0.001	-0.006	10	59.49
10	300	3	5	8	1	0.005	0.001	1000	0.001	-0.006	10	40.05
11	300	3	5	8	1	0.005	0.001	10000	0.001	-0.006	10	448.97
12	300	3	5	8	1	0.005	0.001	100000	0.001	-0.006	10	4306.63
13	300	4	7	8	1	0.005	0.001	100000	0.001	-0.006	10	4585.39
14	300	4	7	8	1	0.005	0.001	100000	0.001	-0.006	12	4540.81
15	300	3	7	8	1	0.005	0.001	20000	0.001	-0.006	12	900.54
16	300	3	7	8	1	0.005	0.001	20000	0.001	-0.006	12	857.29
17	300	3	7	8	1	0.005	0.001	40000	0.001	-0.006	12	1700.70
18	300	3	7	8	1	0.005	0.001	40000	0.001	-0.006	14	1709.73
19	300	3	7	8	0	0.005	0.001	40000	0.001	-0.006	14	1208.55
20	300	3	7	8	0	0.005	0.001	40000	0.001	-0.006	12	1212.31
21	300	3	7	8	0	0.005	0.001	30000	0.001	-0.006	10	823.23
22	300	3	7	8	0	0.005	0.001	40000	0.001	-0.006	10	1226.59
23	300	3	7	8	0	0.005	0.001	100000	0.001	-0.006	10	2971.82

24	300	3	7	8	0	0.005	0.001	100000	0.001	-0.006	12	2839.17
25	300	3	7	8	0	0.005	0.001	50000	0.001	-0.006	10	1477.82
26	300	3	7	8	1	0.005	0.001	50000	0.001	-0.006	10	2191.18
27	300	3	7	8	1	0.005	0.001	50000	0.001	-0.006	10	2113.99
28	300	3	7	8	0	0.005	0.001	10000	0.001	-0.006	10	285.72
29	300	3	7	8	0	0.005	0.001	50000	0.001	-0.006	10	1429.73
30	300	3	7	8	1	0.005	0.001	50000	0.001	-0.006	9	2180.90
31	300	3	7	8	1	0.005	0.001	50000	0.001	-0.006	11	2213.33
32	300	3	7	8	1	0.005	0.001	50000	0.001	-0.006	10	2180.93
33	300	3	7	8	1	0.005	0.001	50000	0.001	-0.006	10	2193.92
34	150	3	7	8	1	0.005	0.001	50000	0.001	-0.006	10	2065.61
35	150	3	7	8	1	0.005	0.001	50000	0.001	-0.006	10	2086.70
36	100	3	7	8	1	0.005	0.001	50000	0.001	-0.006	10	1955.19
37	150	3	7	8	1	0.005	0.001	50000	0.001	-0.006	10	2795.99
38	150	3	7	8	1	0.005	0.001	50000	0.001	-0.006	10	58832.16
39	150	3	7	8	1	0.005	0.001	50000	0.001	-0.006	10	2123.48
40	200	3	7	8	1	0.005	0.001	50000	0.001	-0.006	10	2218.72
41	250	3	7	8	1	0.005	0.001	50000	0.001	-0.006	10	2437.28
42	300	3	7	8	1	0.005	0.001	50000	0.001	-0.006	10	2519.92
43	350	3	7	8	1	0.005	0.001	50000	0.001	-0.006	10	2701.01
44	150	4	7	8	1	0.005	0.001	50000	0.001	-0.006	10	2200.28
45	200	4	7	8	1	0.005	0.001	50000	0.001	-0.006	10	2399.96