** **

**PATAN COLLEGE FOR PROFESSIONAL STUDIES**

**Research Methodologies and Emerging Technologies**

**Topic Identification using clustering algorithm big document vs small documents**

**Contextual Report**

**Submitted by                                                                        Submitted to**

Lalit Sunar (1720846)                                                       University Of Bedfordshire

Luton

Table of Contents

[Abstract 1](#_Toc30766817)

[Topic identification using clustering algorithm Big vs. small Documents 2](#_Toc30766818)

[1. Introduction 2](#_Toc30766819)

[1.1 Background 3](#_Toc30766820)

[1.2 Purpose of Study 4](#_Toc30766821)

[1.3 Aim 4](#_Toc30766822)

[1.4 Objective 4](#_Toc30766823)

[1.5 Target User 4](#_Toc30766824)

[1.6 Deliverables 5](#_Toc30766825)

[1.7 Methodology 5](#_Toc30766826)

[1.8 Research Question 5](#_Toc30766827)

[1.9 Signification of study 5](#_Toc30766828)

[1.10 Limitation of study 6](#_Toc30766829)

[2. Literature Review 6](#_Toc30766830)

[2.1 Technical Background of Topic identification using Clustering 6](#_Toc30766831)

[2.1.1 Artificial Intelligence 6](#_Toc30766832)

[2.1.2 Clustering 7](#_Toc30766833)

[2.1.2.1 Classification of Clustering 8](#_Toc30766834)

[Flat Clustering 9](#_Toc30766835)

[Hard Clustering: Assigns each document/ object as a member of exactly one cluster. 9](#_Toc30766836)

[2.1.2.2 Clustering Techniques 9](#_Toc30766837)

[2.1.3 Document Clustering 10](#_Toc30766838)

[2.2 Relevant Work done is Topic identification using clustering 11](#_Toc30766839)

[2.2.1 Similar System – 1 11](#_Toc30766840)

[2.2.1.1 Brief Overview 11](#_Toc30766841)

[2.2.1.2 Working of the system 11](#_Toc30766842)

[2.2.1.3 Processing and Similarity calculation 12](#_Toc30766843)

[2.2.1.4 About GAHC algorithm 12](#_Toc30766844)

[2.2.1.5 Test Result 12](#_Toc30766845)

[2.2.1.6 Conclusion 13](#_Toc30766846)

[2.2.1.7 Critical analysis of the system 13](#_Toc30766847)

[2.2.2 Similar system -2 13](#_Toc30766849)

[2.2.2.1 Brief Overview 13](#_Toc30766850)

[2.2.1.2 Find of system from literature review 14](#_Toc30766851)

[2.2.1.3 Materials and Methods system -2 14](#_Toc30766852)

[16](#_Toc30766853)

[2.2.1.4 Conclusion -2 17](#_Toc30766854)

[2.2.1.5 Critical Analysis of system -2 17](#_Toc30766855)

[2.4 Research Question Analysis 17](#_Toc30766856)

[2.4.1 Comparison of different clustering algorithm with their complexities 17](#_Toc30766857)

[2.4.2 Preprocessing of clustering in topic identification of documents 18](#_Toc30766858)

[2.5 Finding 19](#_Toc30766859)

[3. Data Analysis & Market Research 19](#_Toc30766860)

[3.1 Survey question, response and their analysis 20](#_Toc30766861)

[3.1.5 I often deal with unlabeled or random labelled (document without file name) document. 23](#_Toc30766862)

[3.1.6 Features of your current operating system(window, Linux, mac os) is enough to sort or group your documents. 24](#_Toc30766863)

[3.1.7 Do you think often the topic you give to your documents best describe the content of your documents contains? 24](#_Toc30766864)

[3.1.8 Do you think Sorting and grouping documents according to their contents would be great? 25](#_Toc30766865)

[4. Project Plan 26](#_Toc30766866)

[5. Planning artefact 29](#_Toc30766867)

[5.1 Requirements 29](#_Toc30766868)

[5.1.1 Hardware Requirement 29](#_Toc30766869)

[5.1.2 Software Requirement 29](#_Toc30766870)

[5.1.3 Functional Requirement 30](#_Toc30766871)

[5.1.4 Non-Functional Requirement 30](#_Toc30766872)

[5.2 System Design 30](#_Toc30766873)

[5.2.1 Use case Diagram 31](#_Toc30766874)

[5.2.2 Activity Diagram 31](#_Toc30766875)

[5.2.3 Sequence diagram 32](#_Toc30766876)

[5.2.4 Low-fi 32](#_Toc30766877)

[5.2.5 Mid-fi 34](#_Toc30766878)

[5.3 Testing Strategy 35](#_Toc30766879)

[5.4 Evaluation Strategy 35](#_Toc30766880)

[6. Conclusion 35](#_Toc30766881)

## Abstract

Document are the part of our daily life. From small office to big corporates deals with huge amount of data. Not just big corporates instead text or small documents are also being product at the personal level. Each document has topic. But often people have to deal with untitled document and some people also don’t think that their title defines their documents. Using clustering technique and using its algorithms, topic can be identified and detected. A topic can be produced and identified by analyzing its content. Clustering various algorithm for topic modeling. K-mean algorithm has been one of the widely used algorithms for topic identification. It’s variants and also other algorithm such as hierarchical clustering technique. Different document has different length of words. Increase in word length can also increase the dimensionality of the data. And analyzing the content of big documents can be harder than smaller one. So, big document and small document are compared in this research. And this also show how increase dimensionality of the content effects the performance of the trained model.

# Topic identification using clustering algorithm Big vs. small Documents

## Introduction

Human start their journey from total oblivion. They cannot tell right from wrong, high from low. But as they start their journey, they began to give label to things. From their ability called pattern recognition, they begin to make meaningful mental heuristic about world. And that heuristic is built from finding pattern out of nothing. As they experience new thing, they begin to cluster their information to connect them with similar things. And this defines every aspect of human life. Likewise, machine can now also perform clustering to find meaningful pattern in random data with power of suitable algorithms. (Gulhane & Deshmukh, 2016) defined **Cluster analysis** or **clustering** asa way of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). Machine can performer more which were impossible such as document classification, customer segmentation, stat Analysis, Fraud detection and so on. So, in other words, it helps us to give meaning insights to our data.

Text and documents are the big part of our daily live. Millions of documents are being produced daily by offices, schools, hospitals etc. Since texts are ways of conveying information and knowledge, every text has some content on it. Its content makes it, what it is. And title of the document provides us an information about its content. Even if we are not aware of its content, from its title we can get idea or essence of document.

So, in other words the topics are the essence of the document. But sometimes we deal with the documents without topics. And it’s hard to give it title without analyzing its content. But now with the help of clustering, we can analyze the content of our document to find suitable topics for our documents. Every document is composed of different words or keywords which gives idea of it. By analyzing these words, we can get idea of which words can be suitable topic for our document. And it can be done by using document clustering.

This research is going to discuss about document clustering. With the help of document clustering topics can be identified. Here small document refers to small parts such as abstract of the document and big documents refers to full documents. Since, everyone know that there are no right answers. Even two different writers can choose different topics to same contents. But main idea behind the topic is to find the small but relevant word to convey the summary of content. So, it might not be always perfect and as expected by people but it’s accuracy will be measured by degree of words which defines the content of the documents. So, while talking about the topic identification, keywords of those documents should analyze. Sometimes topics can also be different than things which are given in the contexts. Sometimes topic could be deduced from the keyword of the question which might not contain in the document, but they are trying to refer. This first part of research going to analyze what it is.

In Section 2, a literature review will be done to find out some research question to broad our understanding. In section 3 the analysis will be done on the data collected from the survey and find it’s meaning. In section 4, project plan will be discussed and see previous work was done and also how to implement our research into development. And in section 5, Planning of artefacts is discussed where requirement, design and testing & evaluation strategy is discussed.

## Background

There are many researches which has been done in past. Topic identification Using Artificial intelligence now more than just arithmetic and simple logical problems can be solved. Artificial Intelligence is a field of computer science which provides way of making machine more intelligent than ever. There is certain thing like facial recognition, diseases detection, object recognition which serves to greater purpose such as self-driving cars, recommendation systems, modern surveillance and so on. A problem can be solved which require intelligence. Now, those can be performed which could not be done with the previous technology. So, Artificial Intelligence really a revolution in technology.

Machine learning is one of the applications for Artificial Intelligence which helps to build system which can learn and improve itself.  Among them unsupervised learning method called clustering helps us to find reference from datasets. In clustering the data points are divided into number of groups. And group can be formed on the basics of their similarities. With these powerful technologies system that can perform topic identification can be built.

document clustering will used for topic identification. Document clustering can help automatically group a large documents corpus into clusters without clusters without consuming too much human labours. But one problem is that unsupervised clustering algorithm like K-means still cannot output the main topics of the formed clusters. Simply treating the keywords set as the features of documents provides only distance information between documents. (Sun, 2014)

## Purpose of Study

Topic of the certain documents provides us the insight of the its content. It provides us a way to summaries our long documents into the words which can gives us idea. Since huge number of documents is being dealt every day. Documents are now part of our daily life. Each human action is being documented like our records, our performance and so on. According to the (Business insider, 2014) In 2013 around 500 billion office documents were produced by everyone around the globe. Documents Clustering has played a significant role in many fields like information retrieval, data mining, understanding etc. Most of the problems of documents clustering are high dimension, big volume and complex semantics. So, purpose of my study is to build topic identification system which can analyse content of the text to find topics.

## 1.3 Aim

* To identify topics of big and small documents using clustering techniques.

## 1.4 Objective

* To identify best algorithm for text-based clustering
* To learn Python for machine learning
* To learn about document clustering using machine learning
* To implement best algorithm possible algorithm for identification of topics
* To build server and web interface for trained modal.

## 1.5 Target User

Since, written digital document and text are part of huge variety of office, so this system can be useful for small and big offices. And it can many people. But main Target user for this project are.

* Student who want to analyze contents of their essay or articles.
* Content creators or blog writers to find topic and also to find out frequent keywords
* Office personal with deal with unlabeled documents.

## 1.6 Deliverables

According to need and nature of my project there are 3 deliverables for my project. And they are given below.

* A flask server: A web server builds with the micro framework flask to communicate with modal and frontend. This will as the part where client-side web application can interact with the system. Flask will be used to serve trained model by handling request sent by users.
* Angular web application: A web application build with framework of JavaScript called Angular. It will work as interface where use can paste text or upload document to get perform topic identification.
* AI Modal: By the use of the python library the system will be trained with given document by using clustering algorithms and then suitable topic will be selected by given list of words.
* Report: This report with all the design and development will also the part of deliverables

By the complete integration of these technologies, system will be created.

## 1.7 Problem and challenges

Since, in 2nd year, AI was only touched in surface and clustering and unsupervised learning are huge topics of machine learning. So, learning it will be part of challenge. And while searching about it, there were only research but not actual implementation. So, lack of review of practically implemented system can also be challenge. Since, it is machine learning project and lack of mathematical background can be also challenging. As a researcher and developer, Nodejs is preference of mine but since python will be use used for training modal. So, using python for server side will make it easy to implement. Since, lack of experience in python’s server side implementation can be also challenging part of this work.

## 1.8 Methodology

As stated by (Kothari & Garg, 2014), Quantitative research is based on the quantitative measurement of some characteristics. It is applicable to phenomena that can be expressed in terms of quantities.

By analyzing the nature of this project and system. Quantitative research methodology is going to be used. Since relation between several things will be examine. And this research is going to examine things such as algorithm, different system as well as will be concern with statistical result. So, Quantitative research is going to used.

## 1.9 Research Question

1. Comparison of different clustering algorithm with their complexities?
2. How clustering can be in our system?
3. How clustering can be applied in big and small documents?

## 1.10 Signification of study

There is certain significance of this study.

* It helps big organizations to sort their documents
* It helps to analyse contents
* It helps to draw reference from the words used in the documents
* It also can be used to categorize documents

## 1.11 Limitation of study

As Everybody know that the even document with the same content can have different topics. And The topic can be varying by the person who is writing that documents. Also, since, this is system is only base on only clustering not on heuristics. So, this only select or suggest title which are present in document. In life scenario, sometime title can be also words which are not present in content. For example, a child from America can write essay about how he spent his 4th of July with father and he can talk about parade and title can be Independence Day even though he never used that keyword in that document. Another huge limitation of the system is that only can be used to identify topics of documents written in English. In sense it can said that it is limited by language barrier.

## Literature Review

## Technical Background of Topic identification using Clustering

## Artificial Intelligence

According to (Copeland, 2019) Artificial Intelligence is the ability of a digital computer or computer-controlled robots to perform tasks commonly associated with intelligent beings. The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristics humans, such as the ability to reason, discover meaning, generalize, or learn from past experience. Since the development of the digital computer in 1940’s, it has demonstrated that computer can be programmed to carry out very complex task- as for example discovering proofs for mathematical theorems or playing chess-with great proficiency.

In the mid-1980s at least four different groups reinvented the back-propagation learning algorithm first found in 1969 by Bryson And Ho. The algorithm was applied to many learning problem in computer science and psychology, and the widespread dissemination of the result in the collection parallel Distributed Processing caused great excitement(Russel & Norvig, 2014)

There a long history of AI. its influence can be seen since classical era, story where human always wanted to build machine with intelligence. And many classical philosophers tried to analyze inner working of human intelligence. But formal development in this area was only started after 1956 A.D. In the conference at Dartmouth College in Hanover, new Hemisphere, the term **Artificial Intelligence** was used. But there was certain problem all along the way and implementing it wasn’t simple because before making intelligent system first they should be intelligent technology to support. So, even though there were many mathematical advancements, development was still frozen. Since, every research and development should support by government funding and there was lack of it. So, time between 1974-80 is considered as the AI winter. And the history of AI has many up and down. After 1980 after government of country such as United Kingdom started to fund, it was revived. And also, Japan show some effort to support. But it had its up and down. Later, time between 1987 to 1994 it was again was at its down condition because of collapsing market of early general-purpose computer as well as lacking funding. But later due to advancement of technology, this field reached to new height. In 1997, Deep Blue of IBM beat grand master of chess. And there also Alpha Go also beat world champion. And there is always rapid development in many fields such as vision, supervised learning, unsupervised learning, data mining and so on. Due to its high provide and possibility it is considered as one of the most promising field in technology right now. (Lewis, 2014).

## 2.1.2 Clustering

Clusters are considered as the group containing data objects that are similar to each other, while data object in different clusters are not. Though there’s not universally agreed upon and precise definition of term cluster. One of them is that cluster can be seen as a set of entities which are alike, and entities from different clusters are not alike.

As pointed out by (Wunsch & Xu, 2008), there are procedure of cluster analysis with following four basic steps:

1. **Feature selection or extraction**

Feature selection choose distinguishing features form set of candidates while features extraction utilizes some transformation to generate useful and novel features from the original ones. Clearly, feature extraction is potentially capable of producing features that could be of better use in potentially capable of producing features that could be of better use in uncovering that data structure. However, feature extraction may generate features that are not physically interpretable, while feature selection assures that retention of the original physical meaning of the selected features. In the literature, these two terms sometimes are used interchangeably without further identifying the difference. Both features selection and feature extraction are very important to the effectiveness

1. **Clustering algorithm design or selection**

This step usually consists of determining an appropriate proximity measure and constructing a criterion function. Intuitively, data objects are grouped into different clusters according to whether they resemble one another or not. Almost all clustering algorithms are explicitly connected to some particular definition of proximity measure. Some algorithms even work directly on the proximity matrix. Once a proximity measure is determined, clustering could be construed as an optimization problem with a specific criterion function. Again, the obtained clusters are dependent on the selection of the criterion function. The subjectivity of cluster analysis is thus inescapable.

1. **Cluster Validation**

Given data set, each clustering algorithm can always produce a partition whether or not there really exist a particular structure in the data. Moreover, different clustering approaches usually lead to different cluster of data, and even for the same algorithm, the selection of parameter or the presentation order of input pattern may affect the final results. Therefore, effective evaluation standards and criteria are critically important to provide users with a degree of confidence for the clustering result.

1. **Result interpretation**

The ultimate goal of clustering is to provide users with meaningful insights from the original data so that they can develop a clear understanding of the data and therefore effectively solve clear understanding of the data and therefore effectively solve the problem encountered.

## 2.1.2.1 Classification of Clustering

As implied by (Deokar, 2013) Clustering algorithm may be classified as listed below:

Flat Clustering: This cluster creates a set of clusters without any explicit structure that would relate clusters to each other; It’s also called exclusive clustering Hierarchical clustering: creates a hierarchy of clusters.

### Hard Clustering: Assigns each document/ object as a member of exactly one cluster.

**Soft Clustering:** Distribute the document/object over all users.

**Algorithms**

(Deokar, 2013) also emphasize upon the in which algorithm they fall under.

Agglomerative (Hierarchical clustering)

K-means (Flat clustering, Hard Clustering)

EM Algorithm (Flat Clustering, soft clustering)’

Hierarchical Agglomerative Clustering (HAC) and K-means Algorithm has been applied to text clustering in straightforward way. Typically, its usages normalized, TF-IDF-weighted vectors and cosine similarity.

## 2.1.2.2 Clustering Techniques

As from previous section 2.2.1, there are main algorithms which are being used. Each algorithm is based on certain approach. Below two of the main techniques are given.

* 1. **K-meaning Technique**

It is partitioning clustering techniques. The technique segregates the given data set into k disjoint subsets(clusters) through an iterative process, congregating to confined local minima. (Arora, et al., 2016) K-means clustering technique is a way to organize the data items based on some features into k group. Where K is positive integer. Grouping is completed by reducing the total amount of squares of distances between cluster centroid and data. (Patel & Thakral, 2016)

Algorithm:

* + 1. Set center of the clusters
    2. Attribute the nearest group to every data point
    3. Set place of every group to the mean value of all data-point which fit to that group
    4. Recap sets i-iii until all object is not classify.
  1. **Hierarchical Clustering technique**

According to (Kaur & Kaur, 2015) hierarchical clustering is an effective method to evaluate the clusters from given dataset as it combines a number of clusters into the form of tree called dendrogram in such a way that the sub clusters are the fundamentally similar to one another. The first phase in making dendrogram is to discover distances between the objects of dataset. When all the distances between the clusters have discovered, the merging or splitting operation is applied on the given. Agglomerative algorithm is always more preferable than divisive algorithm.

## 2.1.3 Document Clustering

Clustering can be defined as the grouping data objects by their similarities and separating them according to their dissimilarities. So, it can be deduce that the document clustering is way of grouping documents according to their similarities and forming their clusters. Where documents which can similarities form clusters and belong to same cluster and clustering having different properties are separated in different clusters. It falls under the unsupervised learning of machine learning. It is also very useful technique for statically analyzing data. There are many fields where it can be used such as image analysis, pattern recognition and outlier.

As point out in (Li et al., 2014) , that document clustering techniques have been receiving more and more attention as a fundamental and enabling tools for efficient organizations, navigation, retrieval, and summarization of huge volumes of text documents. In that paper it was argued that with good document clustering method, computer can automatically organize a documents corpus into meaningful hierarchy, which enables an efficient browsing and navigation of the corpus.

## Relevant Work done is Topic identification using clustering

### Similar System – 1

The paper was added to IEEE explore at 2014, June 05. Title of this paper is **Topic Detection based on Group Average Hierarchical Clustering**. This was published on journal of **International conference on Advanced Cloud and Big Data**. The authors of this research are Ni GAO, Ling GAO, Yiyue HE, Hai Wang, Qian SUN from the Department of Information Science and Technology, Northwest Univercity Xi’an, Chine.

### 2.2.1.1 Brief Overview

This system was mainly focused on Group Average Hierarchical Clustering (GAHC) which was suitable for processing of big data on the network. (Gao, et al., 2014) point out that core idea of this algorithm is to divide big data into smaller groups. Vector modal was used to represent news documents and time & place were considered for similarity calculation model.

Research also pointed out that nowadays the automatically and accurately detecting and tracking and reporting of a topic of a new event is critical problem. So, their main aim was to collect and organize the scattered information about special events effectively. It was also pointed out that

### 2.2.1.2 Working of the system

In this system they firstly related document of news were collected and processed. And then secondly those documents were represented by vector space model Thirdly, the similarity of documents based on weight of time and space was calculated. And at last the topic structure was obtained by using GAHC.

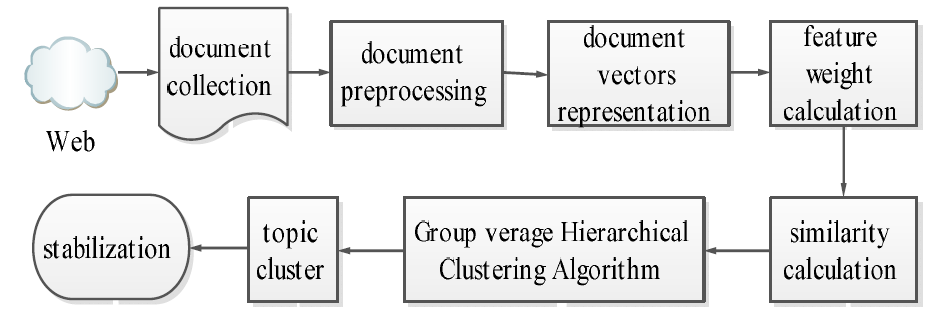
In this system News event can had many properties such as title, time, human name and place name. Vector space was composed by words and their corresponding weights and documents vector.

Figure the knowledge learning process of topic detection

### 2.2.1.3 Processing and Similarity calculation

First elements of news were acquired by employing information extraction technology combined with natural language understanding for document Vector s Representation. And then Feature weight were calculated. Then similarity calculation model Based on weight of time and space calculated by suing cosine function computing.

### 2.2.1.4 About GAHC algorithm

It was argued that traditional hierarchical clustering algorithm aims to build the hierarchy of documents. But he points out that the problem that document in each layer can not achieve the level of abstraction in practice. And also, a regular PC can’t meet the computing requirement of data processing. So, it was suggested that GAHC was a better algorithm which was based on Retrospective Event Detection (RED) than ordinal hierarchical clustering.

### 2.2.1.5 Test Result

Research showed that the effectiveness of GAHC based on similarity calculation model considering weights of time and space was better than based on traditional similarity calculation model considering content. Other’s algorithm which was used were not snot efficient and saleable as compared to GAHC.

### 2.2.1.6 Conclusion

After applying GAHC and reducing the complexity of document clustering algorithm by similarity calculation. In conclusion their experimental result showed that GAHC improved the performance of topic detection.

### 2.2.1.7 Critical analysis of the system

### This research and experiment showed that GAHC was suitable for processing the big data on the network. It has really overcome issue of efficiency and scalability which couldn’t be solve by other algorithms. And their implantation other processes to reduce the complexity of document was wise decision since it was processing big data, these small of small scale can have huge impact of processing at big level. But if look at it, they only worked in algorithm without considering possible alternative algorithms from other technique such as k-means or it’s variants which can also be used. But there is also high detailed implantation of the system which can help to build similar system. So, with little downside it is good quality system.

### Similar system -2

This system is based on the research paper. The topic of this research paper is **A study on Topic Identification using K means clustering algorithm: Big vs. Small Documents.** It was published in **Advances in Computational Science and Technology** in 2017. The author of this paper are Pema Gurung Bengaluru and Rupali Wagh from Christ University, Bengaluru.

### 2.2.2.1 Brief Overview

According to (Gurung & Wagh, 2017) this paper revolves around the idea of topic identification using k-means clustering algorithm. Here researchers discuss about the different algorithms such as K-means clustering, fuzzy clustering, Hierarchical and flat clustering which can be used in this system. They explained that the fundamental idea in clustering was to group document into different clusters. And they also talk about the how each document can be represented by the vectors with assigned weights in words documents. They used tf-idf(term frequency-inverse document frequency) scheme. They explained that in topic modeling cluster of words where they emphasize that each word in the clusters have a probability of occurrence for a given topic and how topic modelling can be used for extracting hidden meaning form the set of text documents. Here it can said that this system implements then system k-mean algorithms to small & big document and how dimension of the document can affect accuracy and the performance of the system.

### 2.2.1.2 Find of system from literature review

After review different literature review they pointed out that due of quality of cluster which can be form after grouping of the documents the topic identification of the big document is still challenging. After feature and inherent complexities of data increases, that also affects the quality of the result. And they used conference papers and their abstract as datasets. They compared result implemented in abstract (small documents) vs full papers(big documents). Since both small and big document should have same topic but there was difference in result

### 2.2.1.3 Materials and Methods system -2

Researchers obtained clusters from the document term matrix. They emphasize that document term matrix was representation of document collection frequency of the word in the document determined the matrix value. The implementation of the system can be seen below:

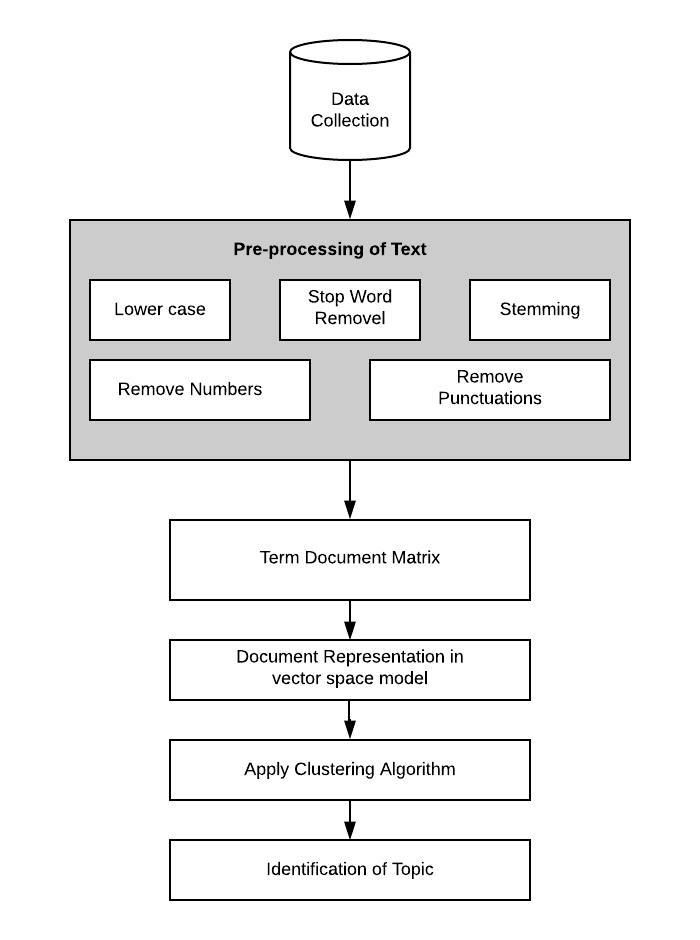


Figure Generic Model of Text Clustering for topic Identification

From above model it can see how researcher implement system. First, they preprocess text, so that it can be clean and easily further processed. And then it was converted into term matrix. And after that Term Document Matrix was converted into document representation in vector space model. And after that clustering algorithm was applied to identify the topic.

### A screenshot of a cell phone Description automatically generated

Figure Methodology

### 

As show inf fig.3 it can seen that how researcher used methodology where first they compare corpus. And document where split into small documents(abstract) and big document (text from introduction to conclusion).

### 2.2.1.4 Conclusion -2

At the conclusion researcher tried to emphasize on how high dimensionality of big document can have impact on result. They applied K-means to both data with low dimensionality and high dimensionality and then result were observed.

### 2.2.1.5 Critical Analysis of system -2

This paper provide insight into how topic identification system should implemented. There clear implementation of the system. And it’s results also showed how dimension of the document can affect the result of the result. But according to the (Steinbach, et al., 2000) variants of k-means such bi-section k-means can also have high accuracy and efficiency with limited time constraints. But overall system and its implementation is really quite impressive. There are clear and simple set of diagrams and illustration which make it really easily to understand. And it also focusses on preprocessing of the document before clustering. So, this paper and its implementation is overall good.

### 2.4 Research Question Analysis

### 2.4.1 Comparison of different clustering algorithm with their complexities

While performing clustering, there are many things that can cause variation the accuracy. So, as it is known that our whole system works by the algorithm it uses. So, better choice of algorithm can give best the performance and accuracy. There are many algorithms out there for clustering of documents like k-means and its variant such as bisecting k-means, fuzzy clustering, hierarchical. While analyzing these algorithms it should keet in mind that tradeoffs are everywhere. There are always place where it has to be compromised to get what really essential for our project. For example, Hierarchical clustering algorithms are considered good algorithm for clustering but if time complexity is considered then k-means and its variant outperform other’s clustering techniques. (Gupta & Rajavat, 2015) pointed out that hierarchical based clustering is more efficient than partitioning based. By both methods they found out clusters and by comparing they got both methods pros and cons. They saw that BIRCH (balanced iterative reducing and clustering using hierarchies) is efficient clustering for large database and it scans all database at one scan, and it is also not so much sensitive with noise, but it takes more time than K-means. So, they were applying BIRCH (balanced iterative reducing and clustering using hierarchies) for large data set.

Even though they prefer BIRCH to be more efficient but even they admit that it is suitable for large data set, but time complexity is issue for it. And if both dimensions are considered, word length and time complexity than k-means can be better option.

### 2.4.2 Preprocessing of clustering in topic identification of documents

While clustering document then first thing to do is to document preprocessing. The text documents are initially in the unstructured format. So, first it should preprocess to form numeric vectors. According to (Kadhim, et al., 2014) In text processing step after reading the input text documents, it divides the text document to features which are called tokenization, it represents that text document in a data representation as a vector space whose components are the feature and their weights which are obtained by the frequency of each feature in that text document, after that it removes the non-informative features such as (stop words, numbers and special characters). Then they emphasize that main objective of this step is that it improves the quality of features and at the same time reduces the difficulty of mining process.

1. Tokenization: Tokenization In our sense doesn’t not only separates into basic processing units but also interprets and groups isolated tokens to create a higher token. Raw texts are preprocessed and segmented into textual unites. The data must be processed in the three operations: First operation is to convert document to word counts which is equal to bag of word (BOW). The second operation is removing empty sequence i.e. this step comprises cleansing and filtering (example, whitespace, collapsing, stripping, extraneous control characters).(c)
2. Stop words Elevation: There are certain words which are part of every documents. Common features such ca conjunctions such as or, and, but and pronouns he, she, it etc. They should be removed from words, can also remove very rare words example, words that only occur in m or fewer document, for example m=6.(c)

1. Stemming: It is the process of removing affixes (prefixes and suffixes) from features i.e. the process derived for reducing inflected (or sometimes derived) words to their stem. The stem needs to be identified to be original morphological roots of the words and it is usually sufficiently related through words map to the similar stem. This process is used to reduce the number of features in the feature space and improve the performance of the clustering when the different forms of features are stemmed into a single feature.
2. Vector Space Model: In this process the present document clustering after performing tokenization keywords, removing stop words and performing stemming keywords in the document are remained. For example, suppose if the term and the document matrix. Row represents words in each document and column represent documents. Each cell represents the words count in the particular document. If the number of documents is increasing the size of the matrix will be increased. And it will lead to high dimensionality.

### 2.5 Finding

Through the literature review and after analyzing different system. It was found about that K-means algorithm was good choice of topic identification using clustering. Hierarchical algorithms were also promising but if speed, accuracy and efficiency is considered then K-means seems wise choice. So, for implementation K-means algorithm will be used.

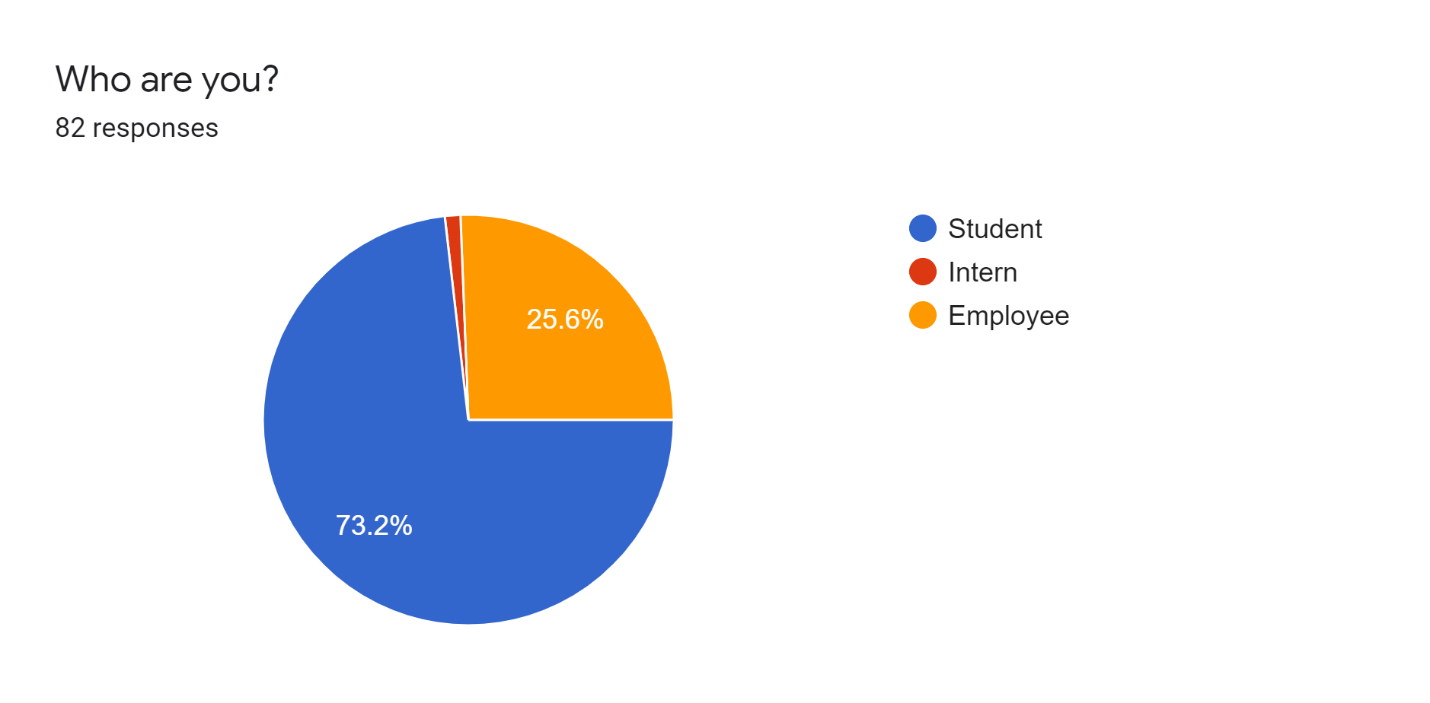
## Data Analysis & Market Research

Product, system or service are built for the potential customers. While building system certain assumption of about targeted user will use it. But there are certain assumptions has to be tested. So, for that purpose Market Research is needed. Market research helps us to determine whether our assumption about our potential customer will find it useful or not. So, survey and interview were conducted. For this project, survey has been done. There are total nine questions which were based on assumption about system.

In my survey there were total 82 number of participants. While taking survey, things such as privacy, purpose of survey and the its significance where discussed. Below there are analysis of certain data from the response.

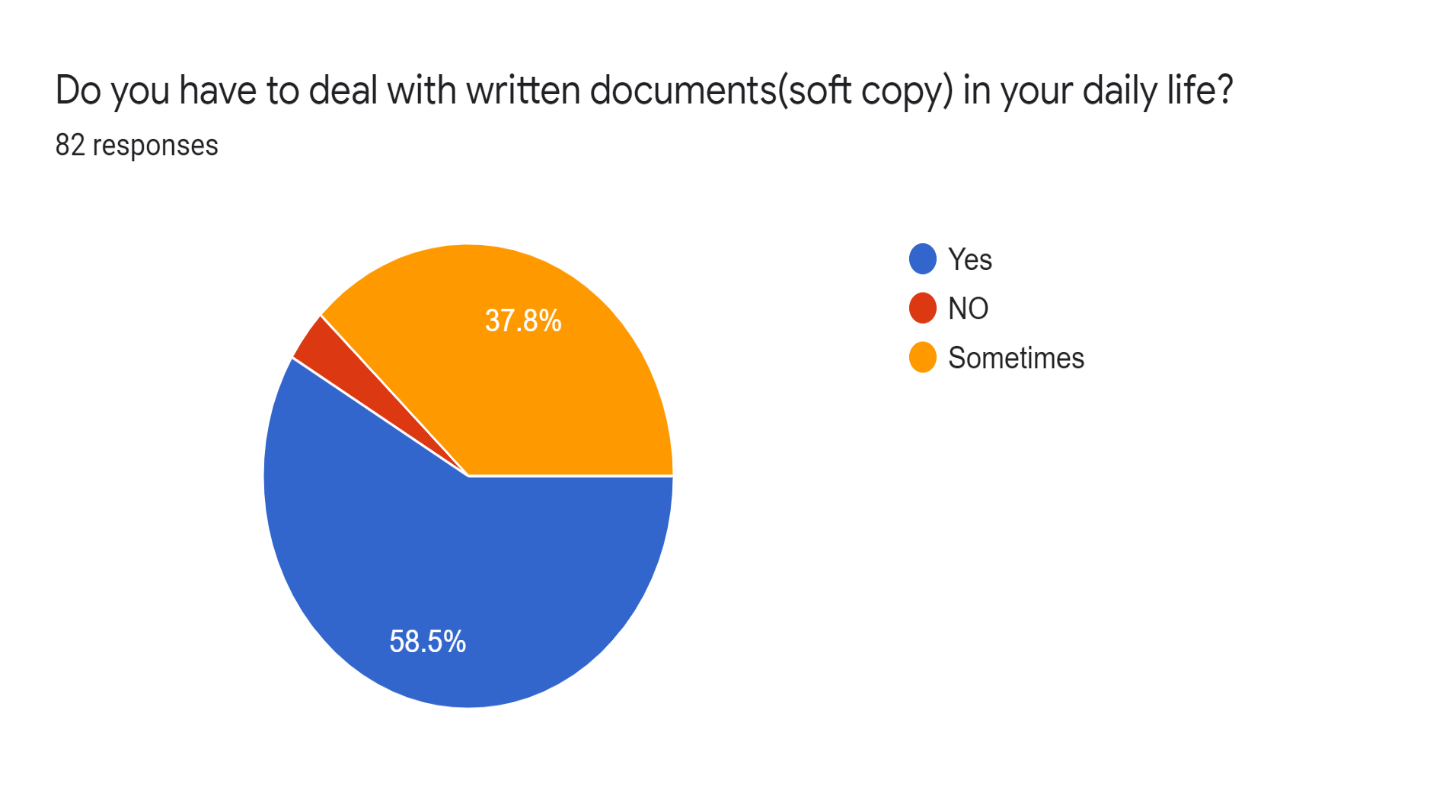
## Survey question, response and their analysis

**3.1.1 Who are you?**



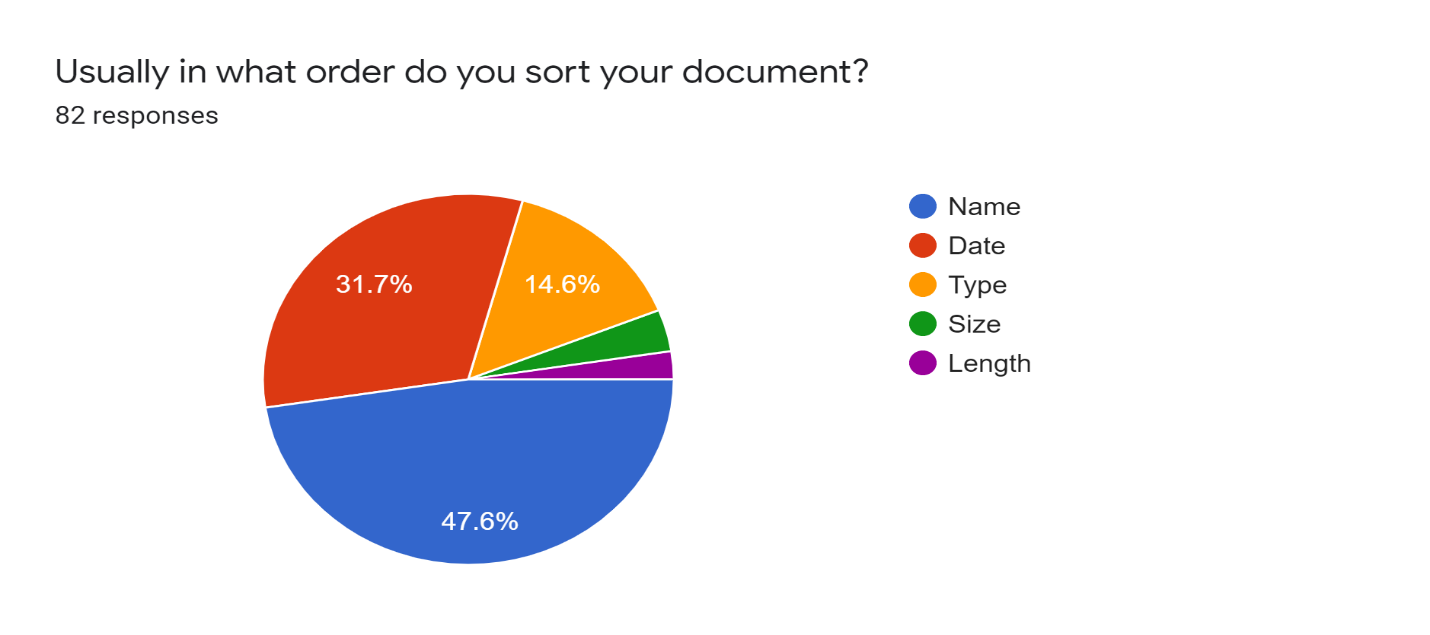
In this question our main target was to identify our participants. Since, different type of the person has to deal with different type of the documents. So, type of the participants can also affect the result.

From above data it can seen that our participant is mostly student (73.2%) and then Employee (25.6). And them intern about 1.2% were intern.

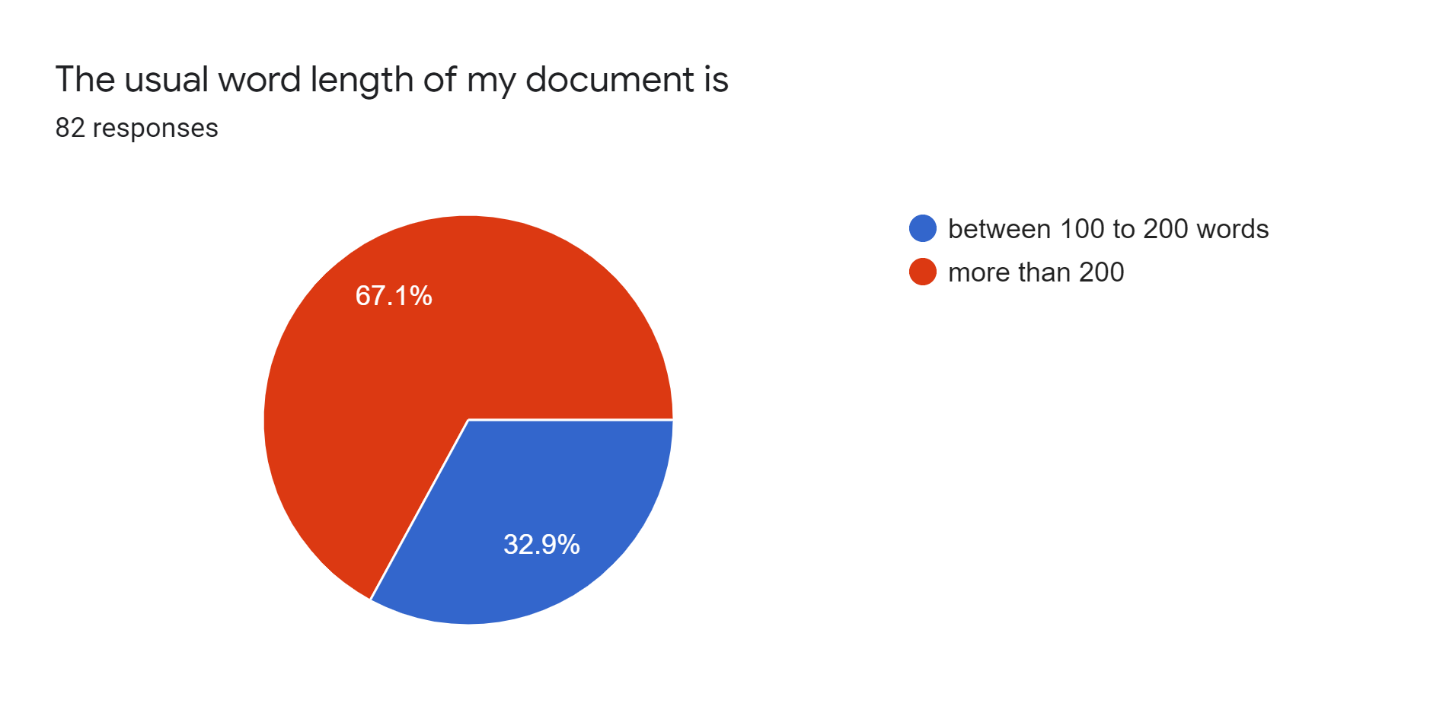
**3.1.2 Do you have to deal with written documents(soft copy) in your daily life?**

The main purpose was this question was to determine how many of our participants were handling data in their daily life. High number od daily user show, how many participants are daily dealing with digital documents. And according to survey majority i.e. about 58.5% dealing with digital documents in their daily life. And about 37.8% of people said they only deal with these documents for sometimes. And about 3% of them said they don’t.

**3.1.3 Usually in what order do you sort your documents?**

Since, Sorting will also be the part of my system. So, main purpose of this question was to find out in which order does majority of the people sort their documents. According to above response of survey it can be seen that 47.6% of the participants sort their file by order of name, 31.7% of people sort their file according to date and about 14.6% sort their document in order of type. And remaining portion sort their file according to size and length. So, from above data, it can be concluded at majority of the people use name order which is mostly order of number and alphabet to sort their file.

**3.1.4 The usual word length of my document is.**

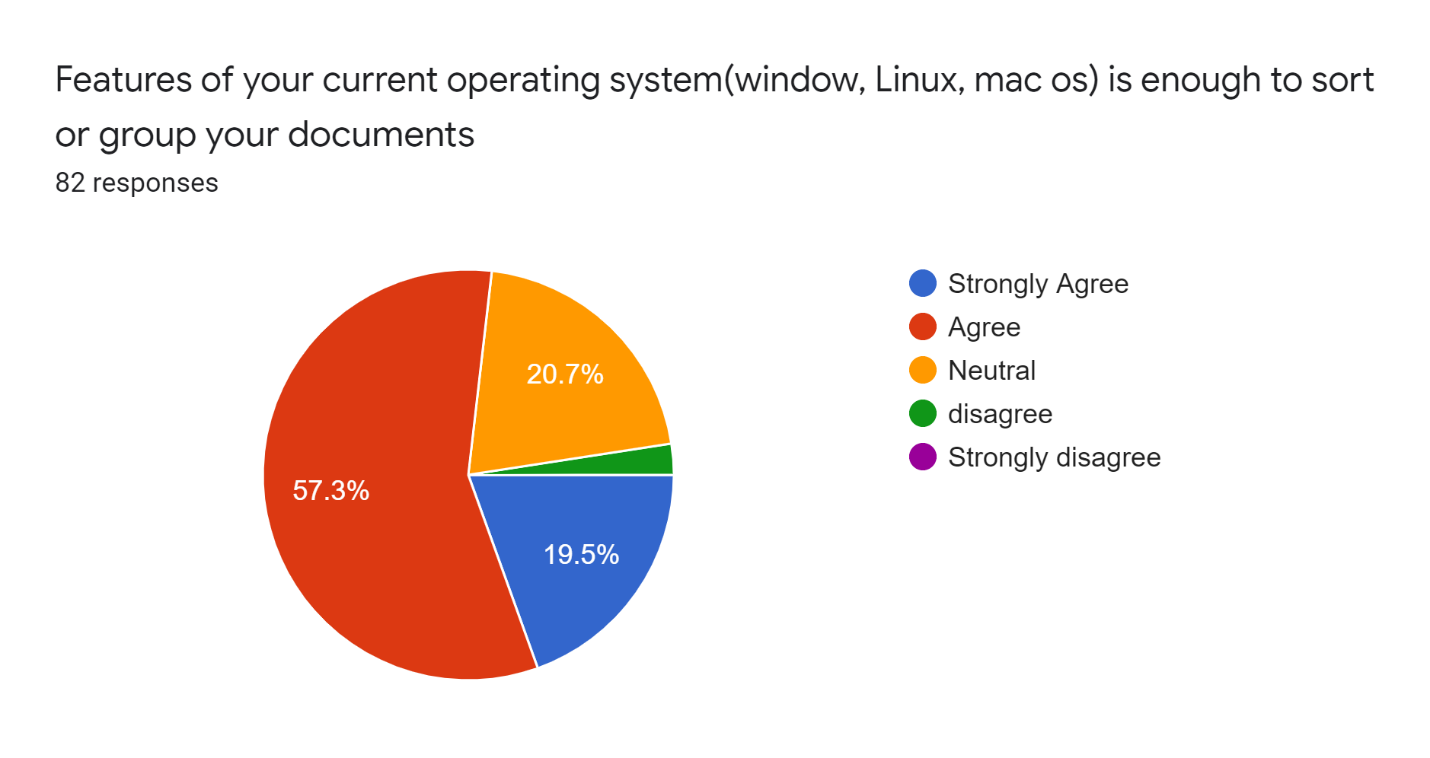
****

Since, primary aim of our system is to compare between big vs small documents. When two documents were compared, small represents abstract of the file. So, the word length of the abstract usually lies between 100 to 200 words. So, this question show that majority of the people deal with more than 100 to 200 words. Which mean that our system should be able to identify also big document as per need of general users.

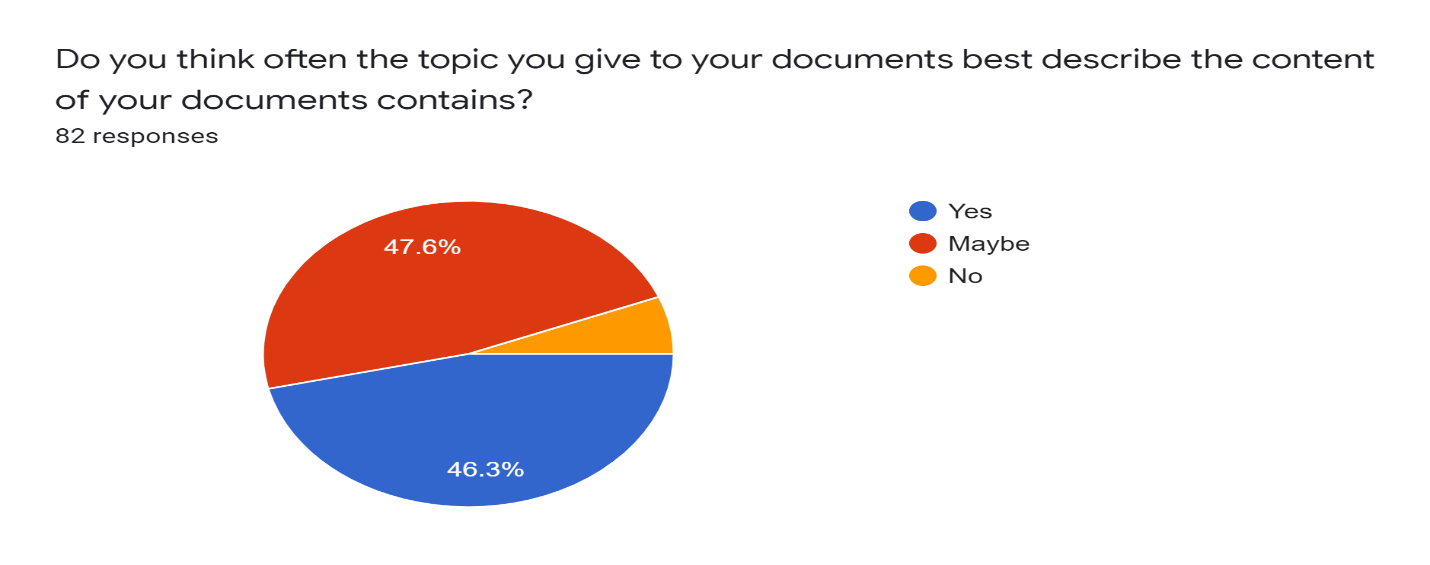
### C:\Users\codester\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\BE84AA91.tmp3.1.5 I often deal with unlabeled or random labelled (document without file name) document.

The main purpose of this question was to find out how many people often deal with random or unlabeled documents. As it is known that from our system, topics can be identified. So, system can be also useful for people which deals with unlabeled documents. Above from above response it can be seen that since sometimes can be seen as neutral. If result of rarely and Never is combined then it can be seen can see that 22% of the participants never deal with unlabeled documents. But majority about 34.2% of the people deal with unlabeled or random data.

### 3.1.6 Features of your current operating system(window, Linux, mac os) is enough to sort or group your documents.

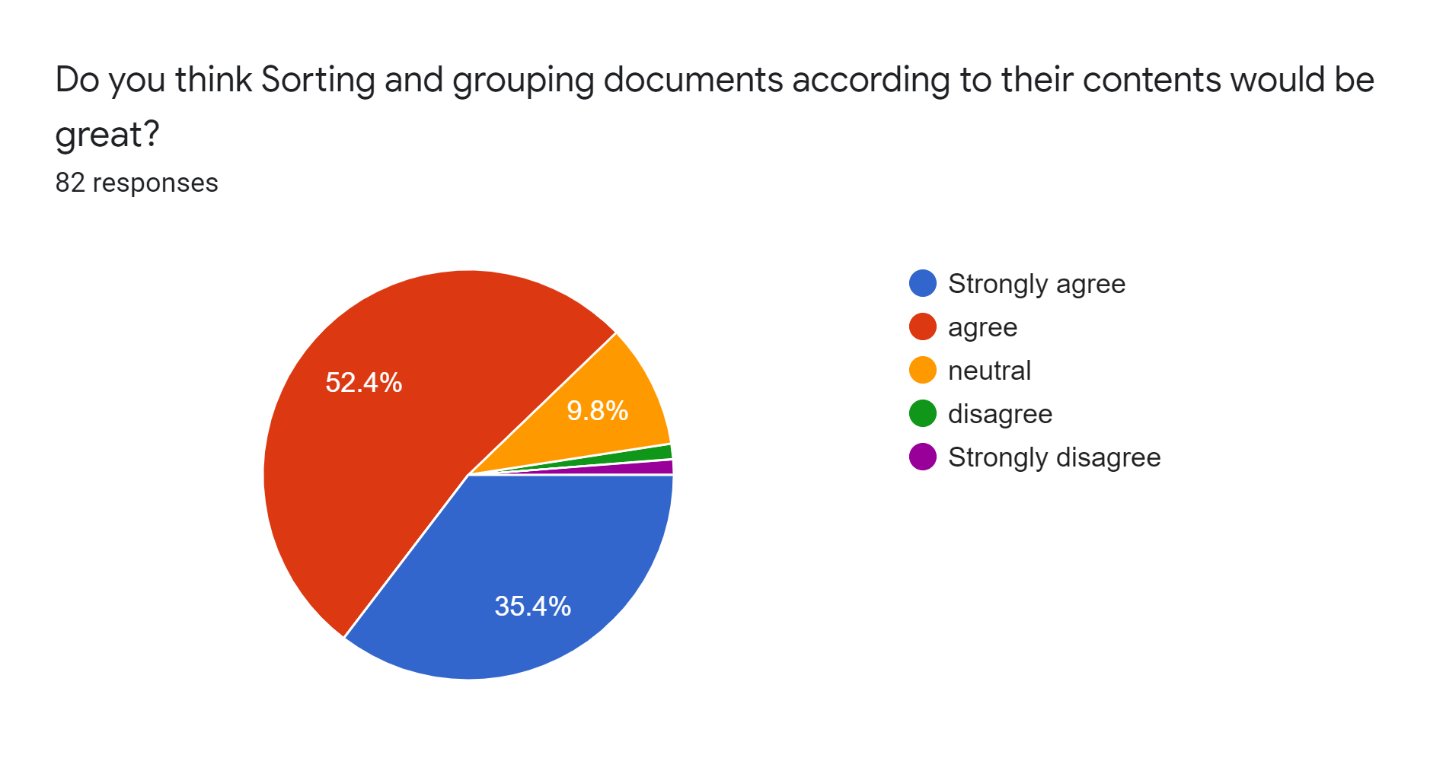
The main purpose of this question was to find out whether people are satisfied with the features which they are getting form their operating system. And from above it can be can see that Majority of people were agree on this. 57.3% of the people were agree and 20.7% of people were neutral. 19.5% of people were strongly agree. So, it can be clearly seen that people are quite satisfied with sorting methods provided by their Operating System.

### 3.1.7 Do you think often the topic you give to your documents best describe the content of your documents contains?



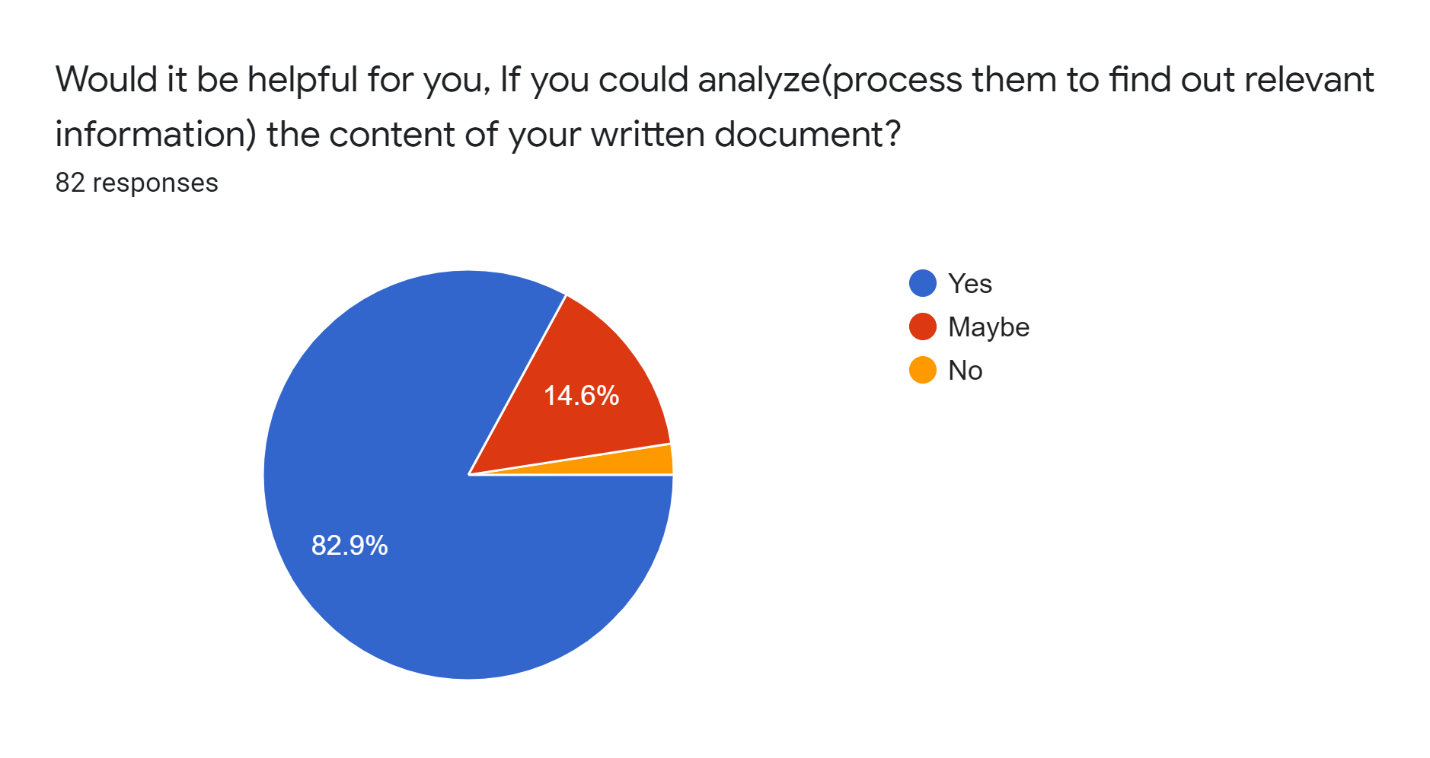
So, know that our system can identify topics. So, it can also help people who want to give topics to their contents. In above question it can be seen that 47.6% people were not sure about this and 46.3% think that it does and remaining said no. From above it can be see many people thinks their given topics were right but most people were confuse about their choice of topic.

### 3.1.8 Do you think Sorting and grouping documents according to their contents would be great?



By topic modeling it can be seen also sort document according to their content by category. So, to see whether this feature will be appreciated by the user or not. And according to the response it can be seen that 52.4% of people were agree that they think that sorting and grouping documents according to their contents would be great and 35.4% strongly agree and 9.8% were neutral. And only small portion of the participants were disgrace. From participants response it can be seen that feature could be good for targeted user.

**3.1.9 Would it be helpful for you, If you could analyze (process them to find out relevant information) the content of your written document?**

Since, my research was based on document clustering. So, it can be used for data mining. So, if my system will be able to analyze documents then it can be helpful. So, from response it can be seen that 82.8% of the people said yes, they think it would be helpful, if they could analyze their contents. And only even 14.6% said may be.

## Project Plan

Since this contextual report is research part and implementation of the system is still yet to do. So, in this section work done is going to be reviewed the and also create plan for the future. First in section the work that has been done and how time did it take to get work done will be reviewed. And after that plan for the future will be done. In the future plan what are going to do for implementing our system will be discussed.

* 1. **Project Review**

Our first responsibility was to come up with title for the final year project. And for that it has to be the area where the researcher’s interest and the place where research is wanting to bring change. After that our proposal was created. Then contextual report was stared. This report was only based on research. First Introduction of the research was started. Where the small overview of what is our system and how doesn’t it work was given. And in Literature review it has been seen how the work which has been already done in our area and existing system. And when it has been seen what has been done and what are being doing, it become easier for us to implement change and experiment with ideas. And Then data from market Research was analyzed. And after that in this section it will be recap of our system and created future. And in next section planning of our artefacts was done. Design of system, testing plan and as well was evaluation plan was done.

Below in Gannt chart it can be seen our work and active is chronological order.

And different date with done working is also given below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Task Name** | **Start Date** | **End Date** | **Duration** | **Status** |
| **Project proposal** | 14-Oct-19 | 29-Nov-19 | 45 |  |
| Title Selection | 14-Oct-19 | 23-Nov-19 | 20 | done |
| Proposal report assignment 1 | 15-Nov-19 | 29-Nov-19 | 14 | done |
| **Contextual Report** | 2-Dec-19 | 24-Jan-20 | 53 |  |
| Introduction | 4-Dec-19 | 17-Dec-19 | 13 | done |
| Literature review | 17-Dec-19 | 12-Jan-20 | 26 | done |
| Design and implementation | 11-Jan-20 | 15-Jan-20 | 4 | done |
| Market research | 27-Dec-19 | 18-Jan-20 | 22 | done |
| planning artefact | 17-Jan-20 | 19-Jan-20 | 2 | done |
| project plan | 20-Jan-20 | 21-Jan-20 | 1 | done |
| Conclusion | 22-Jan-20 | 23-Jan-20 | 1 | done |

* 1. **Future Plan**

Since, task which were already done were recap. Now in this section, what will be going are to done will be discussed. Since, Artificial Intelligence and clustering is new and vast. So, there are certain plan which has to be made for implementing. First thing is to learn more about python, machine learning, clustering in practice, mathematics. Since trained model will need web interface that’s why some expertise of backend and frontend of web technology is also needed. So, below there are certain task that should be done in order to implementing and build system.

* Learning machine learning and clustering
* Learn about python
* Learn about frontend backend technology to implement system in web
* Design system
* Complete requirement of the system
* Add some additional features

And my future plan with task and estimated time is given below in table and Gannt Chart.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Task Name** | **Start Date** | **End Date** | **Duration** | **Status** |
| **Project** | 3-Feb-20 | 10-May-20 | 97 | incomplete |
| Development | 3-Feb-20 | 10-May-20 | 97 | incomplete |
| System setup | 3-Feb-20 | 10-Feb-20 | 7 | incomplete |
| Resource gathering | 8-Feb-20 | 15-Feb-20 | 7 | incomplete |
| create model and learn | 15-Feb-20 | 4-Mar-20 | 18 | incomplete |
| Test And validate | 4-Mar-20 | 22-Mar-20 | 18 | incomplete |
| Deploy in backend | 26-Mar-20 | 9-Apr-20 | 14 | incomplete |
| Create web interface | 9-Apr-20 | 17-Apr-20 | 8 | incomplete |
| Integrete backend and frontend | 17-Apr-20 | 24-Apr-20 | 7 | incomplete |
| Testing | 24-Apr-20 | 30-Apr-20 | 6 | incomplete |
| Evulation | 30-Apr-20 | 7-May-20 | 7 | incomplete |

## Planning artefact

## Requirements

### Hardware Requirement

* Personal Computer

For development of this system, as a hardware requirement

### Software Requirement

* Jupyter Notebook
* Python
* Angular
* Vs code
* Git
* Flask
* Browser
* Spyder

For development purpose jupyter Notebook will be used to train model. And after training to development as well as implementation, python will be useful. Flask will be also need to implement system in web. And it will also need angular to build fronted of the system. And to keep trace as well for version controlling, gitlab will be used.

### Functional Requirement

This system is mainly focus on topic identification but there are also additional function it provide to utilize and make this system more useful.

* Allow user to topic identification
* Sort
* Add category to sort
* Rename existing unlabeled document with identified topics

### Non-Functional Requirement

There are also non-function requirements of my systems. These requirements determine how will our system works.

* Fast topic identification
* Scalable
* Efficiency for small documents

## System Design

To implement our system there are certain protype or blueprint which have to be followed. Since, perfect system cannot be created without planning. For planning perfect system, it is important to have clear maps and blueprints of the systems. UML diagrams such as Use case, Activity and sequence, help us to understand our system. They also help to collaborate and document. Since, after implementation change in the system could be really changing. So, low-fi, mid-fi and hi-fi prototype will be used for prototyping.

### Use case Diagram

Use case diagram shows how user can interact with the system and what type of action can be performed on system. From above figure it can be seen can see that user can do things such as selecting file, View file, detect file title, Rename file and also sort file.

### A screenshot of a cell phone Description automatically generatedActivity Diagram

It is known that activity diagrams represent a series of actions or flow of control in the system. And from above it can be seen that users can select file and they can perform action such as detecting the title, they can also sort document in provided clusters and they can also rename file.

### A screenshot of a cell phone Description automatically generatedSequence diagram

Sequence diagram simply represent the interaction between objects in sequential. And it can be also seen that this system also has some order. So, in sequence diagram it can be clearly seen that objects and sequence in which events are taking place in system. And these events take place between users. System and database.

### Low-fi

Blow are the low-fi of they systems. Show the basics home page and the dashboards.

A close up of a device

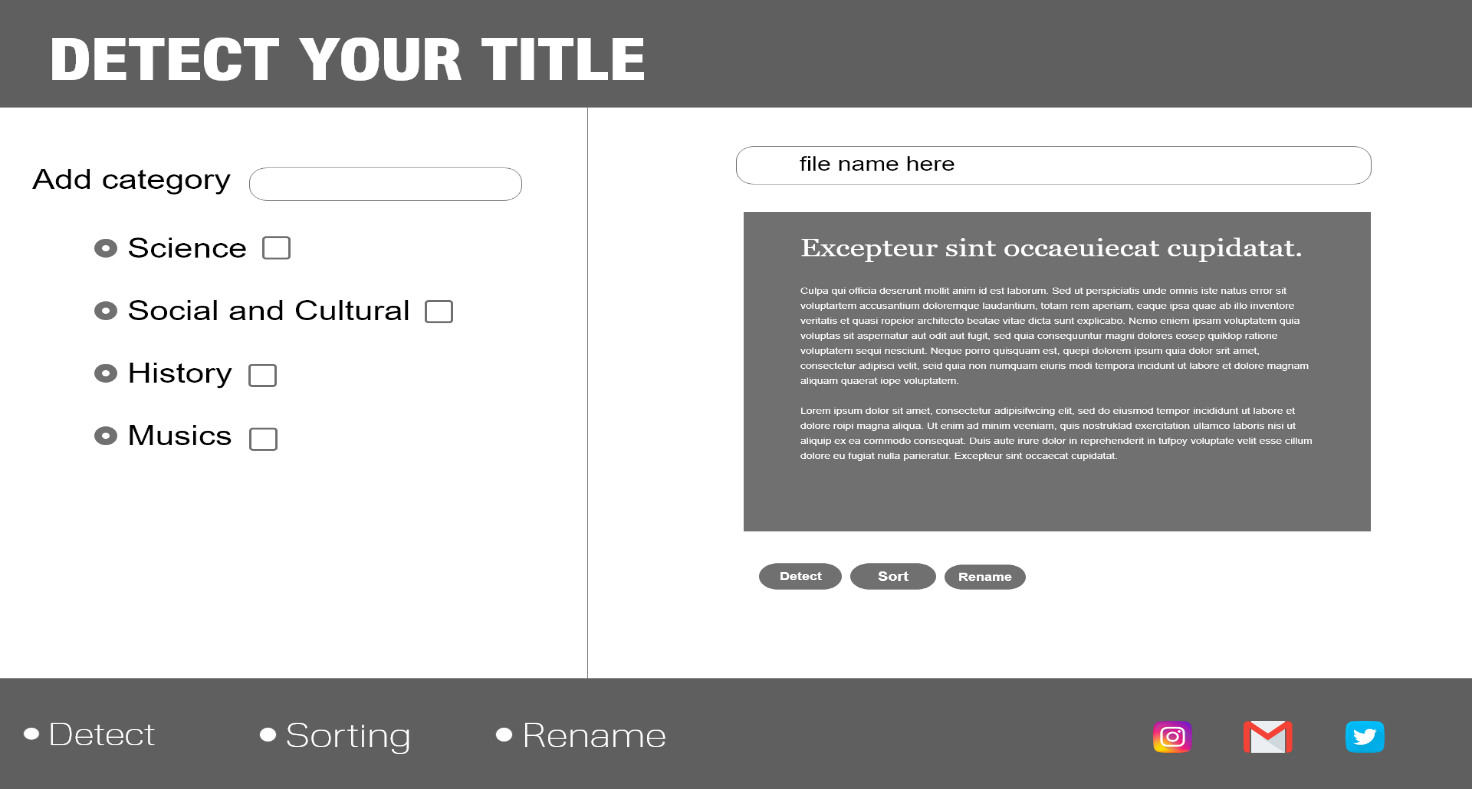
Description automatically generated

A close up of text on a white background

Description automatically generated

### A screenshot of a cell phone Description automatically generatedMid-fi

### 



### Testing Strategy

After the development, testing our software has to be done. And it is known what artificial intelligence model are learning and evolving. So, for the testing of system,

* **Black Box Testing**

In this testing, software will be testing only on the basic of input and the output value it provides. In this testing the tester should be lest concern with how code works but only does it give output what it intended to give.

* **Acceptance Testing**

In this testing, software is tested with some requirement and expectations. Since, our requirement is already clearing mention in use case, tester can test our system according to that requirements. This will help to system to meet its requirement.

* **Functional Testing**

In this testing, only testing what is related with functionality of the system will be tested. Tester only focuses on they are on our list of function. It’s doesn’t include any non-functional things such as accuracy, scalability and performance.

### Evaluation Strategy

Since, testing is not a final part of process. If even after removing all bugs, errors and even if system is working right but if it it’s not right for the users then they don’t care about how good your system is. There are certain aspects of the system which will only revel only after feedback form the general user. It will make sure that if system goes to market, its working will not be ours concern. So, for thing Evaluation Strategy will be implemented. After creating system our system will be evaluation by using Evaluation Strategy such DECIDE frame.

## Development

My deliverables include prediction model for text clustering to generate keyword as title, server to handle request for text clustering and client-side GUI. To achieve this I’ve mainly use python and it’s different libraries for clustering and server. For client-side implementation I’ve used Angular 8. And for test I’ve performed different kind of testing such as Unit testing, Integration testing, functional testing and acceptance test.

### Development

### Programming Language

Since, this project is mainly based on machine learning. So, python and R were main option for this course. Since, python has easy learning curve, python was best option. Building model was first part of the development. Implementing it for use was second part of the development. So, server and client-side development were necessary. And for server-side flask was used and for client-side JavaScript was used. So, python and JavaScript are the main language for this project.

* + 1. **IDE**

For this project there’s only one IDE that I’ve use and that is Spyder. Spyder is powerful scientific environment written in python, for python and designed by and for scientists, engineers and data analyst. With anaconda, it provides huge features and make it able to run machine learning programs. So, for clustering and training model, Spyder is used. Since, flask is also a framework of python, same IDE is used for Flask server side development.

* + 1. **Libraries**

For performing clustering, there were many libraries used which are given below.

* + 1. Sklearn
    2. Matplotlib
    3. Nltk
    4. Numpy
    5. Bs4
    6. Pandas
    7. Seaborn

There were also some libraries which were also used for server-side development. Some flask libraries which were used for development are given below.

* + - 1. Flask\_restful
      2. Flask\_jwt
      3. Docx
    1. **Frameworks**

There were three frameworks which are used in my website. And they are given below:

* + - 1. Flask- A micro web framework of python for server-side development.
      2. Angular 8- An open-source, client-side Typescript based JavaScript framework. It is used for building dynamic web applications.
      3. Bootstrap-A CSS Framework for developing responsive and mobile-first website.

These three frameworks helped to build application in structured way and help in rapid development.

* + 1. **Database**

All records of users and result are saved to database. So, MySQL is used as database for storing and maintain records.

* + 1. **Tools**

There are many tools used for the development of the project. These tools helped to develop and maintain code. These tools also helped to build system in much more simple and efficient way. Some of the used tools are mentioned below:

* + - 1. Vs code Editor
      2. Anaconda
      3. Jupyter Notebook
      4. Gitlab for Version Control and backup
      5. Xampp
      6. Chrome Browser

## Conclusion

Throughout this research we also review different part of the clustering and how we can perform clustering by different ways. So, after researching, critically analyzing and comparing different result, it can be said that K-means algorithm for clustering are superior while looking at the tradeoff for both small and big documents. First different algorithms were reviewed. Different algorithms such as hierarchical and variants of k-means were reviewed in this research. Every algorithm had their own advantages and disadvantages. But after reviewing their own metrics such as time complexity and accuracy, it was found that K-mean gave better performance with both low-dimensional as well as high-dimensional.

# References

Reddy, V. S., Kinnicutt, . P. & Lee, R., 2016. *Text Document Clustering: The Application of Cluster Analysis to Textual Document.* Las Vegas, NV, USA, IEEE, pp. 1174-1179.

Arora, N., Jain, S. & Verma, S. K., 2016. *Range clustering: An algorithm for empirical evaluation of classical clustering algorithms.* Noida, India, IEEE.

Copeland, B., 2019. *Artificial Intelligence.* [Online]   
Available at: https://www.britannica.com/technology/artificial-intelligence  
[Accessed 05 01 2020].

Deokar, S. T., 2013. Text Documents Clustering Using K means Algorthm. *International Journal of Technology and Engineering Science,* 1(4), pp. 282-286.

Gao, N., He, Y., Wang, H. & Sun, Q., 2014. *Topic Detection Based on Group Average Hierarchical Clustering.* Nanjing, China, IEEE, pp. 88-92.

Gupta, M. & Rajavat, A., 2015. Comaprision Of Algorithms For Document Clustering. *Sixth International Conference on Computational Intelligence and Communication Networks,* pp. 541-545.

Gurung, P. & Wagh, R., 2017. A Study on topic Identification using K means clustering algorithm: Big Vs. Small Documents. *Advances in Computational Science and Technology,* 10(2), pp. 221-223.

Kadhim, A. I., Cheah, Y.-N. & Ahamed, N. H., 2014. *Text Document Preprocessing and Dimension Reduction Techniques for Text.* Kota Kinabalu, Malaysia, IEEE.

Kothari, C. R. & Garg, G., 2014. *Research Methodology.* Third ed. New Delhi: New Age International Publishers.

Lewis, T., 2014. *A Brief History of Artificial Intelligence.* [Online]   
Available at: https://www.livescience.com/49007-history-of-artificial-intelligence.html  
[Accessed 06 1 2020].

N. & Kaur, j. P., 2015. *Cluster quality based performance evaluation of hierarchical clustering method.* Dehradun, India, IEEE.

Patel, K. A. & Thakral, P., 2016. *The Best Clustering Algorithms in Data Mining.* Melmaruvathur, India, IEEE.

Russel, S. & Norvig, P., 2014. The return of neural Networks(1986-present). In: *Artificial Intelligence A modern Approach.* Berkeley: Pearson, p. 1092.

Steinbach, M., Karypis, G. & Kumar, V., 2000. A Comparison of Document Clustering Techniques. *TextMining Workshop at KDD2000,* May.pp. 1-20.

Sun, X., 2014. *Textual Document Clustering Using Topic Models.* Beijing, IEEE.

Wunsch, D. . C. & Xu, R., 2008. *Clustering.* 1st ed. Piscataway, NJ: Wiley-IEEE Press .

Li, L., Yang, J., Xu, Y., Qin, Z. and Zhang, H., 2014. Documents clustering based on max-correntropy nonnegative matrix factorization. *Proceedings - International Conference on Machine Learning and Cybernetics*, 2, pp.850–855.