A

PROJECT REPORT

On

Text Clustering

Submitted by

Pratik Turkar

Prathamesh Dubey

Saumya Prasad

Under the guidance of
“Prof. Swati Narwane”

Department of Information Technology
Datta Meghe College of Engineering,
Sector-3, Airoli, Navi Mumbai – 400 708, (M.S.), INDIA

2021-2022

Datta Meghe College of Engineering

(AICTE & Govt. of Maharashtra Recognized, Affiliated to University of Mumbai)

Department of Information Technology

CERTIFICATE

This is to certify that Miss Saumya Prasad of Information Technology class **BE** Div. **B** Roll No. **37** Subject **AIDS II** Mini Project entitled "Text Clustering" has executed the Project work for the Semester 7 carried out by her under my guidance and supervision within the institute.

Signature of the Guide **Signature of Head of Department**

Examined on:

Examiner 1 **Examiner 2**

Acknowledgment

We have taken efforts in this report. However, it would not have been possible without the kind support and help of many organizations and individuals. We would like to extend our sincere thanks to all of them.

We are highly indebted to Prof. Swati Narwane for her guidance and constant supervision as well as for providing necessary information regarding the report and also for the support in completing it.

We would like to express our gratitude towards our parents for their kind co-operation which helped us in completing this report. Our thanks and appreciation also go to those people who have willingly helped us with their abilities.

Yours faithfully,

Pratik Turkar

Prathamesh Dubey

Saumya Prasad

TABLE OF CONTENTS

Abstract

Text clustering divides a set of texts into clusters (parts), so that texts within each cluster are similar in content. It may be used to uncover the structure and content of unknown text sets as well as to give new perspectives on familiar ones. The main contributions of this thesis are an investigation of text representation for Swedish and some extensions of the work on how to use text clustering as an exploration tool. We have also done some work on synonyms and evaluation of clustering results.

Text clustering, at least such as it is treated here, is performed using the vector space model, which is commonly used in information retrieval. This model represents texts by the words that appear in them and considers texts similar in content if they share many words. Languages differ in what is considered a word. We have investigated the impact of some of the characteristics of Swedish on text clustering.

Swedish has more morphological variation than for instance English. We show that it is beneficial to use the lemma form of words rather than the word forms. Swedish has a rich production of solid compounds. Most of the constituents of these are used on their own as words and in several different compounds. In fact, Swedish solid compounds often correspond to phrases or open compounds in other languages. Our experiments show that it is beneficial to split solid compounds into their parts when building the representation.

CHAPTER 1
INTRODUCTION

Classifying means putting new, previously unseen objects into groups based on objects of which the group affiliation is already known, so called training data. This means we have something reliable to compare new objects to — when clustering, we start with a blank canvas: all objects are new! Because of that, we call classification a supervised method, clustering an unsupervised one.

In general, clustering **documents** can also be done by looking at each document in vector format. But documents rarely have contexts. You could imagine a book standing next to other books in a tidy shelf, but usually this is not what large collections of digital documents (so-called **corpora**) look like.

The fastest (and arguably most trivial) way to vectorize a document is to give each word in the dictionary its own vector dimension and then just count the occurrences for each word and each document. This way of looking at documents without considering the word order is called the **bag of words** approach. The Oxford English Dictionary contains over 300,000 main entries, not counting homographs. That's a lot of dimensions, and most of them will probably get the value zero (or how often do you read the words *lackadaisical*, *peristeronic* and *amatorculist*?).

CHAPTER 2
TITLE DEFINITION

Objects inside of a cluster should be as similar as possible. Objects in different clusters should be as dissimilar as possible. Classifying means putting new, previously unseen objects into groups based on objects of which the group affiliation is already known, so called training data. This means we have something reliable to compare new objects to — when clustering, we start with a blank canvas: all objects are new.

This also means that for classifying the correct number of groups is known, whereas in clustering there is no such number. Note that it is not just unknown — it simply *does not exist*. It is up to us to choose a suitable amount of clusters for our purpose. Many times, this means trying out a few and then choosing the one which delivered the best results.

In general, clustering documents can also be done by looking at each document in vector format. But documents rarely have contexts. You could imagine a book standing next to other books in a tidy shelf, but usually this is not what large collections of digital documents (so-called corpora) look like.

CHAPTER 3
REVIEW OF LITERATURE

FINDING FROM LITERATURE SURVEY

Sr.no	Title	Mai n Aim	Algorithm used for Text Clustering	Future Work
1.	A Review on Twitter Sentiment Analysis Approaches	Various approaches for Sentiment Analysis	Selecting the best approach for our project	Machine Learning Approach
2.	Searching Research Papers Using Clustering and Text Mining	To optimize information and fast searching	Self-Organized Mapping Algorithm	To implement Automatic learning and extend search engine to search to generate portable search engine
3.	Extraction of News Content for Text Mining Based on Edit Distance	Extract news from twin pages	Proposed Algorithm based on Edit Distance	To refine the current text clustering system by including more functions.

PROPOSED SOLUTION

Algorithm -

We will use a dataset provided by sklearn to have a replicable corpus. After that, we will use the K- Means algorithm to group the vectors generated by the TF-IDF. We will then use Principal Component Analysis to visualize our groups and bring out common or unusual characteristics of the texts present in our corpus.

Here is what we'll do -

- Import the dataset
- Apply pre-processing to our corpus to remove words and symbols which, when converted into numerical format, do not add value to our model
- Use TF-IDF as a vectorization algorithm
- Apply K-Means to group our data
- Apply PCA to reduce the dimensionality of our vectors to 2 for visualization purposes
- Interpret the data

.

CONCLUSION

The interpretation is quite simple: there are no particular anomalies, except for the fact that there are texts belonging to the technology category that overlap slightly with those belonging to sport, between the dark blue and bright green border. This is due to the presence of common terms among some of these texts which, when vectorized, obtain equal values for some dimensions.

REFERENCES :-

1. https://ieeexplore.ieee.org/document/8376299
2. https://www.semanticscholar.org/paper/A-Literature-Survey-on-Text-Document-Clustering-and-Svadas-Jha/65f4470a3b3685ff422b1a5e396731dd0921003d
3. https://ieeexplore.ieee.org/document/9396915
4. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-