**A MINI PROJECT REPORT**

**ON**

**“NEWS HEADLINES CLUSTERING”**

**GROUP NO-12**

Submitted to

SAVITRIBAI PHULE PUNE UNIVERSITY

in completion of

**LABORATORY PRACTICE II**

**(B.E Computer Engineering)**

**BY**

Richa V. Patel Roll No : 405105

Pradnya V. Chaudhari Roll No : 405119

Jyostna M. Sonawane Roll No : 405135

Pooja V. Patil Roll No : 405178



Department of Computer Engineering

Sinhgad College of Engineering, Pune-41

**Accredited by NAAC with grade ‘A’**

**YEAR 2019-20**

**CERTIFICATE**

Sinhgad Technical Education Society,

Department of Computer Engineering

Sinhgad College of Engineering, Pune-41

**Accredited by NAAC with grade ‘A’**



**“NEWS HEADLINE CLUSTERING”**

Submitted to

SAVITRIBAI PHULE PUNE UNIVERSITY

in completion of

**LABORATORY PRACTICE II**

**(B.E Computer Engineering)**

BY

Richa V. Patel Roll No : 405105

Pradnya V. Chaudhari Roll No : 405119

Jyostna M. Sonawane Roll No : 405135

Pooja V. Patil Roll No : 405178

Prof. T. P. Vaidya Prof. M. P. Wankhade

Guide Head of Dept.

Department of Computer Engineering Department of Computer Engineering

Dr. S.D. Lokhande

Principal

SCOE, Pune

**ACKNOWLEDGEMENT**

We would like to express our deepest appreciation to all those who provided us the possibility to complete this project on “News Headline Clustering”.

We express our sincere thanks to Prof. T. P. Vaidya for her enthusiastic guidance and help in successful completion of the project. We are profoundly grateful to Dr. S. D. Lokhande, Principal, Sinhgad College of Engineering, Pune, Prof. M. P. Wankhade, Head of Department of Computer Engineering. We are thankful to all the staff members of Computer Department for their help and encouragement.

Lastly, extend my thanks to all those, who helped us directly or indirectly in completing this project.

Richa V. Patel

Pradnya V. Chaudhari

Jyostna M. Sonawane

Pooja V. Patil

**ABSTRACT**

One of the most successful application of Data Mining, Text Clustering, has received significant recognition, especially during past few years. Text Mining and lustering has immerged as an attractive solution to address many contemporary needs for extracting important information from raw data. This report describes the news headline clustering project undertaken. It uses K-Means algorithm and various preprocessing techniques in python.

The features of this project are:

1. Grouping similar types of News together.
2. Visualization of Clusters.

**CONTENTS**

1. **INTRODUCTION 1**
   1. Background and Basics 1
   2. Problem Definition 1
2. **SYSTEM REQUIREMENT SPECIFICATION 3**
   1. System Overview 3
   2. Functional Requirements 3
   3. Non – Functional Requirements 9
   4. Deployment Environment 4
3. **ANALYSIS AND DESIGN 5** 
   1. IDEA Matrix 5
   2. Activity Diagram 6
   3. Use Case Diagram 7
4. **METHODOLOGY AND IMPLEMENTATION 8** 
   1. Methodology 8
   2. Algorithm 10
   3. Visualization of Results 11
   4. Graphical User Interface 11
5. **TESTING 13**

5.1 Introduction 13

5.2 Automation Testing 13

5.3 Unit Testing 14

1. **CONCLUSION AND FUTURE SCOPE 15**

6.1 Conclusion 15

6.2 Future Work 16

1. **REFERENCES 16**

**LIST OF FIGURES**

|  |  |
| --- | --- |
| **Fig. No.** | **Figure Name** |
| 3.1 | News Headline Clustering Activity Diagram |
| 3.2 | News Headline Clustering Use Case |
| 4.1 | News Headline Home Page |
| 4.2 | News Headline Cluster Table |
| 4.3 | News Headline Cluster Visualization |
| 5.1 | Test Case Result in Excel Sheet |
| 5.2 | Selenium Testing |
| 5.3 | Unit Testing |

**LIST OF TABLES**

|  |  |
| --- | --- |
| **Table No.** | **Table Name** |
| 3.1 | IDEA Matrix |
|  |  |

**ABBREVIATION**

|  |  |
| --- | --- |
| **Sr. No.** | **Abbreviation** |
| 1 | TF-IDF – Term Frequency Inverse Document Frequency |
| 2 | API – Application Programming Interface |



1. **INTRODUCTION**

**1.1 BACKGROUND AND BASICS**

The increasing scope of the web and the large amount of electronic data piling up throughout the web has provoked the exploration of hidden information from their text content.

News articles published on different news portals throughout the web are the sources of the information. These can also be very good topics for the research on text mining. Clustering of similar news headlines and putting them under a single platform can be a very efficient option to the exploration of the same news article across multiple different news portals, which is a tedious and time-consuming task.

Clustering of the news headlines across the different news portals, uses document pre-processing techniques and creates groups of news headlines that are of similar category and represents them using visual representation techniques like graphs.

**1.2 PROBLEM STATEMENT**

**1.2.1 Problem Statement:**

Various news portals post various news articles every day. The project aims to group these articles under various categories and represent it using clusters. The goal is to analyze raw data of news headlines, process it and determine various categories of news posted over a period of time.

**1.2.2 Scope Statement:**

The scope of the project is to find similarity between various news headlines and group them accordingly. Clustering news which was obtained over a certain period of time provides researches and a broader view and understanding of type of incidents which majorly dominated the year, ex: 2019 was all about the Cricket World Cup and Elections.

* + 1. **Objective:**

The objectives of this project are:

* To provide a single platform to place the clusters of similar news headlines.
* To have a visual of the clusters through graphs for easy understanding.

1. **SYSTEM REQUIREMENT SPECIFICATION**
   1. **SYSTEM OVERVIEW**

The project is a web based text mining and clustering application which works on the dataset of news headlines focusing on Indian news including national, city level and entertainment obtained from start 2001 to end 2018,by Times of India.

The resulting clusters will be displayed on the web page along with the headlines which are grouped in a particular cluster.

* 1. **FUNCTIONAL REQUIREMENTS**

Proper preprocessed data is a very important factor in data clustering. For the clusters to be formed with maximum accuracy data preprocessing is a prerequisite. The basic functions which the project executes are:

* Preprocessing (Tokenization, stop word removal, vectorizing).
* Formation of clusters using k-means algorithm.
* Plotting the clusters.

The system allows the user to upload a dataset of news headlines to determine the categories of news present in it.

Input: News headlines (text file)

Output: Similar types of news headlines grouped into clusters and displayed by table and graph.

* 1. **NON-FUNCTIONAL REQUIREMENTS**
     1. **Interface:**

The project constructed as a web based system. A web interface is designed for the user so that he can upload a dataset, which will be linked to further web pages to display the clusters of the news headlines.

* + 1. **Performance**

The developed system must be able to preprocess the data and group the given headlines into clusters based on the K means algorithm.

* + 1. **Scalability**

The system must provide as many options for visualizing the cluster such as in tabular format as well as a graph.

* 1. **DEPLOYMENT ENVIRONMENT**
     1. **Hardware Requirements:**

The web application has no extra hardware requirement.

* + 1. **Software Requirements:**
* Operating System: Windows 10
* Web Browser: Google Chrome
* Python 3.6 or above
* Flask Library

1. **ANALYSIS AND DESIGN**
   1. **IDEA MATRIX**

IDEA Matrix is a filter and a qualitative decision-making tool. It is used to evaluate and prioritize your project and can be used to illuminate the big issues that need to be addressed by more research and refinement. The letters in the word idea are the initials of the components analyzed regarding the project. It is represented in a tabular form. Along with the component name, its description is written beside it in another column.

|  |  |  |
| --- | --- | --- |
| **IDEA** | **Deliverables** | **Parameters** **Affected** |
| **I** |  |  |
| Innovate | Form Clusters of News Headlines | News Headlines |
| Improve | Clustering technique | Clusters formed |
| Increase | User Understanding of various categories of News Headlines | End User |
| **D** |  |  |
| Drive | Extract useful information from raw dataset of news headlines | Dataset |
| Deliver | Web Based Platform for News Headline Clustering with easy GUI | News Headline Clusters |
| Decrease | Reduce user time to manually segregate News Headlines | User time |
| **E** |  |  |
| Educate | Study and implementation of K-Means Algorithm and Stop words Vectorizer | Project Members |
| Evaluate | Similarity between news headlines | Dataset |
| Eliminate | Need to manually to segregate New Headlines | User convenience |
| **A** |  |  |
| Avoid | Large Dataset | Dataset |
| Associate | -News headlines with its category (Sports, travel, politics etc.) | -Output data (Clusters)  -Number of clusters |
| Accessibility | -Simple GUI  -Accessible to all Users | - User Interface |

Table 3.1: IDEA Matrix

* 1. **ACTIVITY DIAGRAM**

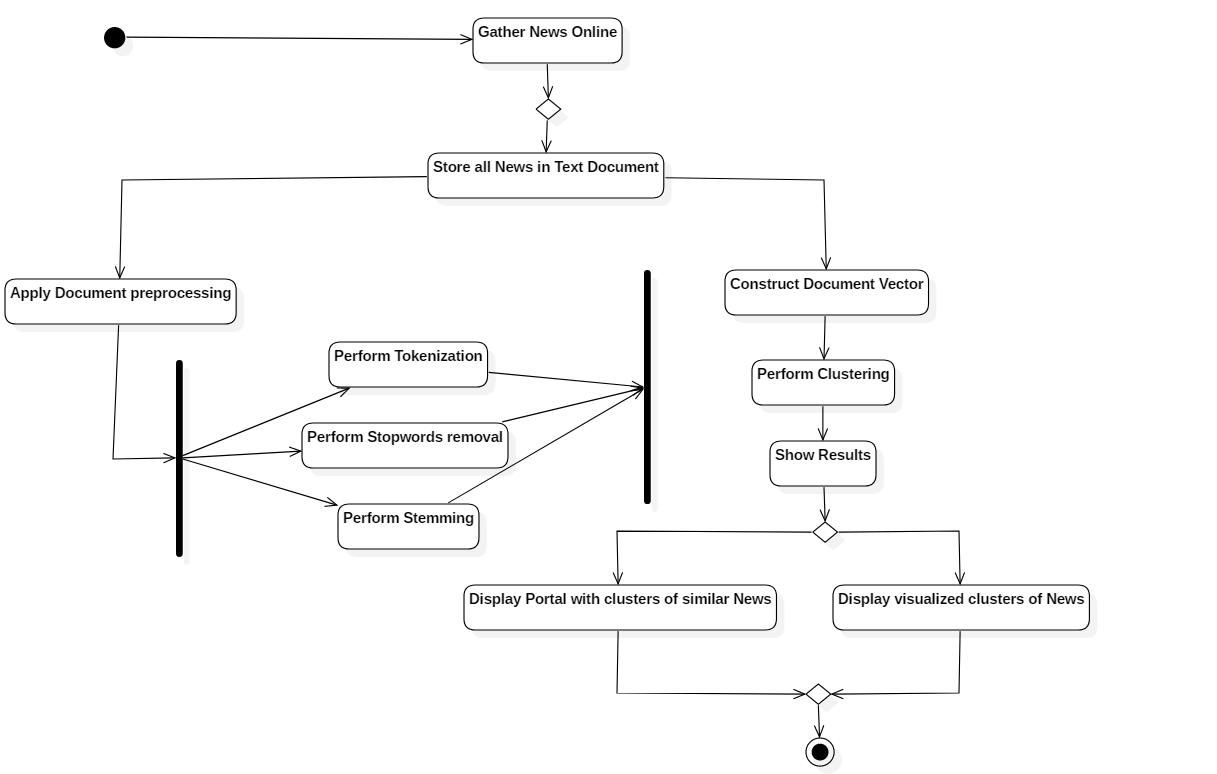


Figure 3.1: News Headline Clustering Activity Diagram

* 1. **USECASE DIAGRAM**

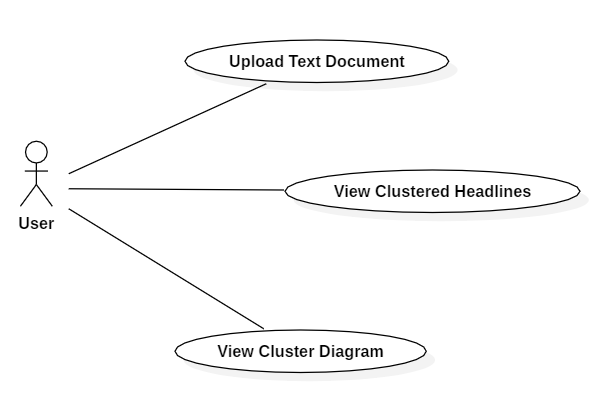
****

Figure 3.2: News Headline Clustering Use Case Diagram

1. **METHODOLOGY AND IMPLEMENTATION** 
   1. **METHODOLOGY**

The overall process used for the project is described. The process includes step-by-step actions performed for the clustering of the news headlines from news portals. The overall process is illustrated.

* + 1. **TOKENIZATION:**

Tokenization in text mining is the process of chopping up a given stream of text or character sequence into words, phrases, symbols, or other meaningful elements called tokens which are grouped together as a semantic unit and used as input for further processing such as parsing or text mining. It is used as a form of text segmentation in Natural Language processing.

* All alphabetic characters in the strings in close proximity are part of one token; likewise with numbers.
* Whitespace characters like space or line break or punctuation characters separate the tokens.
* The resulting list of tokens may or may not contain punctuation and whitespace.
  + 1. **STOP WORDS REMOVAL:**

Stop words are words which are filtered out before processing of natural language data. The general strategy for determining a “stop list” is to sort the terms by collection frequency and then to make the most frequently used terms, as a stop list, the members of which are discarded during indexing. Some of the examples of stop-word are: a, an, the, and, are, as, at, be, for, from, has, he, in, is, it, its, of, on, that, the, to, was, were, will, with etc.

* + 1. **STEMMING:-**

The idea of stemming is a sort of normalizing method. Many variations of words carry the same meaning, other than when tense is involved. The reason why we stem is to shorten the lookup, and normalize sentences. A stemming algorithm is a process of linguistic normalization, in which the variant forms of a word are reduced to a common form, for example,

connection

connections

connective ---> connect

connected

connecting

* + 1. **DOCUMENT REPRESENTATION:**

Document preprocessing is a key process in the document processing and information retrieval systems. To extract the relevant information from the large collection of the documents, it is very important to transform the full text version of the documents to vector form. Such a transformed document describes the contents of the original documents based on the constituent terms called index terms. These terms are used in indexing, the relevant ranking of the keywords for optimized search results, information filtering and information retrieval. The vector space model, also called vector model, is the popular algebraic model to represent textual documents as vectors. Using the vector space model, documents are represented using the term frequency (tf), inverse document frequency (idf) or tf- idf weighting scheme.

* + 1. **CLUSTERING USING K-MEANS:**

After the construction of the document vector, the process of clustering is carried out. The K-means clustering algorithm is used to meet the purpose of this project. From the given set of n data, k different clusters; each cluster characterized with a unique centroid (mean) is partitioned using the K-means algorithm. The elements belonging to one cluster are close to the centroid of that particular cluster and dissimilar to the elements belonging to the other cluster. The letter “k” in the K-means algorithm refers to the number of groups we want to assign in the given dataset. If “n” objects have to be grouped into “k” clusters, k clusters centers have to be initialized. Each object is then assigned to its closest cluster center and the center of the cluster is updated until the state of no change in each cluster center is reached. From these centers, we can define a clustering by grouping objects according to which center each object is assigned to.

**4.2 ALGORITHM**

Input: k: the number of clusters,

Output: A set of k clusters.

Algorithm:

Step 1: Choose k numbers of clusters to be determined.

Step 2: Choose k centroids randomly as the initial centers of the clusters.

Step 3: Repeat

3.1: Assign each object to their closest cluster center using Euclidean distance.

3.2: Compute new cluster center by calculating mean points.

Step 4: Until

4.1: No change in cluster center OR

4.2: No object changes its clusters.

**4.3 VISUALIZATION OF RESULTS**

The results obtained from the entire process, which, in fact are the clusters of similar news headlines, are presented in the form of table and graph on a web page. The sample web page shows clusters of news headlines.

* 1. **GRAPHICAL USER-INTERFACE**

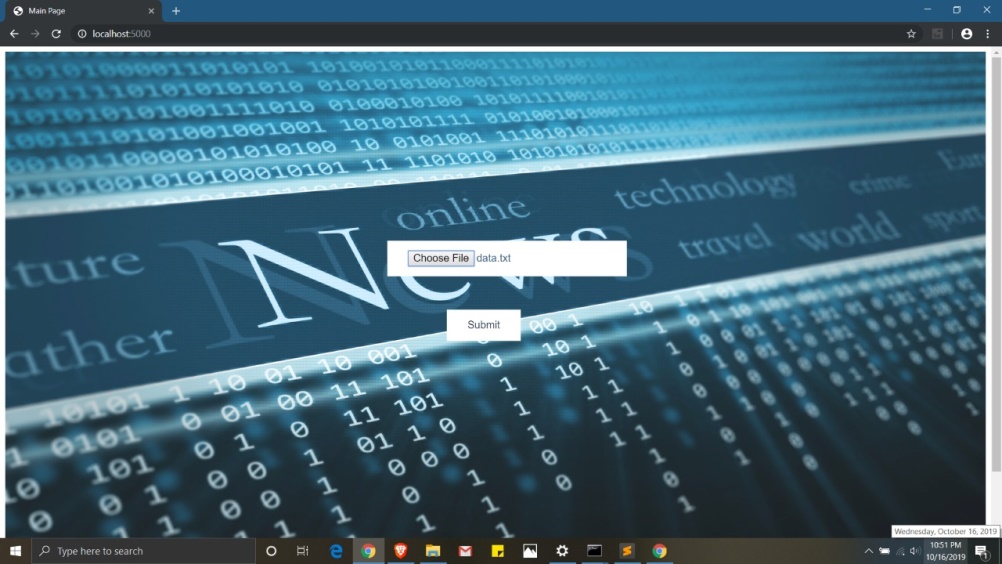
****

Figure 4.1: News Headline Home Page

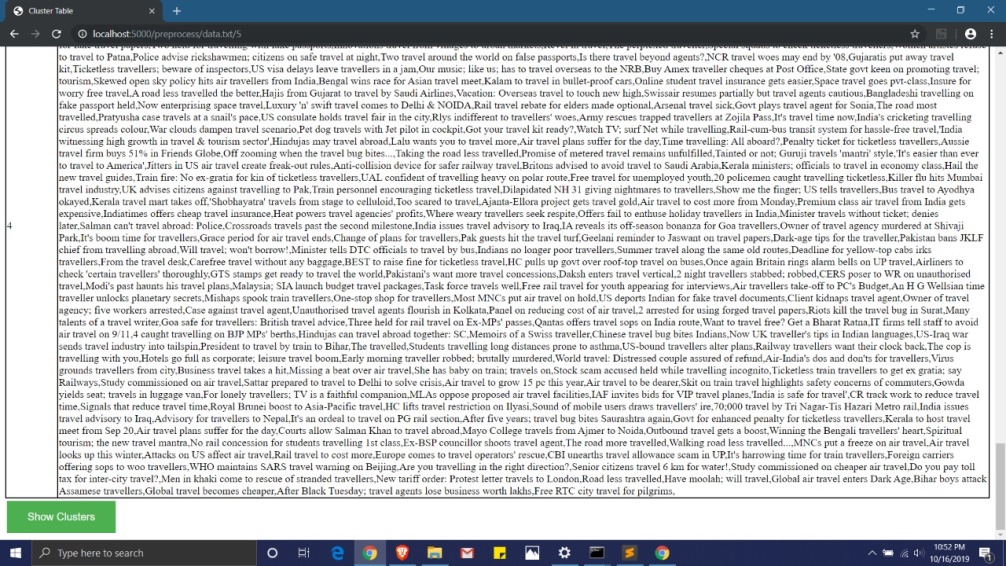


Figure 4.2: News Headline Cluster Table

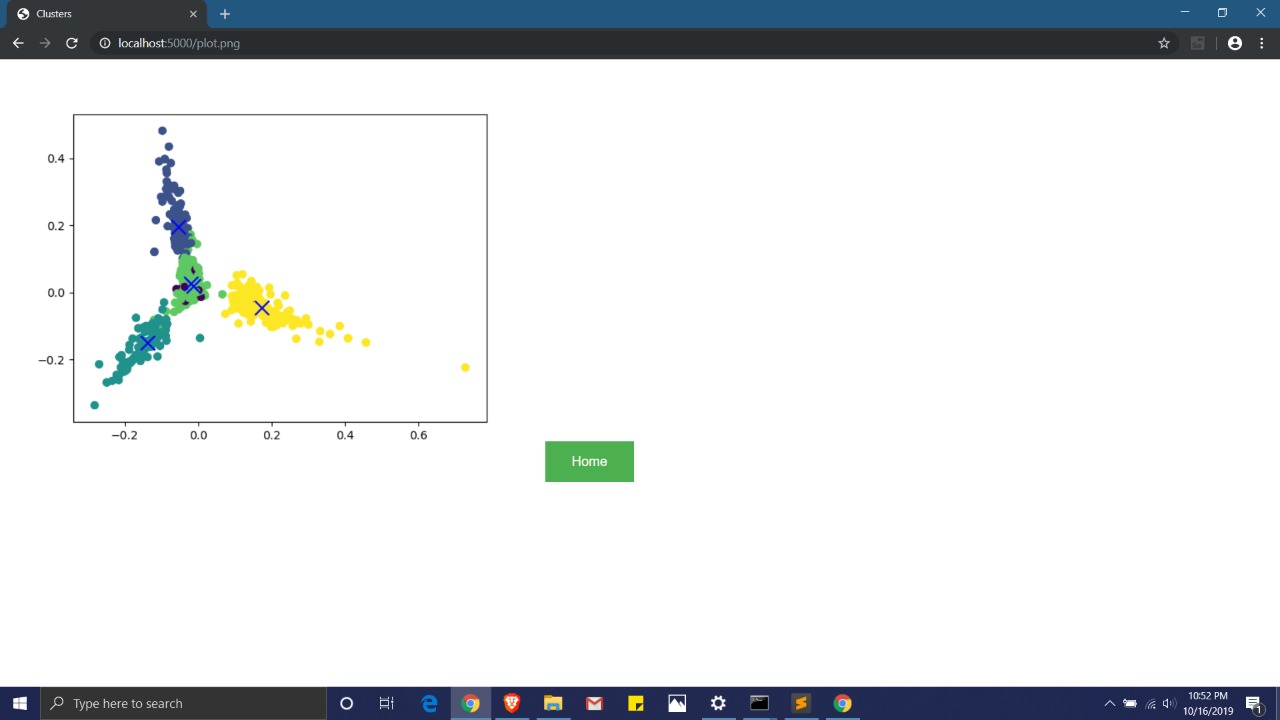


Figure 4.3: News Headline Cluster Visualization

1. **TESTING**

**5.1 INTRODUCTION**

This chapter covers various testing approaches used along with its test cases.

* Automation Testing
* Unit Testing
  1. **AUTOMATION TESTING**

Automation testing was done with selenium automation testing suite. Selenium is a set of different software tools each with a different approach to supporting test automation. The results of the tests being carried out were displayed on the console and also noted in excel sheet dynamically.

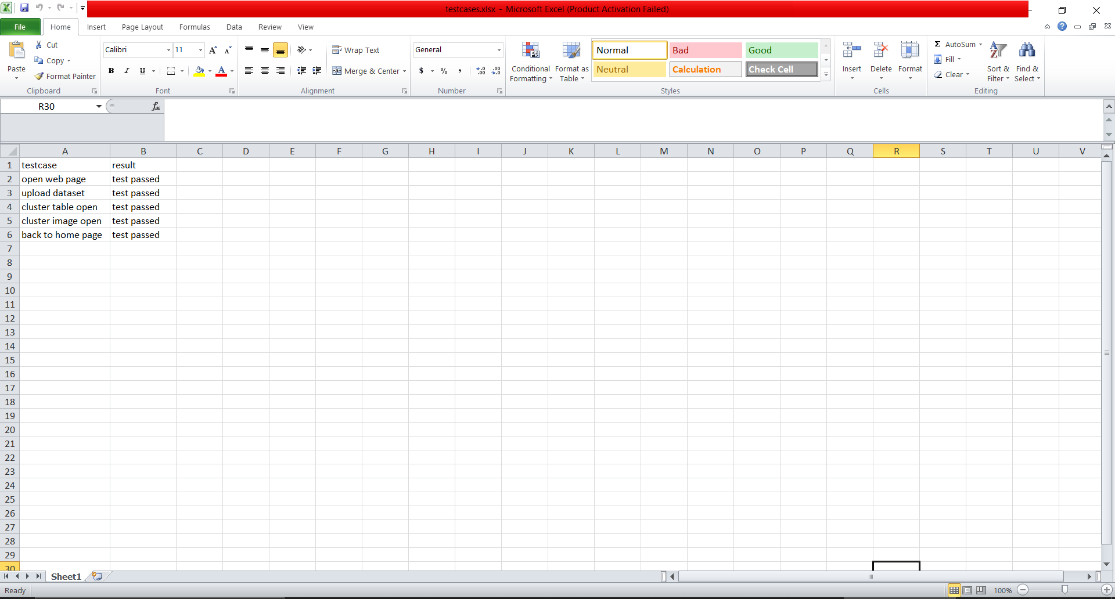
****

Figure 5.1: Test Case Result in Excel Sheet

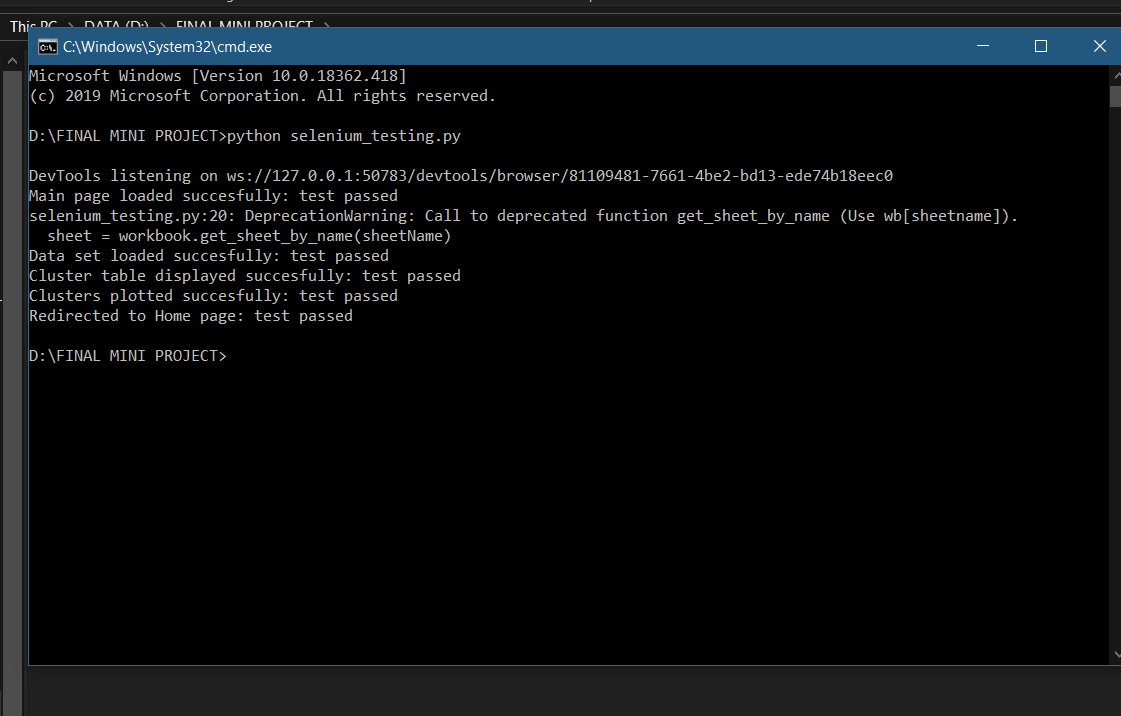


Figure 5.2: Selenium Testing

* 1. **UNIT TESTING**

Unit Testing is a level of software testing where individual units/ components of a software are tested. The purpose is to validate that each unit of the software performs as designed. Unit Testing allows developers to learn what functionality is provided by a unit and how to use it to gain a basic understanding of the unit API. Unit testing allows the programmer to refine code and make sure the module works properly. Unit testing enables to test parts of the project without waiting for others to be completed. Using the detailed design description important control paths are tested to establish errors within the boundaries of the module. In this system each sub module is tested individually

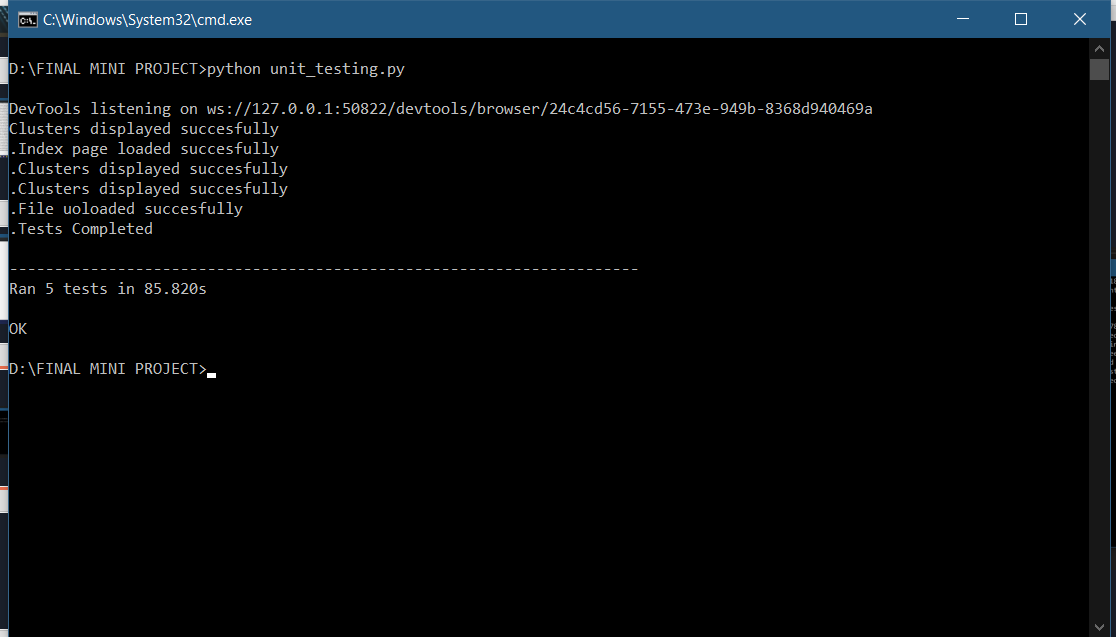


Figure 5.3: Unit Testing

|  |  |  |  |
| --- | --- | --- | --- |
| **Test to be conducted** | **Input** | **Expected Output** | **Test Pass/Fail** |
| Open Home page | Run main program | Open home page | Pass |
| Data Upload | File(.txt) | Accept file and upload it | Pass |
| Display Cluster Table | Data file | Group similar news and display it in table | Pass |
| Show Graph Of Clusters | Clustered news | Graph of Clusters | Pass |
| Back To Home | Click on ‘home’ button | Display home page | Pass |

|  |  |  |
| --- | --- | --- |
| ***Test to be conducted*** | ***Input*** | ***Expected Output*** |
| Flask Web App Test | Dataset(.txt) File | Should support HTML5 and CSS3 |
| GUI Testing | Onclick event on ‘Submit’, ’Show Graph’ and ‘home’ buttons | All fields on page |

1. **CONCLUSION AND FUTURE WORK:**
   1. **CONCLUSION:**

The project involved work on various areas of information retrieval and text mining and focused on the various methods for document pre-processing and document clustering.

Text mining and clustering techniques are really powerful. This project was completely based on these techniques. The system was created for finding the similarities among the news articles headlines. Various techniques were applied for preparing the corpus of a pre-processed document. Lastly, the k-means clustering algorithm was used for creating the clusters of similar news articles headlines. The similar news headlines were grouped into a single cluster. The real world application of the project would help people to find the similar news headlines on different news portal from a single platform. This would not have been possible without the use of text mining and clustering techniques. In general, it is not feasible to manually look for similar news in each of the portals and then compare each of them to find similarities between them.

* 1. **FUTURE WORK:**

After successful implementation of the model, enhancement can be done to cover the news articles over large domains. Web crawlers and parsers can be developed to extract information from the whole news portal sites. Text mining and clustering techniques can be used to create the clusters on the basis of contents and those contents present on the different news portals can be displayed under the single platform. That single platform can be any webpage, web application or any mobile application. In addition, this project can be extended to apply the methods proposed for web documents other than articles. The developed model can be enhanced to use in web content mining.

1. **REFERENCES:**

* **Book**

[1] <Bhaskar Chaudhary>, <Tkinter GUI Application Development Blueprints>, <Packt Publishing Ltd>, <2015>

* **World Wide Web**

# [1] <[Adam Geitgey](https://medium.com/@ageitgey?source=post_header_lockup)>, <Machine Learning is Fun! Part 4: Modern Face Recognition with Deep Learning>, <<https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78>>, < Jul 24, 2016>.

# [2] <Siddharth Modi> <Face recognition technology-BEST PPT>, < <https://www.slideshare.net/SiddharthModi1/face-recognition-technology-27574561>>, <Oct 25, 2013>