# **Cab Transaction Using Facial Recognition and Matching Engine**

B.E. Phase I project report submitted in partial fulfilment of the requirements of the degree of

Bachelor of Engineering

by

Kshitij Shukla EU2152053

Rohan Chavan EU1152082

Saniket Patil EU2152014

Under the guidance of

Dr. Kiran Bhandari

(Assistant Professor)



Department of Computer Engineering

St. John College of Engineering and Management

University of Mumbai

###### 2019-2020

**CERTIFICATE**

This is to certify that the project entitled **“Newsfeed Data mining and interactive insights extraction”** is a Project Phase I report of

**“Eric Thomas” (EU1142002)**

**“Pratik Penchka” (EU1142056)**

submitted in partial fulfilment of **“Bachelor of Engineering”** in **“Computer Engineering”** as laid down by University of Mumbai during the academic year 2019-2020.

Dr. Kiran Bhandari

Guide

Dr. Rahul Khokale Dr. G. V. Mulgund

Head of Department Principal

**Abstract**

With the rapid proliferation of the Internet and mobile devices, vast amounts of user-generated content have been accumulated through social network services, and massive amounts of news continues to be created and posted online. In this paper we try to find. We found the following differences: First, the news responds to official events but content on social network services is related to personal interests. Second, the news contains lots of redundant data which mostly not related to the news itself. Third, items from the news can be identified with a single key word in searches, whereas more keywords are needed to extract the desired information from social network services. At the same time, we found that the words mentioned in both the news and on social network services were similar, and both were used for commercial purposes. And while we search news on the search engine, they generate results randomly on its page and not according to the relevance or either how much the news is covered in that news site or forum. So, through our system we are going to classify the news on the categories like sports, international news, politics, entertainment etc. And show the most relevant news that the user wants to be in the news.

Project Report Approval for B. E.

This project synopsis entitled ***Newsfeed Data mining and interactive insights extraction*** by ***Eric Thomas, Pratik Penchka,*** is approved for the degree of ***Bachelor of Engineering*** in ***Computer Engineering*** from ***University of Mumbai***.

Examiners

1.---------------------------------------------

2.---------------------------------------------

Date:

Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

---------------------------------------

Signature

Eric Thomas (EU1142002)

---------------------------------------

Signature

Pratik Penchka (EU1142056)

Date:

**Table of Contents**

|  |  |  |
| --- | --- | --- |
|  | **Abstract** | **Iii** |
|  | **List of Figures** | **Vii** |
| **Chapter 1** | **Introduction** | **1** |
|  | 1.1 Motivation | 1 |
|  | 1.2 Problem Statement | 2 |
|  | 1.3 Objectives | 2 |
|  | 1.4 Scope | 2 |
| **Chapter 2** | **Review of Literature** | **3** |
|  | 2.1  The Methodology of Online News Analysis | 3 |
|  | 2.2 Characteristics Analysis of Data from News and Social Network Services | 5 |
|  | 2.3 Editorial Analytics: How media are developing audience data and metrics | 7 |
| **Chapter 3** | **Requirement Analysis** | **9** |
|  | 3.1 Functional Requirements | 9 |
|  | 3.2 Non-functional Requirements | 9 |
|  | 3.3 Minimum software/hardware Requirements | 10 |
|  | 3.3.1 Hardware Requirements |  |  |
|  | 3.3.2 Software Requirements |  |  |
| **Chapter 4** | **Design** | **11** |
|  | 4.1 DFD Diagrams of Newsfeed Data mining and interactive insights extraction. | 11 |
|  | 4.1.1 DFD Level 0 of Code Construction using Flowcharts |  |
|  |  |  |
|  |  |  |
|  | 4.2 UML Diagrams of Newsfeed Data mining and interactive insights extraction. | 14 |
| **Chapter 5** | **Report on Present Investigation** | **17** |
|  | 5.1 Select Language | 17 |
|  | 5.2 Input Data | 17 |
|  | 5.3 Classify Nodes | 17 |
|  | 5.4 Generate Output | 20 |
| **Chapter 6** | **Results and Discussions** | **21** |
|  | 6.1 Results | 21 |
|  | 6.2 Discussions | 25 |
| **Chapter 7** | **Conclusion** | **27** |
|  | **References** | 28 |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Figure Name** | **Page No.** |
| 4.1 | DFD LEVEL 0 for Newsfeed Data mining and interactive insights extraction | 12 |
| 4.2 | DFD LEVEL 1 for Newsfeed Data mining and interactive insights extraction | 12 |
| 4.3 | DFD LEVEL 2 for Newsfeed Data mining and interactive insights extraction | 13 |
| 4.4 | Use Case Diagram for Newsfeed Data mining and interactive insights extraction | 14 |
| 4.5 | Sequence Diagram for Newsfeed Data mining and interactive insights extraction | 15 |
| 4.6 | State Diagram for Newsfeed Data mining and interactive insights extraction | 16 |
| 6.1 | System Implementation 1 – Comparing redundant dataset | 22 |
| 6.2 | System Implementation 2 – tokenising the string | 23 |
| 6.3 | System Implementation 3 –Removal of redundant data and punctuation. | 24 |
| 6.4 | Generated Program | 25 |
| 6.5 | Program Execution | 26 |

###### 

**Chapter 1**

**Introduction**

This chapter discusses the problem posed, defines the topic and explains the aim and scope of the work presented in this project. The motivation of the project describes the driving force behind the project, while the problem statement describes the problem that the project aims to solve. Objectives are the individual targets that the project aims to work on, while the scope describes the upper limits of the intent of the project.

* 1. **Motivation**

The motivation for this proposed system is that there is a lot of news on the internet. A search for a particular keyword could possibly give thousands of search results. Now the reader would most probably select one out of the top 5 which may not have any significance to that particular news. Our approach to this problem is to filter out the articles with redundancy or irrelevant news in them with regards to the keyword that the reader has typed in. This would be done by using search feature in news feed API. Because many times the news is filled with redundant news and also after few lines the content of the news also changes that not any how matches the actual searched keyword. So, our system analyses the whole news article and gives the news article with less redundant data. Also, it searches for the keyword which is most used and combines both the methods and providing the user with the most relevant news. The system uses machine learning program in which a dataset is trained to analyze the whole news and give the best possible result.

**Chapter 2**

**Review of literature**

This chapter gives details about the studied literature papers based on the system statement. Three papers have been surveyed, each of which describe a system that is close in either functionality, concept or approach to the proposed system. The systems in each of these surveyed papers are explained, along with the sections that have been adopted by the system and the sections in which the system outperforms the system proposed by the system.

* 1. **The Methodology of Online News Analysis**

The authors, Helle Sjøvaag and Eirik Stavelin, demonstrates that methodologies designed for measuring broadcast news do not suffice in the ephemeral online news environment. They had divided their research on 4 problems that they face while doing collecting the data for sample or data set. So, the first problem was obtaining the full population of data and what exactly is the object we are studying? Then the issue is thus one of inclusion and exclusion. The second problem faced is with defining the unit of analysis; The structure of newspapers and broadcasts are sufficiently standardized and their syntax is sufficiently well understood that defining the nature of a message’s context is straight forward. In contrast, the nonlinear nature of the WWW causes problem because of boundaries and environment of messages and involves more complex semantics. The third and one of the most challenging problems is with designing an operable coding scheme news media do not only carry different characteristics across publication platforms, but they also share a great many features, particularly in terms of genre traits, communicative forms, journalistic standards, and overarching institutional practices. So, this causes a lot of complications while writing a code for the analysis. The fourth and the final problem according to this paper is establishing causes a lot of complications while writing

**Chapter 3**

**Requirement analysis**

This chapter gives details about various requirements of the proposed system. The different types of requirements used in the classification system consists of functional and non-functional requirements and hardware/software requirements. The requirements need to be satisfied in order to enable the user to work with the proposed system.

* 1. **Functional requirements**
* The system shall accept the keyword which the user has entered in the API.
* The system shall generate a search according to the keyword.
  1. **Non-functional requirements**
* The system shall be easy to use for even non-programmers.
* The system shall generate an error-free program, almost instantly.

**Chapter 4**

**Design**

This chapter specifies the structure of how a software system will be written and function, without actually writing the complete implementation. The design of the system is shown with the help of A. Data Flow Diagram (DFD) and B. Unified Modelling Language Diagram (UML). A data flow diagram (DFD) illustrates how data is processed by a system in terms of inputs and outputs. UML diagram consists of Use Case Diagram, which is used to identify the events responsible for state changes, Sequence Diagram, which describes interactions among classes in terms of an exchange of messages over time and State Transition Diagram, which describes the behaviour of a single object in response to a series of events in a system.

**4.1 Data Flow Diagram**

A data flow diagram (DFD) illustrates how data is processed by a system in terms of inputs and outputs. As its name indicates its focus is on the flow of information, where data comes from, where it goes and how it gets stored.

DFD Layers. Draw data flow diagrams can be made in several nested layers. A single process node on a high level diagram can be expanded to show a more detailed data flow diagram. Draw the context diagram first, followed by various layers of data flow diagrams.

**Chapter 5**

**Report on Present Investigation**

This chapter provides details regarding the complete process followed by the proposed system. All of the steps are explained in detail. This chapter describes the System Architecture followed by the system. This system architecture has been designed to perfectly depict the various processes and explain them in great detail. The following steps are followed while implementing the proposed system:

* 1. **Redundant Datasets:**

The proposed system has been given set of data in which there are list of redundant data which are mostly present or widely used in the news article which is of no use and can be eliminated to get the results.

* 1. **Tokenizing:**

And then we take a string and then we convert it into lowercase first and then we break it down into tokens.

* 1. **Removing the redundant words and punctuation:**

In this we remove the redundant words from the strings by referring the dataset in which all redundant letters are present andalso, we are removing the punctuation mark which is present in the string.

**Chapter 6**

**Results and Discussions**

This chapter describes the results obtained after the implementation of the proposed system. The current implementation is demonstrated in this chapter as well.

* 1. **Results**

The experimentations and results confirm that we are able to tokenize the words. With regard to the experimentation and the corresponding results, it can be concluded that the system implementation is successful in exploring the feasibility of constructing a prototype compiler to translate flowchart to code. Moreover, it confirms that the approach for the process of translating flowchart to code is a better alternative to the traditional, classical programming

**Chapter 7**

**Conclusion**

The Newsfeed Data mining and interactive insights extraction is implemented in Python which analyse the news articles and provide us the most optimum search result. First of all, user enters the keyword on which he/she wants the result. The data the analyses on the basis of the keyword that is searched. It uses the procedure that clear the redundant data (i.e. words) which are used in the news article which is not needed is eliminated and the rest of the data is tokenize and analyse. Then it started categorizing according the topic and then uses the program to get the most optimal result. Also, the program provide u the hot topic which is moving around the internet for that day. So, the uses get the most relevant and the topic related data that ‘he/she is looking for.

* 1. **Problem Statement**

To create and implement a system resolves this issue by analyzing the news feed searching for key words by training the ML for data set to remove redundancies from the news feed and give the reader a relevant and useful article top read.

* 1. **Objective**
* To perform analysis of news feed in current as well as during a specific period of time if needed.
* To extract relevant news which is consistent throughout the article
* To perform analysis of news feed in current as well as during a specific period of time if needed.
* To show the most relevant results according to the keyword which users want.
* To implement categorization of topics.
  1. **Scope**

The system is used to get the most relevant news from the news that are available on the internet, because many of the time the headlines and the content of the news doesn’t match. So, the system uses machine learning and train datasets to analyse the news. So, the user finds the actual news article according to the keyword searched.

a code for the analysis. The fourth and the final problem according to this paper is establishing intercoder reliability**.** Quantitative content analyses should always report validity as measured by inter-coder agreement. Reliability reporting increases the validity of the analysis and contributes to the further improvement of the method within media and communication studies. A premise for reliability measures is testing the extent to which two or more coders can be said to agree on the meaning, and hence coding, of latent content. News articles published on 10 preselected dates in 2009, abstracted from data set 1. That data set consists of a representative selection of text-based news items published across a constructed 10-day period.

The dates were preselected from the second week of each month. In three cases sampling had to be moved to the next week due to the lack of complete set of front-page snapshots for the date in question. As time restraints limited sampling to 10 days, May and July were excluded as the least representative months (May because of the national holiday, and July because the general summer holiday left fewer journalists on staff). Every week day was represented and the selection was predetermined – a Monday from January, a Tuesday from February, and so on. Because of lower publication frequency at weekends, the sample contains only one Saturday and one Sunday. Event-sensitive days such as Easter and Christmas were avoided. Data set 2 underwent manual coding and quantitative content analysis.

Items analyzed were limited to the text-based news articles published on nrk.no, and did not include redistribution of audiovisual content from broadcasting to the Internet. Audiovisual news content on nrk.no amounted primarily to radio and television productions ported to the website for streaming. As we were looking to analyze the news dissemination of NRK’s online news unit, we chose to overlook audiovisual content not primarily intended for the online platform and to focus NRK’s textually mediated online news.

The hereto relevant aims of the project were to assess the degree to which nrk.no presented a continuous and updated news agenda online during 2009; the thematic distribution of its news content; its front page priorities; the depth and perspectives of the news content; and the degree to which nrk.no used interactive or other Internet-specific tools in its news dissemination.

**2.2 Characteristics Analysis of Data from News and Social Network Services**

The authors,Beakcheol Jang and Jungwon Yoon, propose that the supplementary computer technologies have developed rapidly since the invention of the Internet. Due to this advancement has provided people with various new avenues for communication. For example, in the past, information or knowledge was transmitted through print newspapers or by word of mouth as traditional information-sharing methods, but it is now possible to quickly share vast amounts of information with others through the Internet.

With such mobile devices as smartphones and tablets, people create vast amounts of content every day on websites, blogs, and social network services (SNSs) such as Twitter and Facebook. They provide researchers with various items of inquiry, e.g., the structure of social networks, the dissemination of information therein predictions, and their influence on other resources. Newspapers, a traditional means of information, also upload their contents onto websites to make them available on the Internet. The news has a close relationship with knowledge, information, and opinions because it provides articles on topics that interest people. The news has been used in a variety of research, e.g., its prediction, influence and other features. Due to the importance of the news and SNS, many researchers have studied them extensively. Most claim that it is important to understand how the news in sequences SNS and vice versa but none of those has investigated the characteristics of the contents of each.

They collected data concerning the news using NAVER, which is a popular portal sites in Korea. So, they collected SNS data from Twitter. The data was thus gathered were too large to analyze; we thus reduced their size by specifying their topics to the disease. Fifteen keywords most closely related to influenza and collected news and SNS datasets based on this were used. This is because influenza is the most threatening disease to humans at present. They observed the following differences by analyzing data from the news and SNS. Gathering the data was challenging to find the same topic in the news and on SNSs. Therefore, the former responds to official events whereas the latter to personal interests. The news covers a specific topic continually, whereas the transition from one topic to another on SNSs is rapid.

**2.3 Editorial Analytics: How media are developing audience data and metrics**

The author, Chanchai Supaartagorn, proposes a system that accepts flowcharts as an input In the paper, Editorial Analytics: How media are developing audience data and metrics, the authors, Federica Cherubini and Rasmus Kleis Nielsen, propose that leading digital news organisations are developing distinct forms of editorial analytics tailored to help them pursue their particular goals. So, these forms of editorial analytics differ from more rudimentary and generic approaches in being aligned with the editorial priorities and organisational imperatives (whether commercial, non-profit, or public service) of specific news organisations, in informing both short-term day-to-day decisions.

They say best-practice editorial analytics are tailored to the priorities and goals of a given organisation as well as the context in which it competes. Instead, news organisations need to think about how they can develop their analytics capability by making sure they combine the right set of tools, an organisational structure that incorporates the expertise to use them, and a newsroom culture that embraces data-informed decision-making.

They have noticed that most sophisticated audience teams are keenly aware that analytics are not perfect. the data never tell the full story, and quantitative analysis always has to be supplemented by editorial expertise and other forms of qualitative judgement. So they have come to a conclusion that even the best editorial analytics continue to be constrained by the difficulties involved in defining and measuring many of the things that news organisations aim to achieve and is beset by a whole range of data-quality and data-access issues, exacerbated by rapid changes in the media environment.

Many journalists also want analytics, as an earlier period of skepticism seems to have given way to interest in how data and metrics can help newsrooms reach their target audiences and do better journalism. that is encouraging, because analytics and data metrics will continue to evolve, and if journalists are not part of that process, the tools and techniques developed will continue to reflect and empower commercial and technological priorities more than.

**Minimum software/hardware requirements**

* + 1. **Hardware requirements**
* Processor : Intel core i3 and higher
* Operating System : Windows 7,8,10/ Mac OS/ Linux
* RAM : 4GB and higher
* Hardware Devices : Keyboard and mouse
* Storage : At least 2 GB free on the drive.
* Display : Standard Output Display
  + 1. **Software requirements**
* Language Used : Python 3.6
* Tools : Anaconda
  + 1. **DFD Level 0 Newsfeed Data mining and interactive insights extraction**

Context Diagram: A context diagram is a top level (also known as "Level 0") data flow diagram.

It is the most condensed representation of the system functionalities.

****

Fig 4. 1 DFD level 0 of Newsfeed Data mining and interactive insights extraction

It only contains one process node ("Process 0") that generalizes the function of the entire system in relationship to external entities. It is given further detail in DFD Level 1 diagram.

* + 1. **DFD Level 1 Newsfeed Data mining and interactive insights extraction**

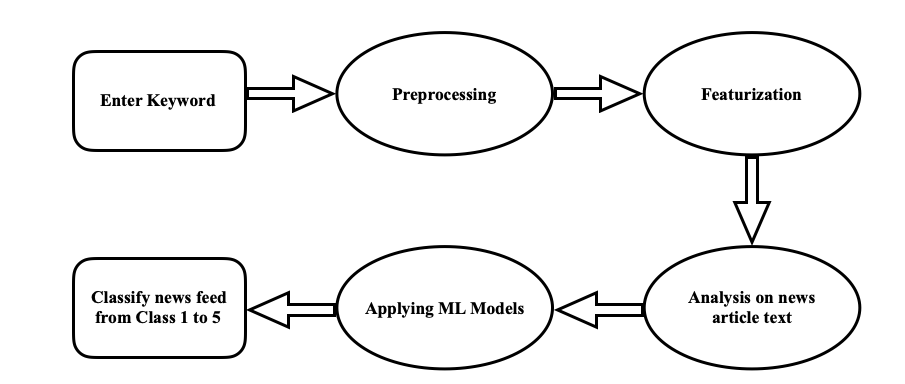


Fig 4. 2 DFD level 1 Newsfeed Data mining and interactive insights extraction

The first level DFD shows the main processes within the system. Each of these processes can be broken into further processes until you reach pseudo code. The processes include entering keyword, pre-processing stage includes tokenizing the words, analysing the news from the article text, applying the trained data set on the article and classifying them into the 5 classes. The output would be the articles that are most relevant with respect to the keyword.

The diagram represents the process of the system from input to the output. The process in the system is shown step by step in details.

* + 1. **DFD Level 2 Newsfeed Data mining and interactive insights extraction**

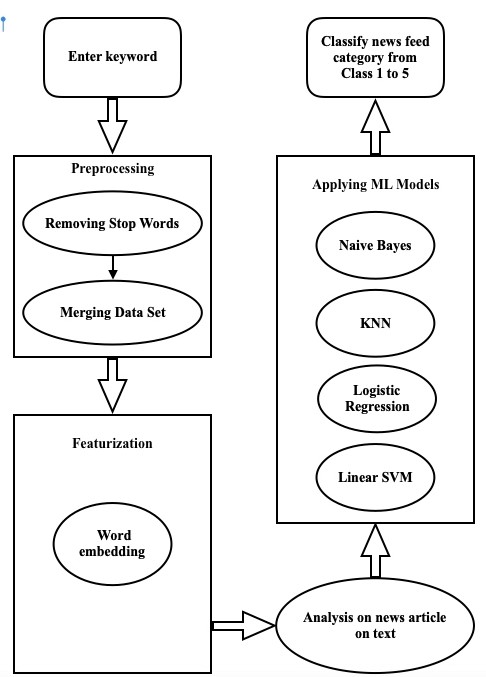
****

Fig 4. 3 DFD level 2 of Newsfeed Data mining and interactive insights extraction

The user gives input as flowchart and the node classification is done in several steps first is input/output node, second is procedure node, followed by conditional node, file handling and function call node. After that analysing of command is done. Syntactically right code is written in a separate file which reflects the user logic.

* 1. **UML Diagrams**

UML stands for Unified Modelling Language. UML 2.0 helped extend the original UML specification to cover a wider portion of software development efforts including agile practices.

* + 1. **Use Case Diagram**

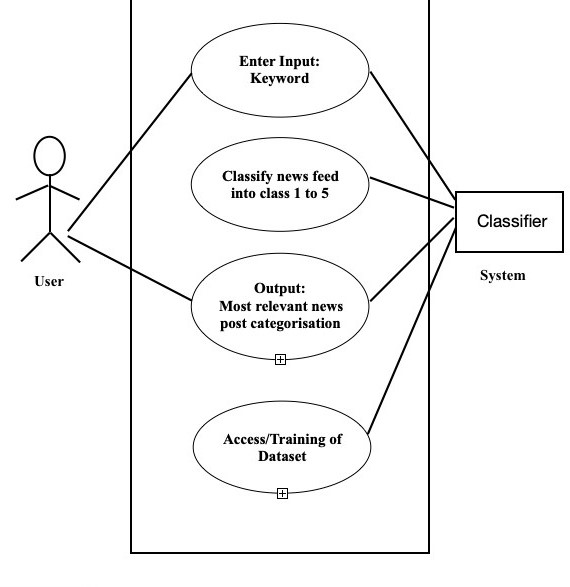
****

Fig 4. 4 Use Case diagram for Newsfeed Data mining and interactive insights extraction

A use case diagram is a dynamic or behavior diagram in UML. Use case diagrams model the functionality of a system using actors and use cases. Use cases are a set of actions, services, and functions that the system needs to perform. In this context, a "system" is something being developed or operated, such as a web site. The "actors" are people or entities operating under defined roles within the system.

* + 1. **Sequence Diagram**

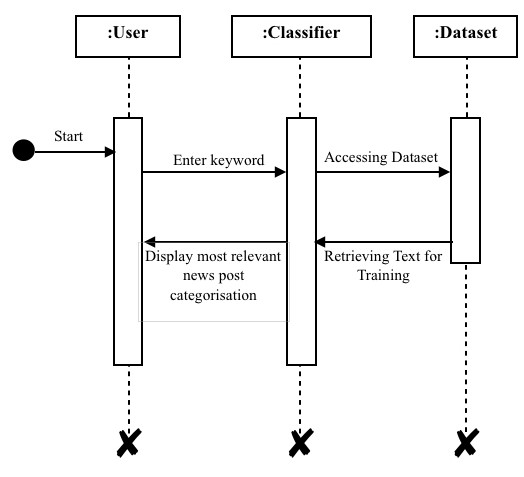
Sequence diagrams describe interactions among classes in terms of an exchange of messages over time. They're also called event diagrams. A sequence diagram is a good way to visualize and validate various runtime scenarios. These can help to predict how a system will behave and to discover responsibilities a class may need to have in the process of modelling a new system.

Fig 4. 5 Sequence Diagram for Newsfeed Data mining and interactive insights extraction

The sequence diagram here shows 3 components in the system. The interaction of these components, User, Classifier and Dataset has been described. The user gives an input, which has to be a flowchart. This keyword is searched in the article and each article, then it is passed into the dataset. The trained dataset then gives the most relevant news article post categorization.

A state diagram shows the behaviour of classes in response to external stimuli. Specifically, a state diagram describes the behaviour of a single object in response to a series of events in a system.

* + 1. **State Diagram**

Sometimes it's also known as a Harel state chart or a state machine diagram. This UML diagram models the dynamic flow of control from state to state of a particular object within a system. A state diagram shows the actual changes in state, not the processes or commands that created those changes.

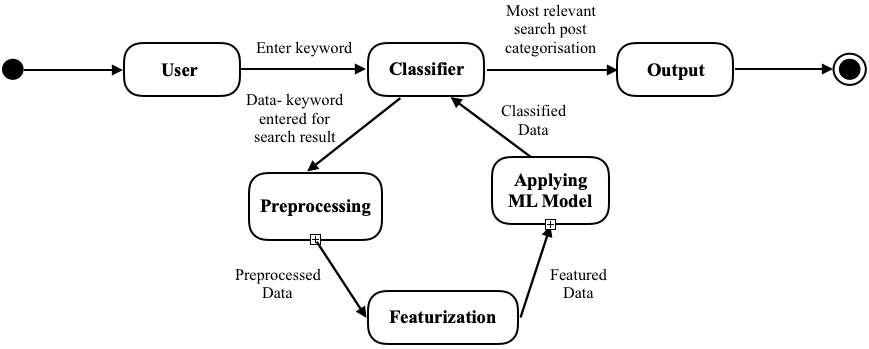


Fig 4. 6 State Diagram for Newsfeed Data mining and interactive insights extraction

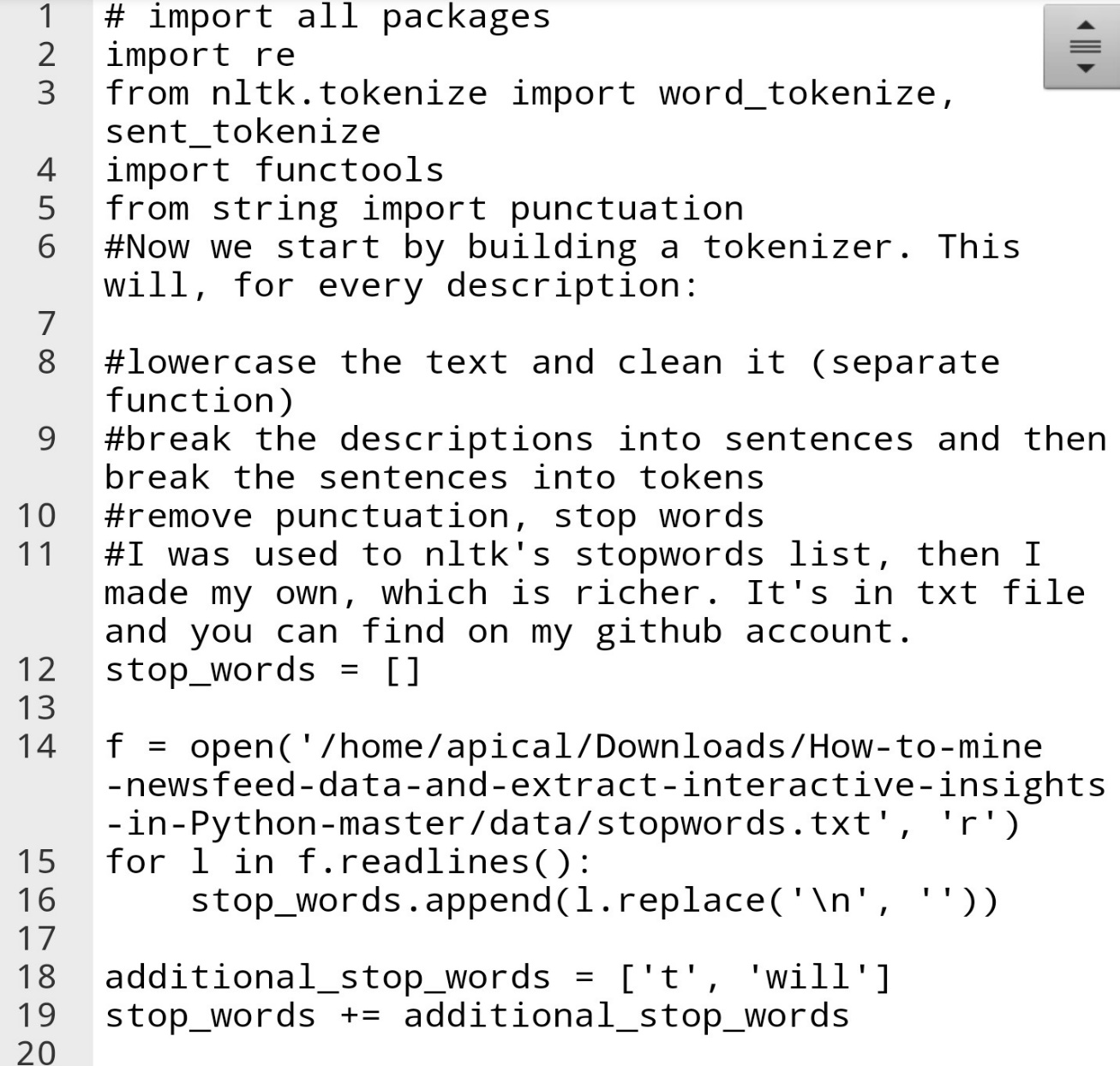
The state diagram shows the various state transitions that the processing within the system undergoes. The various states are the user input state such as keyword, classifier, processing state, applying ML model, featurization. The process starts with accepting the keyword from the user. It is accepted and given to the processing state which tokenizes the article that’s found. The articles with the most meaningful articles are found, and then useful article is extracted. The useful data is then used to make the dataset better and give improved results each time.

* 1. **Collecting data:**

We collect the remaining tokens and then again form a string and sent to further processing.

Pictures of the system under implementation are given below, along with a description.   
In the Fig 6.1, Fig 6.2 and Fig 6.3, the proposed system is being implemented.

Fig 6.1.1 System Implementation 1 – Import library and removal of redundant words



In Fig 6.1.1, the system has just been started up, and import the required data. An input has been initialized, and single process has been performed. It converts the string or the sentences in the tokens for the analysis. Also, we created a data set in which we collected set of words where most of the data which is used in the article which is mainly redundant. So the program eradicate those words from the dataset which we are going to tokenize.

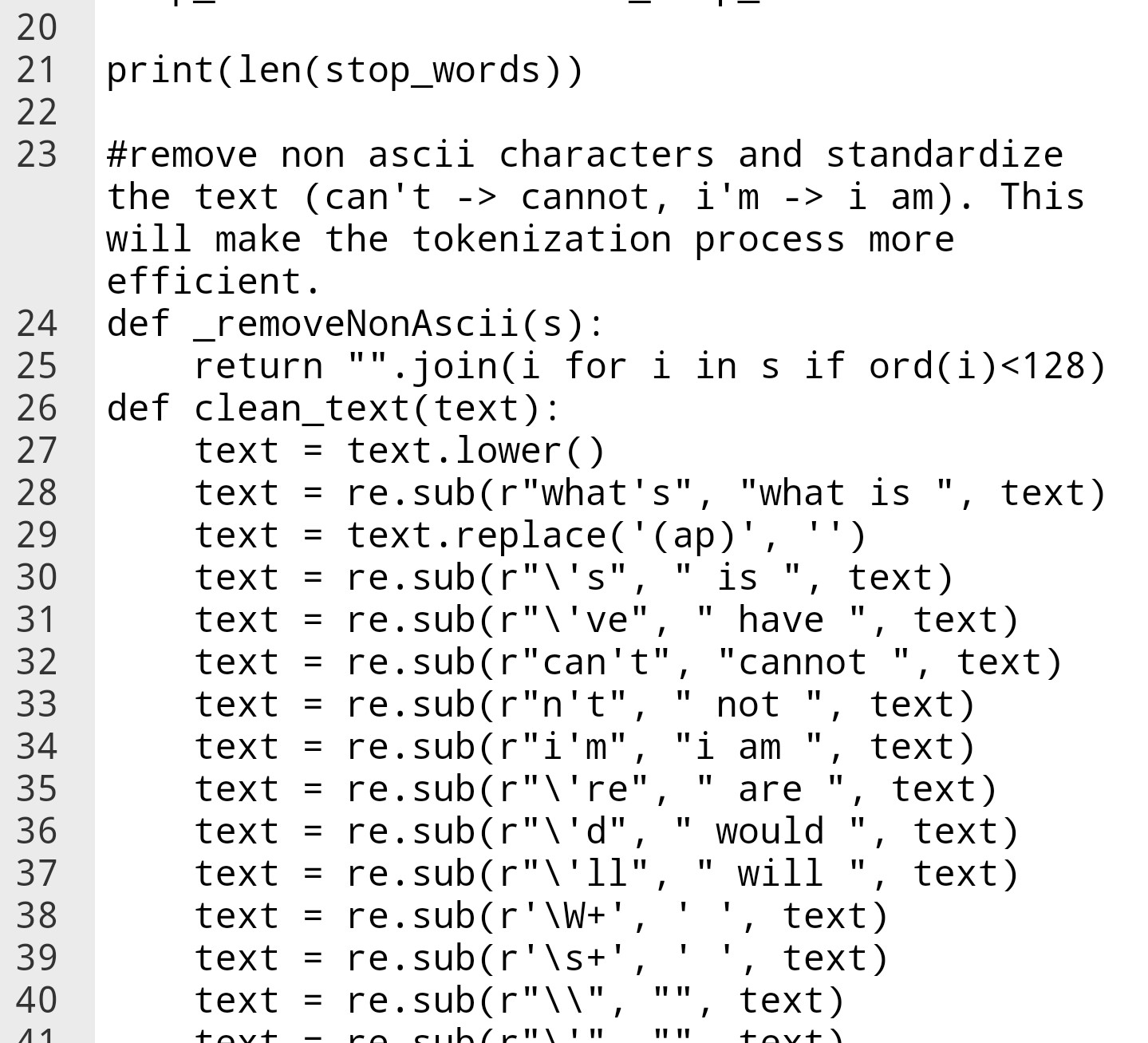


Fig 6. 2 System Implementation 3 – Removal of apostrophe and short words

In Fig 6.2, two more process commands have been initialized. These commands converts all into lower cases for better analysis of the data. And then this program converts non ascii values. And then replaces all the non ascii values to the ascii values.

Fig 6. 2 System Implementation 2 – Procedural Node and Display of result

Then while changing the non ascii values to the ascii values we simultaneously clearing the data that are redundant.

In Fig 6.1.3, the system asks the user, whether or not the code, representing the current progress must be displayed. This is an option given to the user.

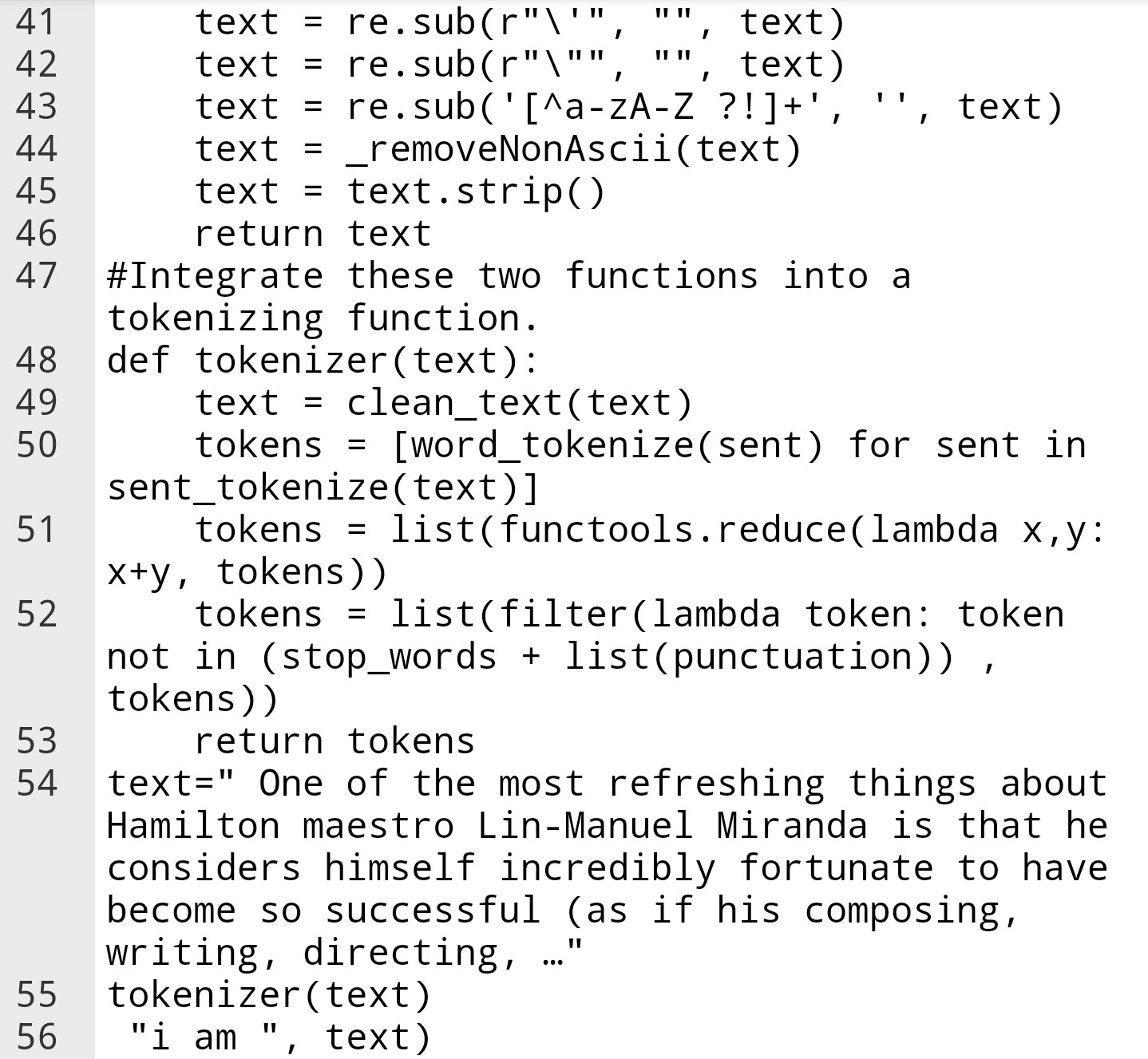


Fig 6.1.3 System Implementation 3 –Creating list of all clear words

However, any words with an apostrophe and abbreviated word are broken down in to simple word with the help of this program. Then we get a data set of all clean and simple word in the form of tokens.

* 1. **Discussions**

In Fig 6.4, the program shows the clearing of the apostrophe and also the punctuation which are not needed for analysing the news article

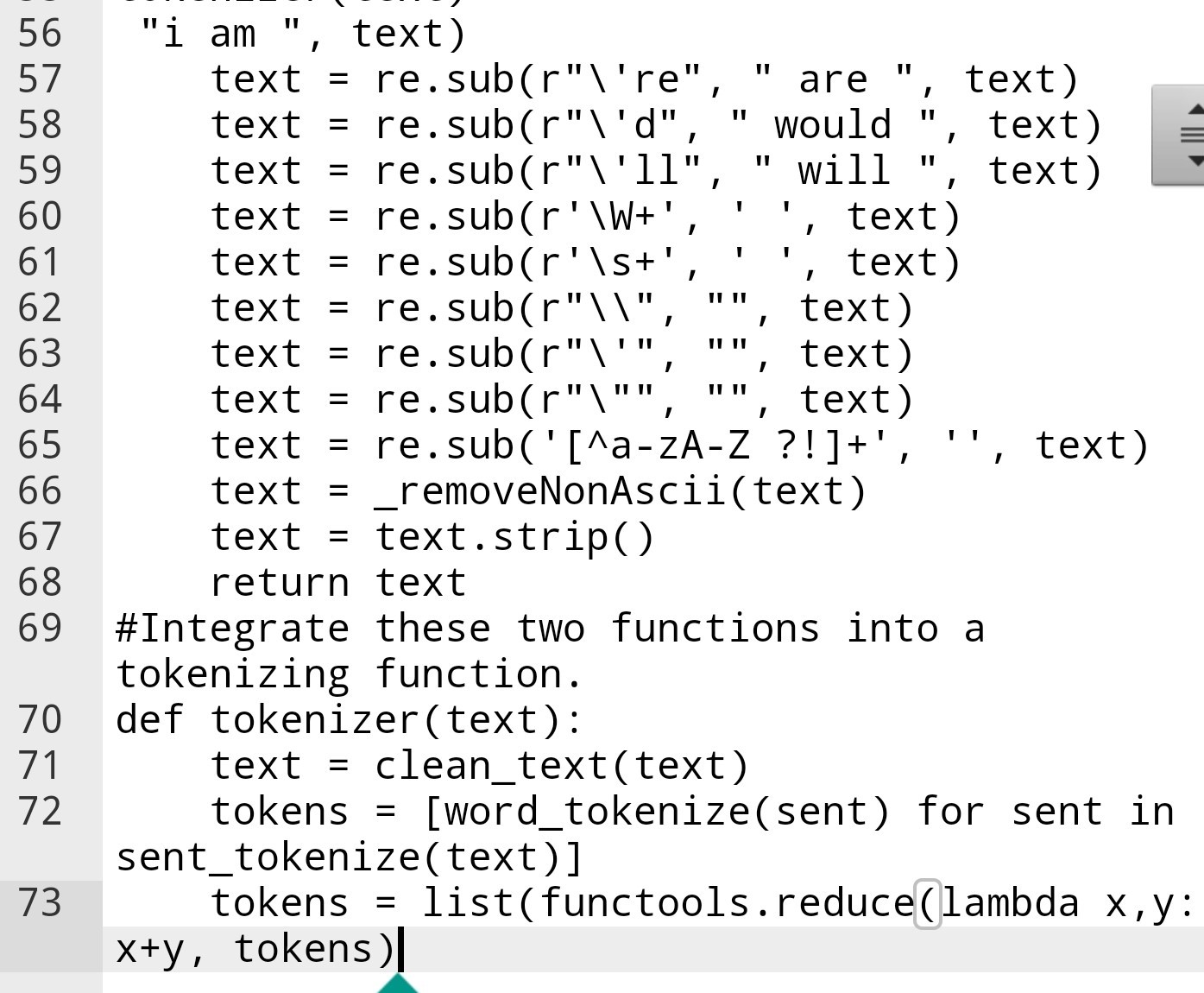


Fig 6. 4 Tokens with clean word and removal of punctuation

In Fig 6.5, the program that was generated is executed. This was performed manually, by giving a random article from BBC to the system. So, the generated words that we observe in the command are the keywords that are taken from the article given as input.

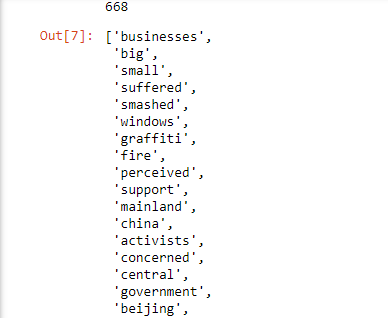


Fig 6. 5 Tokens created

The executed program functions perfectly. The above image confirms that the system is able to divide the articles. The complete article has been cleaned up from all the redundancy that are there are required for humans and not machines. So, the keywords have all been tokenized. The processes have been performed successfully and sequentially. The results have been displayed as per the preferences of the user. The system is able to extract the keywords from the article successfully to see if the article is relevant from the user’s perspective.

**References**

[1] Helle Sjøvaag and Eirik Stavelin, “The Methodology of Online News Analysis: A Quantitative Approach to Ephemeral Media”. 10.1002/9781444361506.wbiems192

[2] Beakcheol Jang and Jungwon Yoon, “Characteristics Analysis of Data from News and Social Network Services”. Digital Object Identifier 10.1109/ACCESS.2018.2818792

[3] Federica Cherubini and Rasmus Klies Nielsen “Editorial Analytics: How News Media are Developing and Using Audience Data and Metrics”. RISJ Publications

[4] R. K. Nielsen and K. C. Schrøder, ``The relative importance of social media for accessing, finding, and engaging with news: An eight-country cross-media comparison,'' Digit. Journalism, vol. 2, no. 4, pp. 472\_489, 2014.

[5] C. S. Lee and L. Ma, ``News sharing in social media: The effect of gratifications and prior experience,'' Comput. Human Behaviour, vol. 28, no. 2, pp. 331\_339, 2012.

[6] Deuze, M. (2003). The Web and its journalists: Considering the consequences of different types of news media online. New Media & Society, *5*(3), 203–230.

[7] Karlsson, M., & Strömbäck, J. (2010). Freezing the flow of online news: Exploring approaches to the study of the liquidity of online news. Journalism Studies, 11(1), 2–19.