**Data pre-processing and Discretization**

**Team Members:**

Charishma Gurram – 800960338

Vivekananda Adepu – 800967951

Mounika Pabbathireddy – 800934319

Hanisha Marothu – 800966169

Nikhita Munlagiri – 800960995

Pooja Gajapaga – 800984655

**Project description:**

With the rapid expansion of ecommerce, more and more products are sold on the Web, and more and more people are buying products on the Web. In order to enhance customer satisfaction and their shopping experiences, it has become a common practice for online merchants to enable their customers to review or to express opinions on the products that they buy. With more and more common users becoming comfortable with the Internet, an increasing number of people are writing reviews. As a consequence, the number of reviews that a product receives grows rapidly Some popular products can get hundreds of reviews at some large merchant sites. This makes it very hard for a potential customer to read them to help him order to make a decision on whether to buy the product. In this application, we are able to recognize all the features of a product and also the opinion carried about it by the customers through the reviews that will be processed. The problem is extracting the features form the reviews which are explicit is a difficult task.

**Project scope:**

The scope of the project is to gather reviews for a particular product and once the reviews have been gathered our task to remove the stop words from all the reviews as they do not give any meaning to the task at hand. Once all the stop are removed from the text we then tag each word with its respective parts of speech and then perform discretization to this.

**Software Development Model: Agile**

Agile is a time boxed, iterative approach to software delivery that builds software incrementally from the start of the project, instead of trying to deliver it all at once near the end.

Since our development team was rather small, we wanted a hands on approach with continuous involvement of all the team members from end to end instead of developing the software using the standard waterfall model. Our small team provided better opportunity to practice Agile methodology.

Agile works by breaking projects down into little bits of user functionality called [user stories](http://www.agilenutshell.com/user_stories), prioritizing them, and then continuously delivering them in short two week cycles called [iterations](http://www.agilenutshell.com/iterations).

We have come up with various user stories and focussed on developing the software iteratively by concentrating on one user story at a time. Some of the user stories that we focussed on were as follows. First to develop a user interface so that testing the code for its results would be much more easier and time saving. Next to make sure that the step in our project which is to remove stop words works correctly and also try different data in order to come up with all possible stop words or as many can be managed. The next user story was the main part of the project which is parts of speech tagging. All the words were given their parts of speech and were represented by the word followed by underscores and then abbreviations for parts of speech.

The last user story was to discretize all the words after parts of speech tagging and to separate nouns and adjectives. Since we are also not very experienced with collecting requirements. We have used agile as a strong point to cover that weakness and update our requirements as we went about with our project.

**Functional Requirements:**

* Gather reviews about the product from online websites
* Select a set of product features to rate on.
* Determine the ratings for the selected features based on the sentiment of the sentence in which it appears.
* Summarize the ratings for the features as the total number of positive and negative points for each of the review.
* Draw a graph based on the summarized ratings and depicting the trend of a particular feature.

**Non-Functional Requirements:**

* Maintainability: The software needs to be easily maintained. There must be proper documentation of the entire software in such a way that it is easily readable by a new developer. In Case of any issues the development/maintenance team should be able to fix the problem quickly and push it to production and ensure that there is very less downtime.
* Usability: The application should be very easy to understand and use. Novice users who are new to technology should be easily able to figure out what to do in the application to accomplish their tasks.
* Availability: The application should be always online providing the latest status of the reviews and delivering the output in a timely manner.
* Scalability: This application can be scaled if the user interest grows. The application should be able to handle considerable number of requests from users and at the same update users quickly.

**System Requirements:**

**User Requirements**

OPERATING SYSTEM: Win XP and above

WEB BROWSER: Any compatible browser.

**Hardware Requirements**

PROCESSOR: PENTIUM IV 2.6 GHz

RAM: 512 MB DD

RAM MONITOR: 15” COLOR

HARD DISK: 20 GB

CDDRIVE LG 52X

KEYBOARD: STANDARD 102 KEYS TREND ANALYSIS

**Software Requirements**

TOOLS USED: Stanford POS Tagger, Vaadin, Eclipse

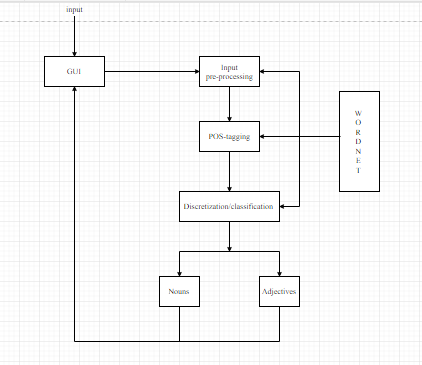
OPERATING SYSTEM: WINDOWS

LANGUAGES: JAVA

DATABASE: FILE SYSTEMS

**Architectural Design Diagram:**

The below diagram shows the architectural design of the project.



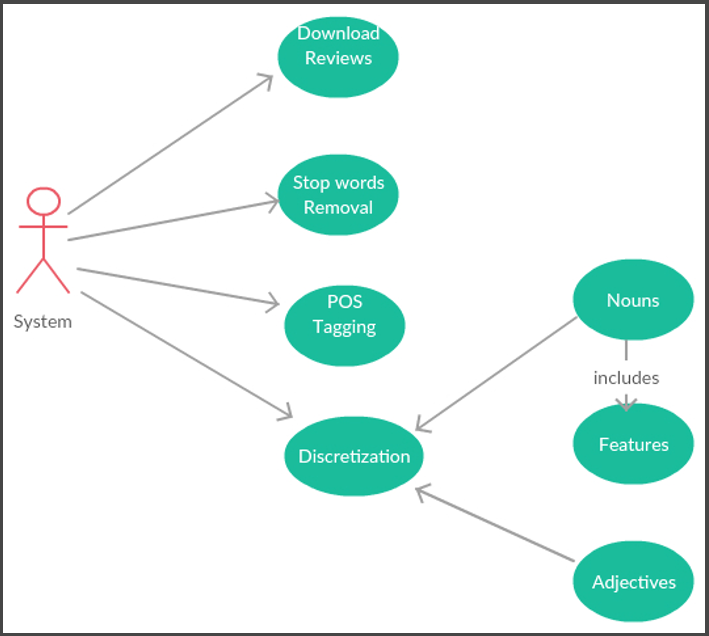
**UML Diagrams:**

The Unified Modelling Language (UML) is a general-purpose, developmental, [modelling language](https://en.wikipedia.org/wiki/Modeling_language) in the field of [software engineering](https://en.wikipedia.org/wiki/Software_engineering), that is intended to provide a standard way to visualize the design of a system.

We have chosen few of the diagrams to depict our project model.

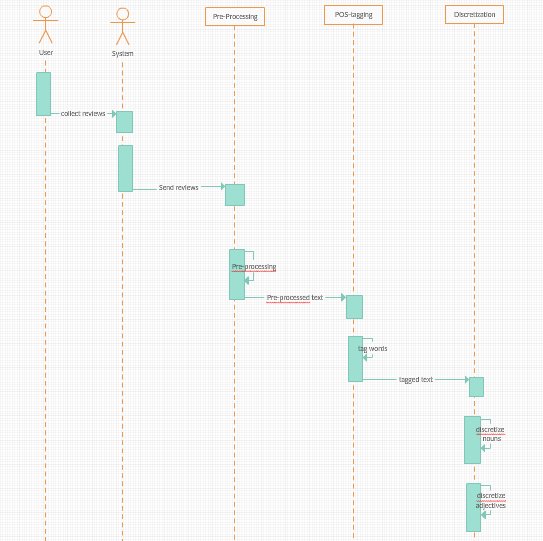
**Use Case Diagram:**

Use case diagrams are a set of use cases, actors and their relationships. They represent the use case view of a system. A use case represents the particular functionality of a system. It is used to describe the relationships among the functionalities and their internal and external controllers where the controllers are known as actors.



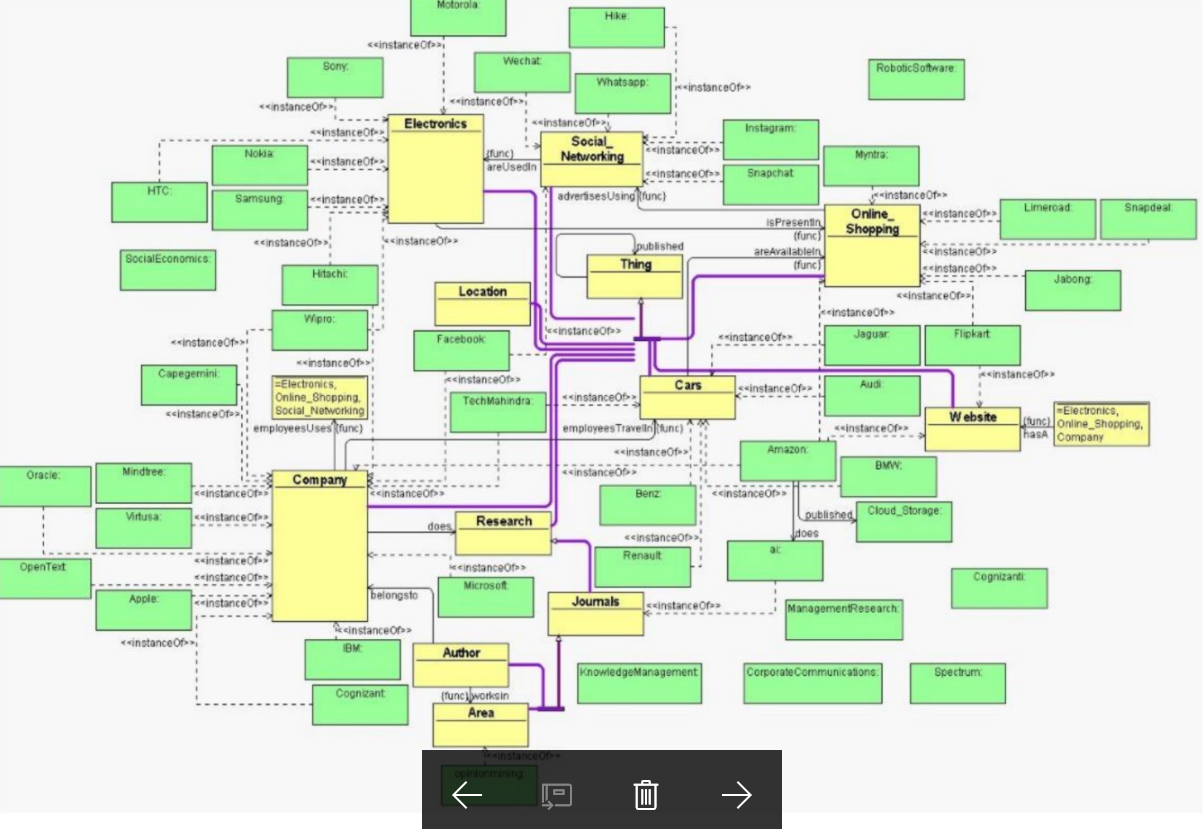
**Sequence Diagram:**

A sequence diagram is an interaction diagram. The diagram deals with some sequences, which are the sequence of messages flowing from one object to another. Interaction among the components of a system is very important from implementation and execution perspective. It is used to visualize the sequence of calls in a system to perform a specific functionality.



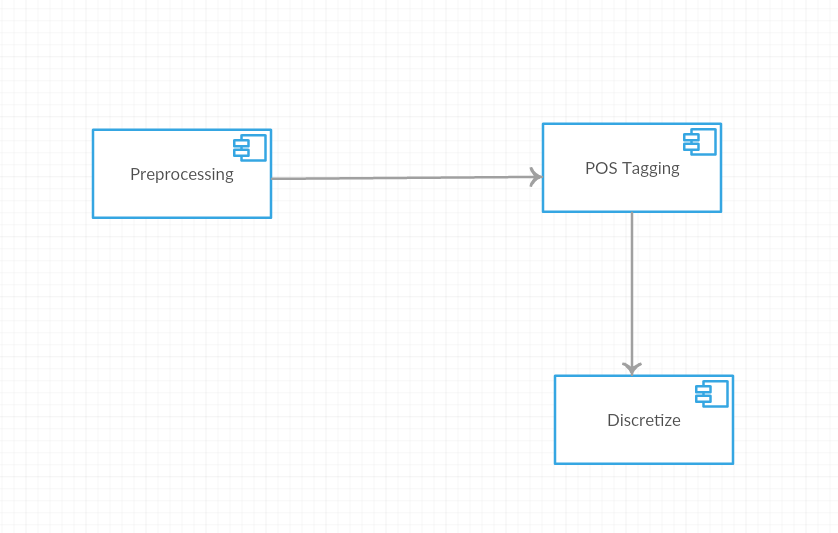
**Class Diagram:**

In software engineering, a class diagram in the Unified Modelling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects



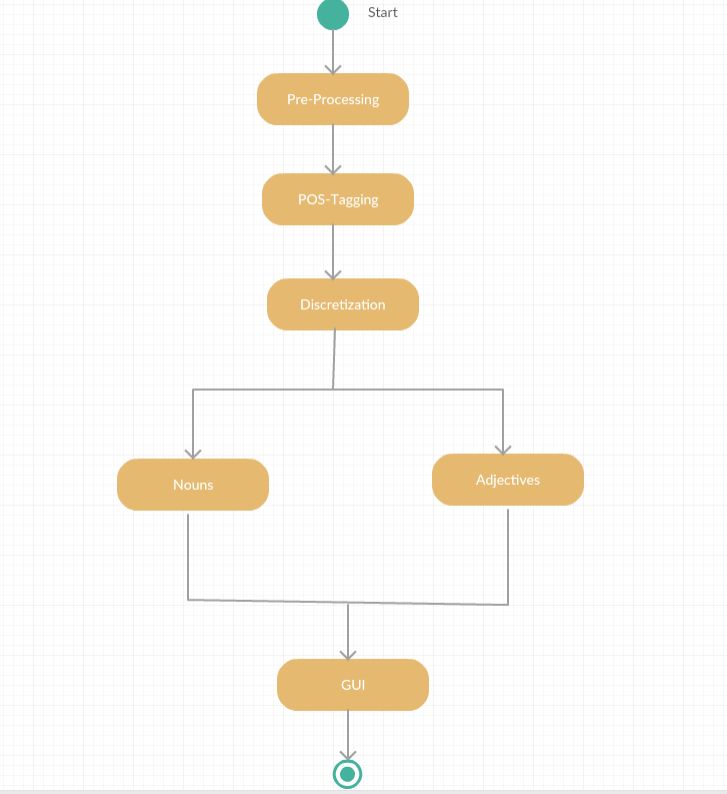
**Component Diagram:**

Component diagram is a special kind of diagram in UML. The purpose is also different from all other diagrams discussed so far. It does not describe the functionality of the system but it describes the components used to make those functionalities.



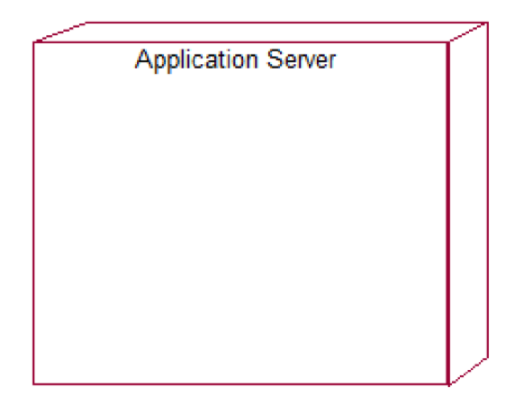
**Activity Diagram:**

Activity diagram is another important diagram in UML to describe dynamic aspects of the system. Activity diagram is basically a flow chart to represent the flow from one activity to another activity



**Deployment Diagram:**

Deployment diagram is a structure diagram which shows architecture of the system as deployment(distribution) of software artifacts to deployment targets. Artifacts represent concrete elements in the physical world that are the result of a development process.



**Implementation Steps:**

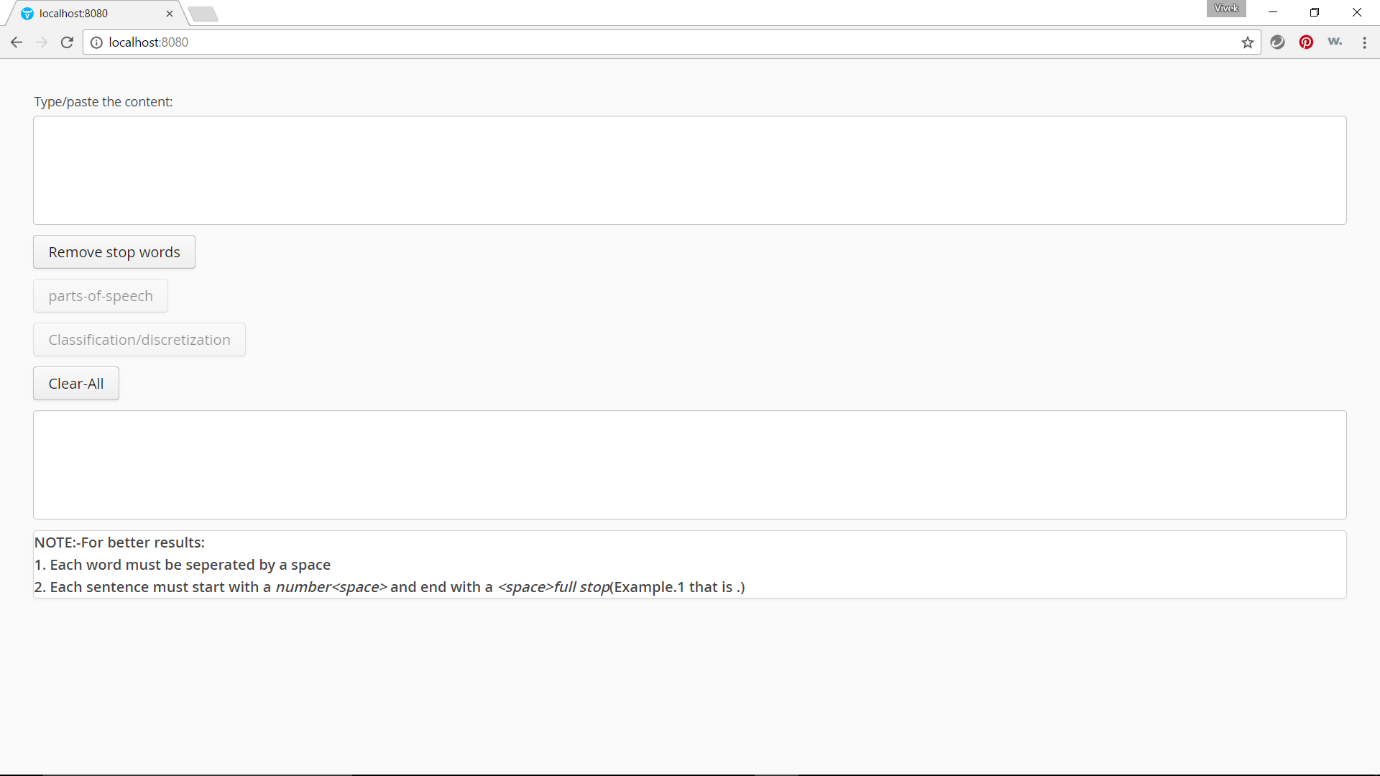
**Step-1:**

**Pre -Processing:**

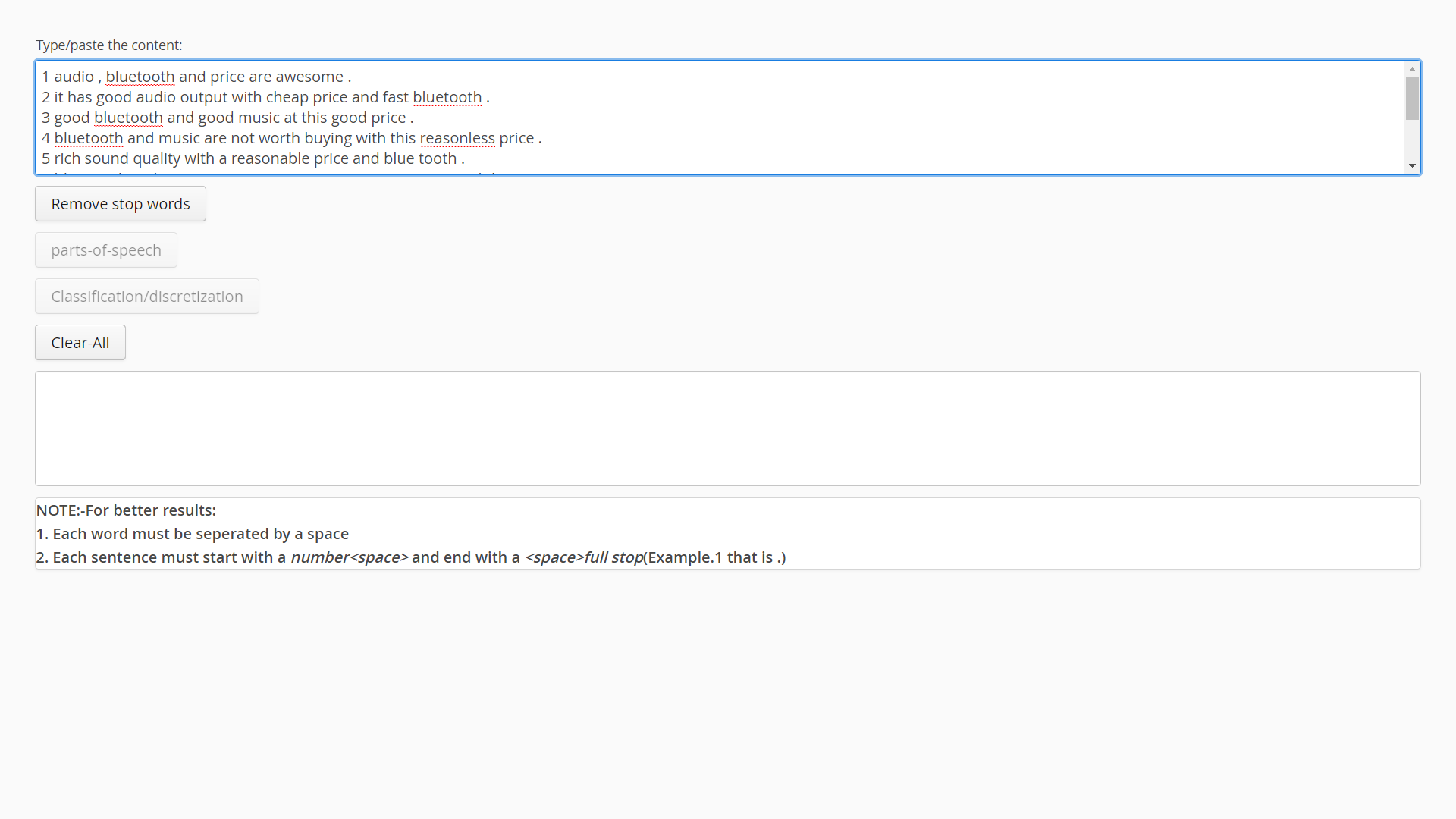
All the stop words present in a given input text file which consists of selected product reviews are removed.

* The text file that contains all the selected product reviews is given as input to the program.
* The program processes the text file and then splits each word of the file.
* Each word of the file is compared with the set of stop words.
* If the word that is present in the file is a stop word, then it is eliminated.
* Since all the stop words are removed the resultant file has no stop words.
* The program creates a file and then writes all the reviews that do not contain the stop words.

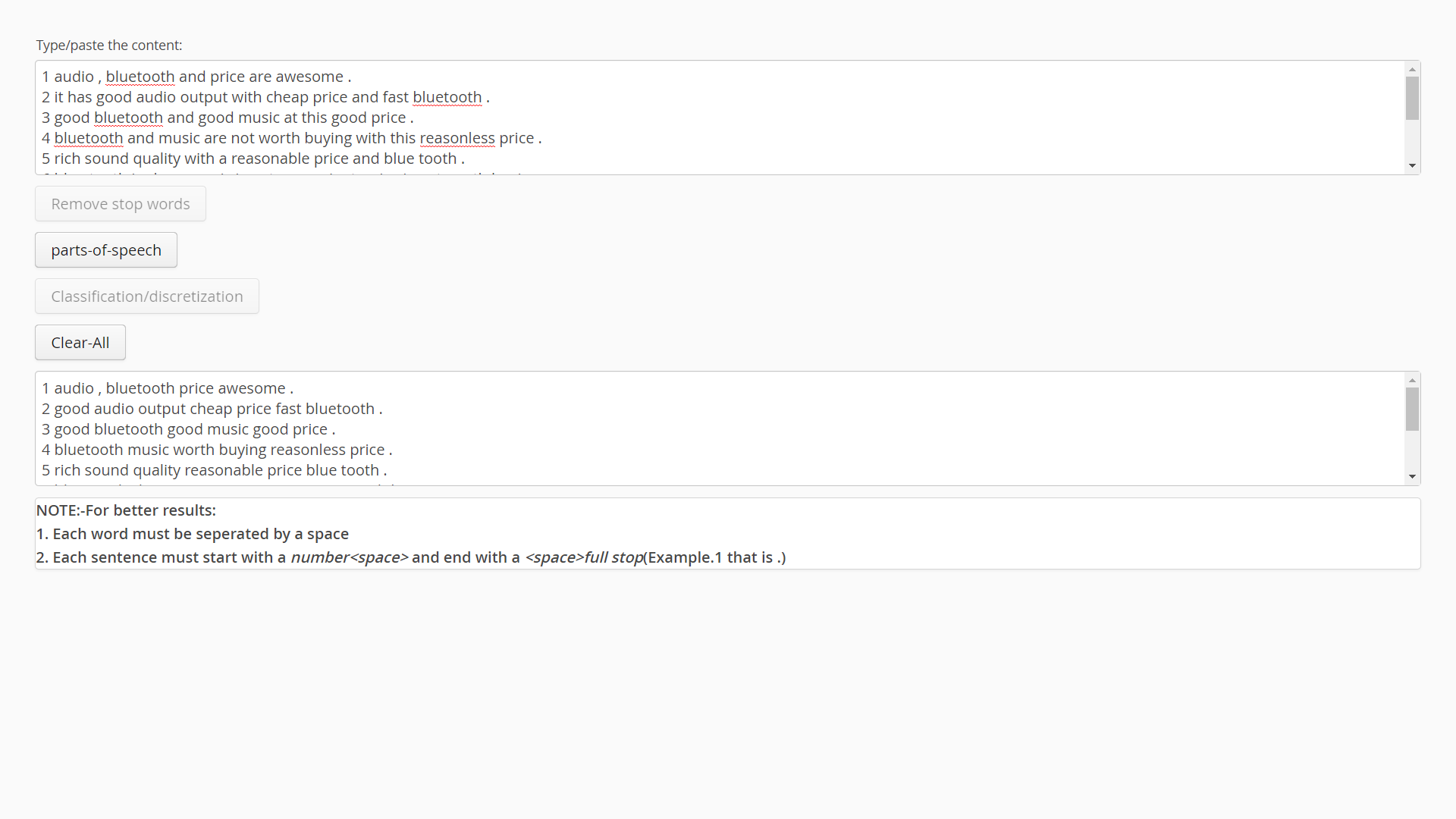
**Our Webpage:**



**Input:**



**Processed Output:**



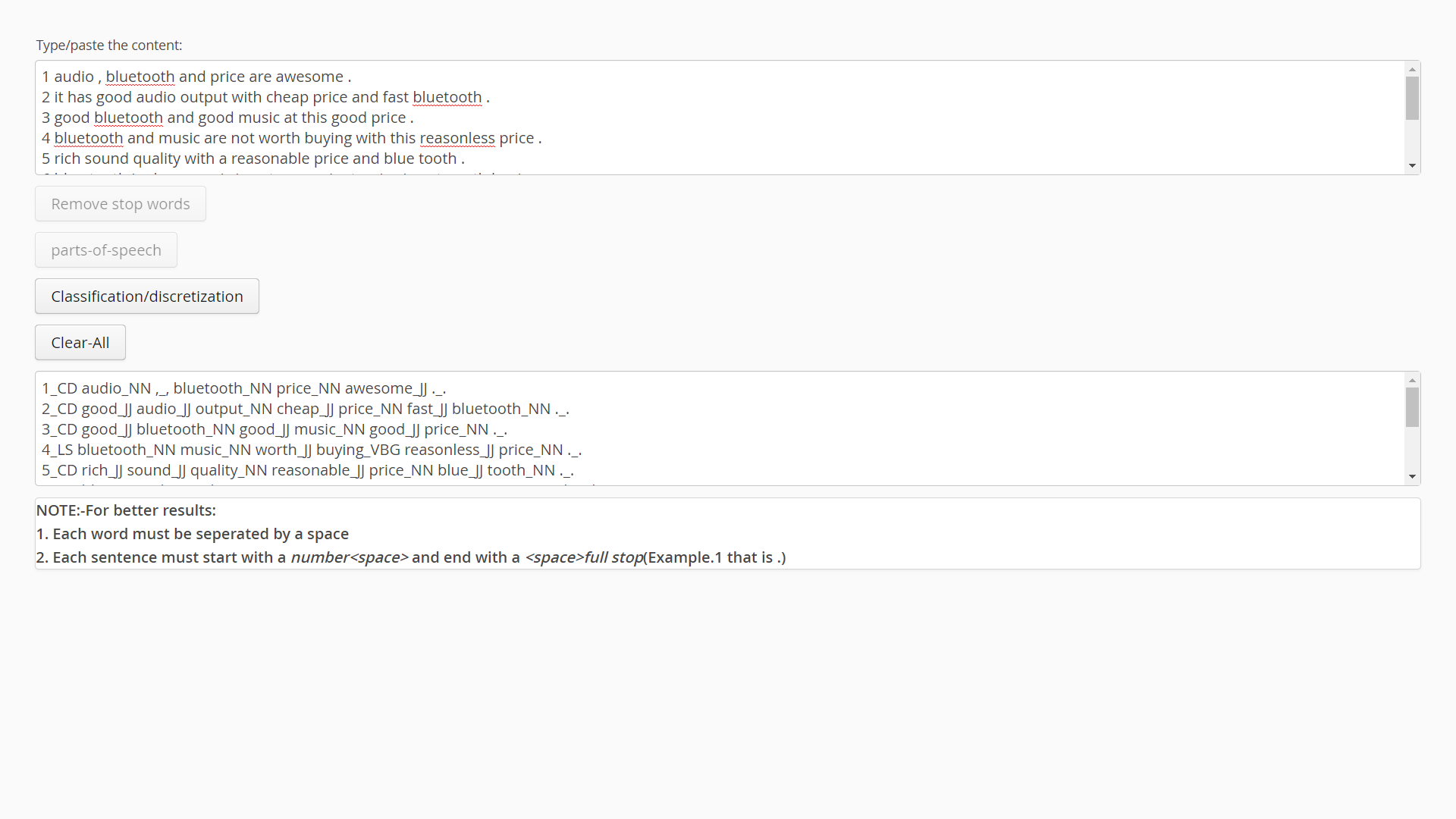
**Step-2:**

**Parts of speech tagging:**

Each word in the text file is tagged with the respective parts of speech.

* The output text file from the step-1 is given to the step-2 for parts of speech tagging.
* In this step the Stanford-postagger.jar file is used to tag the parts of speech of any given text.
* The .jar file can be accessed by using a command in the command prompt.
* The command is embedded in a java program and use process builder that opens the command prompt and directs to the path that is given where the Stanford-POStagger is present.
* The Stanford-POStagger hence tags the content of the text file containing the reviews.
* The output is then written to a text file.

**Processed Output:**



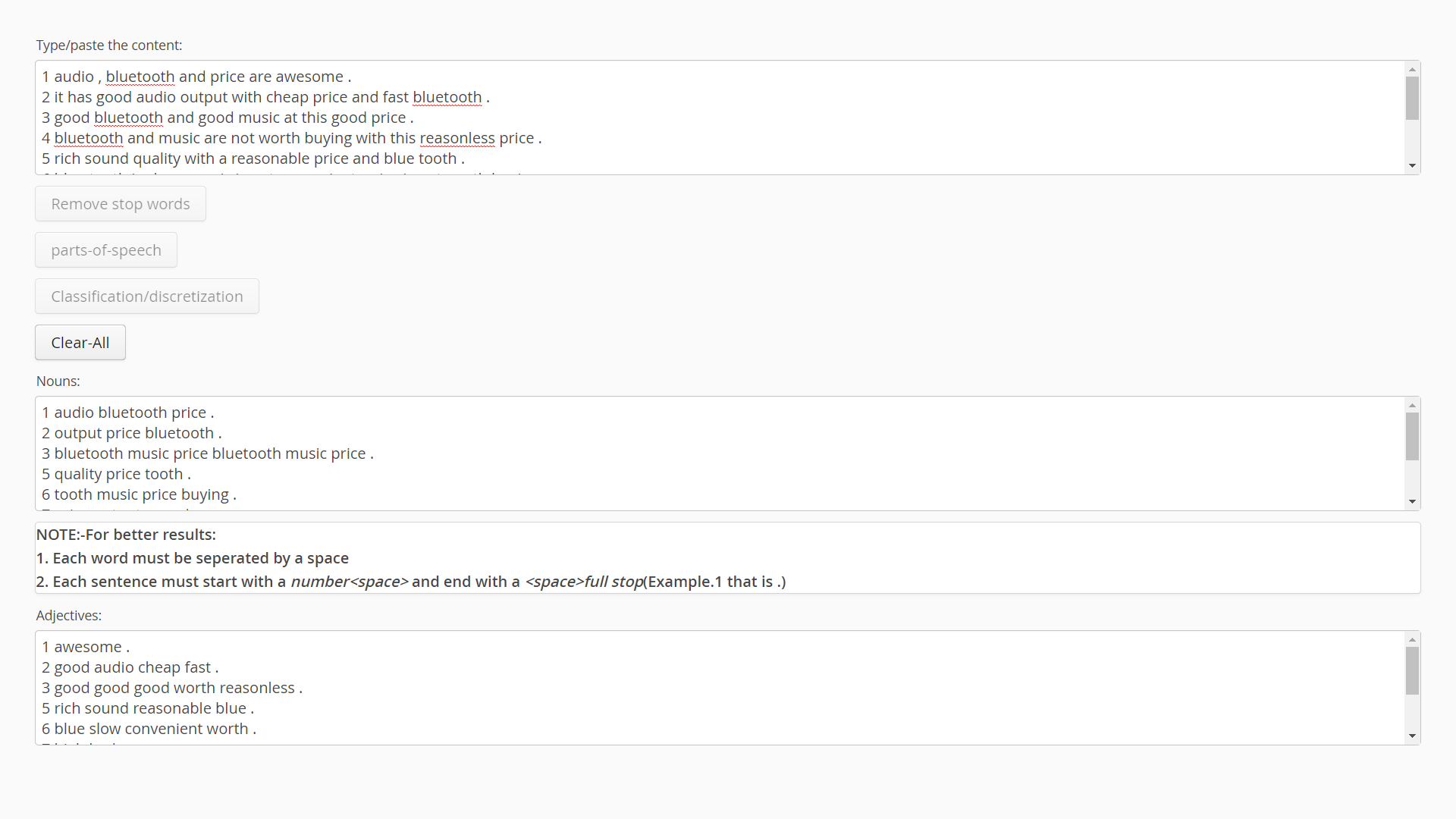
**Step-3:**

**Discretization:**

The adjectives and nouns that are present in the tagged file are separated and stored in two different files.

* The output file from step-2 is dynamically given to step-3 for discretization.
* The Nouns and Adjectives that are present in the tagged file are separated in this step.
* The program checks each element of the file and then retrieves the nouns in the file.
* The Nouns that are retrieved are stored in a text file that is created by the program.

**Processed Output:**



**Input File:**

1 audio, bluetooth and price are awesome.

2 it has good audio output with cheap price and fast bluetooth.

3 good bluetooth and good music at this good price.

4 bluetooth and music are not worth buying with this reasonless price.

5 rich sound quality with a reasonable price and blue tooth.

6 blue tooth is slow, music is not convenient, price is not worth buying.

7 price is not high but this output sound is extremely bad.

8 affordable phone but it has poor sound output.

9 this is very comparable and cheaper budget alternative to the s4 the music features were defective.

10 music is very slow and get stuck sometimes and the size isn't good enough.

11 audio is very low but has a good user interface for beginners.

12 galaxy grand has good audio with a powerful bluetooth.

13 very cool audio features and bluetooth is awesome.

14 extremely low sound and touch response is very good.

15 audio response is not good.

16 music features are magnificant.

17 excellent audio.

18 audio is acceptable.

19 i am very impressed with loud sound feature.

20 camera and display are functioning awesome with great clarity, the phone has got good size also.

21 size is not satisfactory, camera and display are lively.

22 camera takes good pictures and pictures are display is good.

23 picture clarity is not good and display works bad sometimes.

24 the camera is acceptable with pretty good flash.

25 perfect camera with brilliant flash light.

26 camera not agreeable but flash is working good.

27 the camera isn't the best at 5 mp and it has bad audio.

28 camera takes good pictures.

29 the camera takes very colourful photos.

30 camera worked great and pictures were wonderful.

31 the battery only lasts half day and its disappointing, memory is not good enough, keypad is a little tight.

32 battery has been disappointing memory is erratic, great screen.

33 battery has bad issues capability to hold two sim cards is superb installation of applications is inconsistent.

34 low batter life but good dual sim feature and clear call clarity.

35 worst memory and battery because they always low.

36 poor battery life and memory.

37 display is nice but erratic battery.

38 disadvantage is its poor battery life

has the plus: double sim card feature is great.

39 dual sim card is great feature in this but battery is weak.

40 battery and dual sim card are very bad.

41 call clarity is great due to strong dual sim card operations.

42 the dual sim comfortable! enjoys the benefits, it has great apps and easy to handle.

43 battery is acceptable.

44 battery is low.

45 display and dual sim are pretty good.

46 problematic calls and display.

47 troublesome calls and display is not good enough.

48 only bad thing is that you cannot use the two sim cards together.

49 great phone for a reasonable price.

50 budget of phone is impressive.

51 the price was great for the features.

52 upgrade is good.

53 the size is comfortable.

54 great size for travel.

55 the dual sim slots are super.

56 dual sim option which is very lovable.

57 the screen size is perfect for my needs.

58 not able to increase brightness in this poor display.

59 size is perfect.

60 comfortable size.