# Title of the Project: SwiftKey Predictive Text

**Submitted by**

**Name of the Group: Group 5172226**

**Name of the Students**

**Ritvik Kanchan**
**Manoj Yadav**
**Shreya Shinde**
**Apeksha Singh**

**Batch Year (2020-2022)**

**Course Details**

**MSc DSAI (PART II)**

**Institute Details**

**Ramniranjan Jhunjhunwala College**

**Name of the Guide: Prof. : Mr. Cyrus Lentin**

**PROJECT SYNOPSIS**

| | | |
|---|---|---|
| Name of Group | Group 5172226 | |
| Name of Students | 05 Ritvik Kanchan<br>17 Manoj Yadav<br>22 Shreya Shinde<br>26 Apeksha Singh | |
| Program | Batch Year | Academic Year |
| M.Sc. DSAI | 2020-2022 | 2021 |

| | |
|---|---|
| Name of Guide | Prof. Cyrus Lentin |
| Title of the Project | SwiftKey Predictive Text |
| Reason for this topic | In order to study the working of applications based on text analysis model. |
| Project Details | |
| Objectives / Goals | • Creating a Working Web App that success fully takes required inputs and gives required outputs.<br>• Model should be able to predict word based on the last character and suggest it. |
| Please mention key Literature which you plan to Study | • https://www.nltk.org/book/<br>• https://textblob.readthedocs.io/en/dev/<br>• https://searchengineland.com/how-google-instant-autocomplete-suggestions-work-62592<br>• https://dzone.com/articles/how-to-build-a-google-search-autocomplete |
| Expected Outcomes which you expect after Data Analysis | Creating a Web App which predicts word that is going to be entered based on the last character. |

| | |
|---|---|
| | |
| | **Guide Signature** |

## DATA DICTIONARY:

The source data file we have used in the project was already given to us with the problem statement. Data source link: https://drive.google.com/file/d/0B95yMv4YhoSOLVkwbVhaaFBtMFk/view?resourcekey=0-wLDqUjn9P9JuWIxxpoTbWg

The dataset contains text from blogs and it is consists of multiple sentences. Sentences may include numbers, profane words, punctuation marks words with capital letters.. Therefore before using it cleaning and transformation of data is very necessary.
The size of the data file is quite large and appear to require significant memory to run it. As this is before any processing it may be necessary to sample the files rather than use the entire file to avoid hitting memory and processing constraints.

# EXPLORATORY DATA ANALYSIS:

A) Preprocessing data:

After importing and reading the dataset first step is to clean the data, to remove all the unnecessary data Preprocessing will transform the raw data into an understandable format for making it suitable for machine learning models, which will also increase the accuracy and efficiency of a model. It involves the below steps:

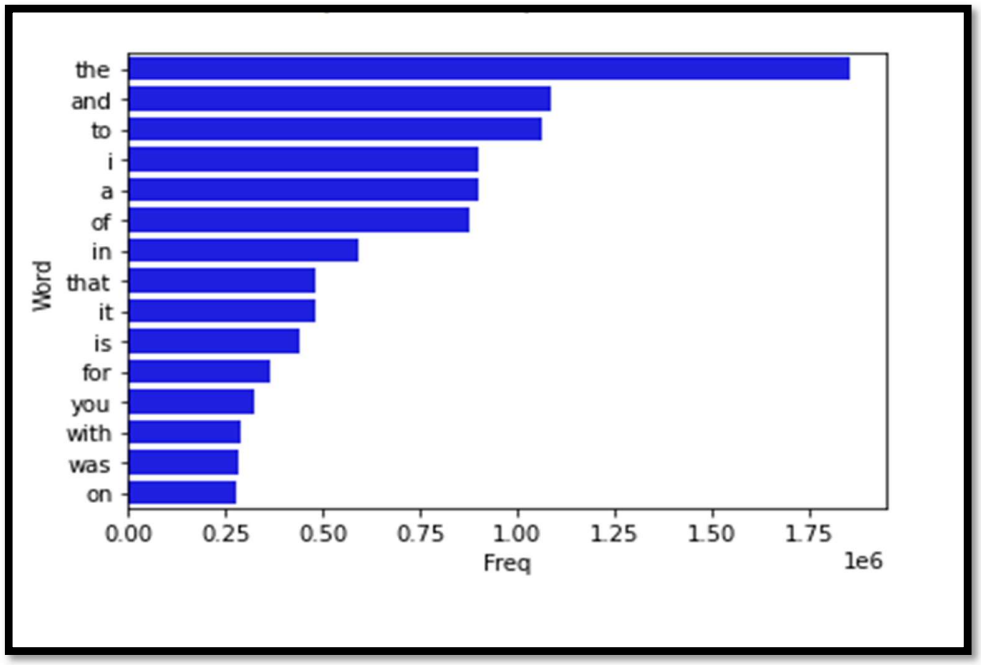
1) Importing libraries: we will import libraries such as pandas, matplotlib, seaborn, wordcloud.
2) Read the data: reading the data from the device
3) As the size of the imported file is too large to display. we are displaying only 3000 lines of text.
4) Data Transformation:
   - Tokenization: tokenization refers to splitting or breaking a larger body of text into smaller lines. Various tokenization functions are in-built into the nltk module itself
   - Convert to lower case: after tokenization we have converted the data into lower case.

5) Data cleaning: we will remove unnecessary text and sign.
   - Removing digits and punctuation:
   - Removing profane words:
   - Removing words with length zero

## VISUAL DATA ANALYSIS :

In the barplot given below, we have plotted frequently occurring words versus no. of time these words are occurring.

Top 15 most occurring words:




In the diagram given below, we have used Word Cloud to represent the most occurring words. The larger the word is the more number times that word is occurring.

## MODELING AND PREDICTION:

We have created an algorithm which will predict next word based on the occurrence of that word. The prediction algorithm takes an input, convert it into lower case.

- If the input is an incomplete word It will consider the last character of the word as input and it will predict the complete word. Then based on that input if the user has an incomplete word, then predict the completion of the word else predicts the next word.
- If the input is a complete word with the help of lastword() function it will return last word of the given sentence and with the help of NextWordPred() function it will predict next word that the user might type.

```
Enter Input :-
frien
Enter the number of words to predict :-9
frien
            Word   Freq
0        friends  16417
1         friend  12533
2       friendly   1571
3     friendship   1058
4    friendships    392
5   friendliness     40
6       friends…     21
7  friendsfamily     20
8     friendlier     18
```

## Data Analysis Findings

In this we are going to predict the different data into different formats. We have 3 sets of data with consists of blogs, twitter, news.

1. Analysis for blogs data:

   Here we will predict with different format to check the prediction is correct or not.

```
Enter Input :-
nam
Enter the number of words to predict :-6
nam
     Word   Freq
0    name  12821
1   named   3021
2   names   2843
3    nama    524
4  namely    383
5  naming    254
```

In this, we have just written half of the word and it has predicted itself my giving no. of options.

```
Enter Input :-
my name
Enter the number of words to predict :-7
  Word  Freq
0   of  1680
1   is   996
2  and   741
3  for   473
4    i   406
5   it   347
6  was   340
```

In this, we have tried to write a sentence and stop before it predicted us the next word.

2. Analysis for news data:

   Here we will predict with different format to check the prediction is correct or not.

```
Enter Input :-
he wa
Enter the number of words to predict :-3
wa
   Word    Freq
0   was  236318
1   way   28155
2  want   21841
```

In this, we have just written half of the word and it has predicted itself my giving no. of options.

```
Enter Input :-
long
Enter the number of words to predict :-4
   Word  Freq
0    as  1652
1  time  1275
2   and   606
3   way   541
```

In this, we have tried to write a sentence and stop before it predicted us the next word.

3. Analysis for twitter data:

   Here we will predict with different format to check the prediction is correct or not.

```
Enter Input :-
has
Enter the number of words to predict :-3
has
      Word   Freq
0      has  47907
1  hashtag   1135
2     hash    409
```

In this, we have just written half of the word and it has predicted itself my giving no. of options.

```
Enter Input :-
@gmail
Enter the number of words to predict :-3
@gmail
No Predicted Words
```

In this, we wrote a special character which doesn't predict it and throws an error.

# Project notes

# Stage 1:

## Data Source:

Data source is the location where data that is being used originates from. Large databases comprising of text in a target language are commonly used when generating language models.

Data: Our data consist of 3 three files of data available:

- Blogs
- News
- Twitter

The data files may have some foreign text but we will use English database to predict the model.

# Stage 2:

## Exploratory Data Analysis:

It performs a thorough exploratory analysis of the data, understanding the distribution of words and relationship between the words in the file.

## Preprocessing data:

1. Pre-processing helps to transform to applied our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set.
2. It is simple cleaning procedures which makes it easier to use the data in subsequent steps. Various steps are performed into it to perform the data which we needed in the model.
3. Various steps are used to process the data are:
    - Importing libraries
    - Reading data
    - Data transformation (Its work with different functions:
        - Tokenization
        - Converting into lower case
    - Data Cleaning:
        - Removing punctuation

- Removing profane words
- Removing numbers, white spaces.

# Stage 3:

## Visual Data Analytics:

Data visualization is the process of translating large data sets and metrics into charts, graphs and other visuals. The resulting visual representation of data makes it easier to identify and share real-time trends, outliers, and new insights about the information represented in the data.

# Stage 4:

## Text Prediction:

It is used to predictive the text present in the data for the next word to be there. We have use different functions to predict the data.

- LastChar() function takes text as input and returns the last char in the text
- lastword() function takes text as input and returns the last Word in the text
- predictnextText() function takes a character or a string and number as input and predicticts the completion of the word (Word that user is going to type). It returns top number of characters with highest frequency based on input number
- NextWordPred() function takes a string and number as input and predicticts the next word that the user might type. It returns top number of frequent words which came after this word based on data used.

# Stage 5:

## Problems to be encountered and overcome:

**Problem 1:**

There were special characters where we have typed any special characters so we get an error.

**Problem 2:**

There was problem of computing the data, where it was taking time to compute for solving that we used flag to avoid the error.

**Problem 3:**

There was also problem with the punctation, where we use it, they show an error for what we create function to resolve it and when we give input with the punctation it shows us "error we can't use punctation marks".

**Problem 4:**

There was problem with the input part, where if we keep it input empty it shows an error but making a function it shows as empty input, please give some input to predict.

**Problem 5:**

There was a problem with the list of word and frequency of word it shows error when it was empty, so created the new variable and a function to overcome it and if its empty it wont shows the graph of word cloud.

## Conclusions:

- In a text prediction task, a sentence, or series of words, is presented to a prediction model. The model's task is to then present the word or words with the highest probability of following the initial series to the user.
- With the help of Natural Language Processing domain, it helps to develop and train the models to approximate the way our human brains use towards language.
- The goal of this project was to create a product to highlight the prediction algorithm built and to provide an interface that can be easily accessed by others.
- It is simple and intuitive to use. Just type in the first few words of as sentence and the suggested next word will immediately show up.
- There are certain cases where the program might not return the expected result. This is obvious because each word is being considered only once. This will cause certain issues for particular sentences and you will not receive the desired output.
- However, certain pre-processing steps and certain changes in the model can be made to improve the prediction of the model.

## References:


- https://towardsdatascience.com/next-word-prediction-with-nlp-and-deep-learning-48b9fe0a17bf
- https://cyruslentin.shinyapps.io/swiftkey/
- https://medium.com/analytics-vidhya/nlp-word-prediction-by-using-bidirectional-lstm-9c01c24b2725
- https://towardsdatascience.com/exploring-the-next-word-predictor-5e22aeb85d8f

# Appendix:

1. Text Prediction Model File link

https://github.com/delta2127/Swiftkey_Project_Model

2. Website link

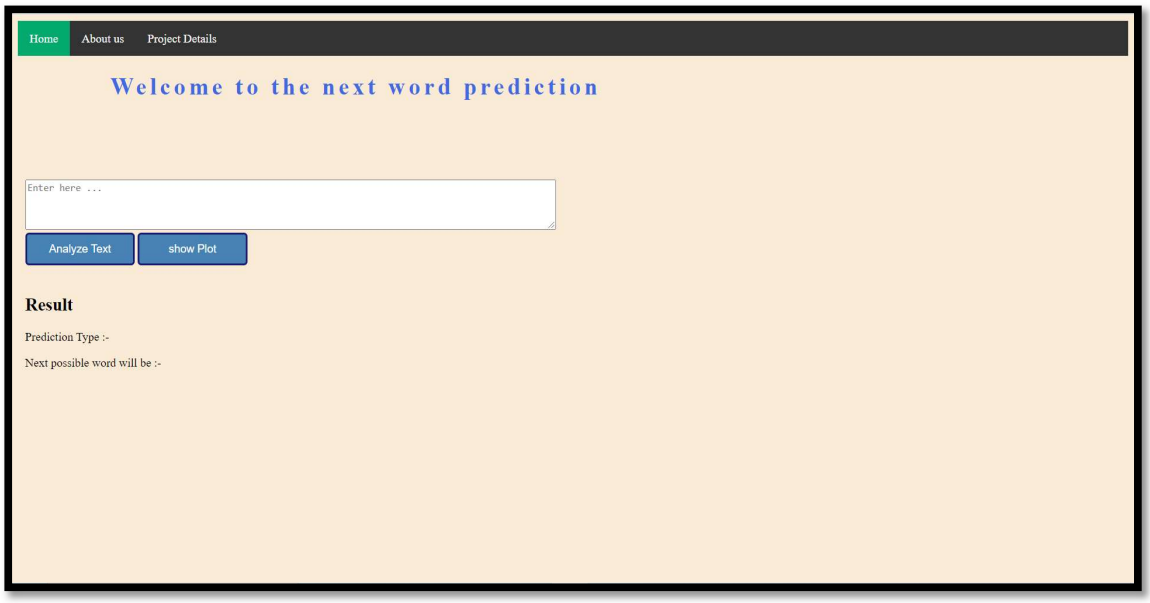
http://nextword002.herokuapp.com/

3. Project File link

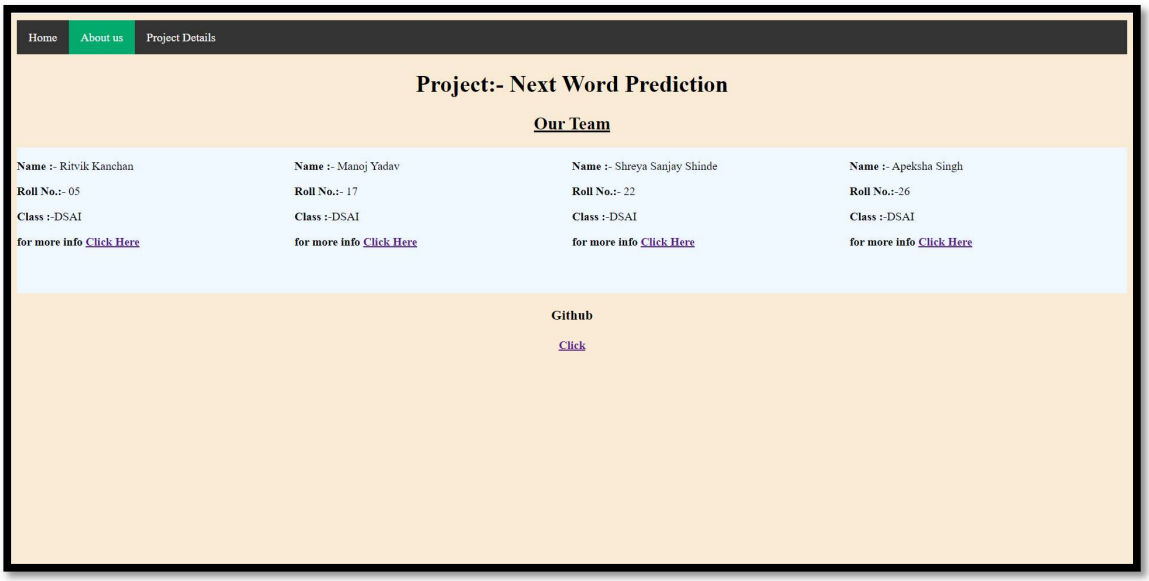
https://github.com/Manoj123-github/WebMaster/tree/main/firstProject

Output of application:

I. Home page

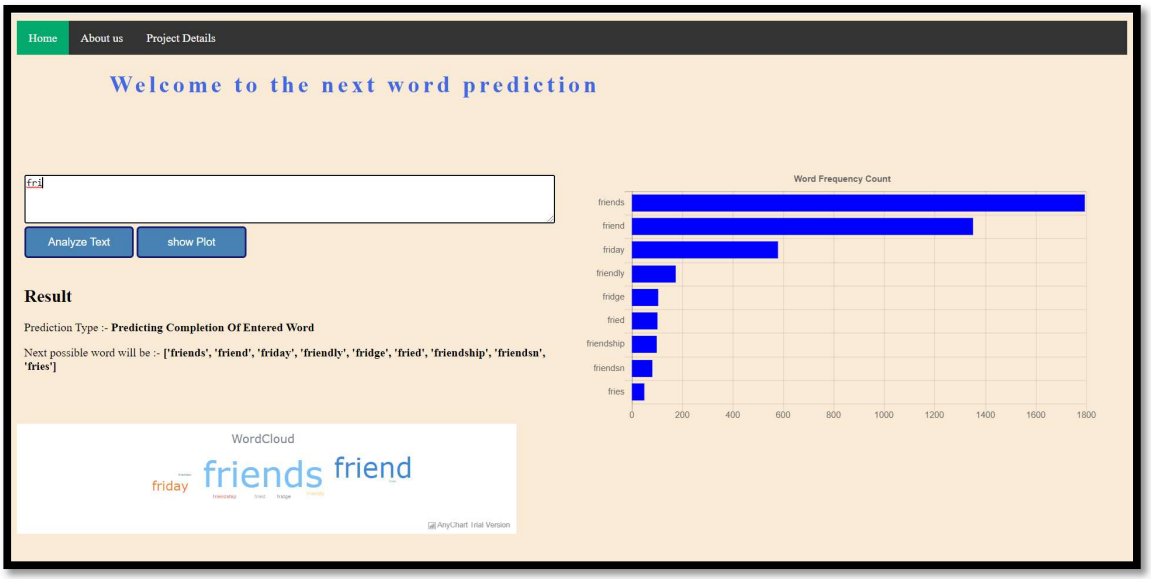


## II. About page

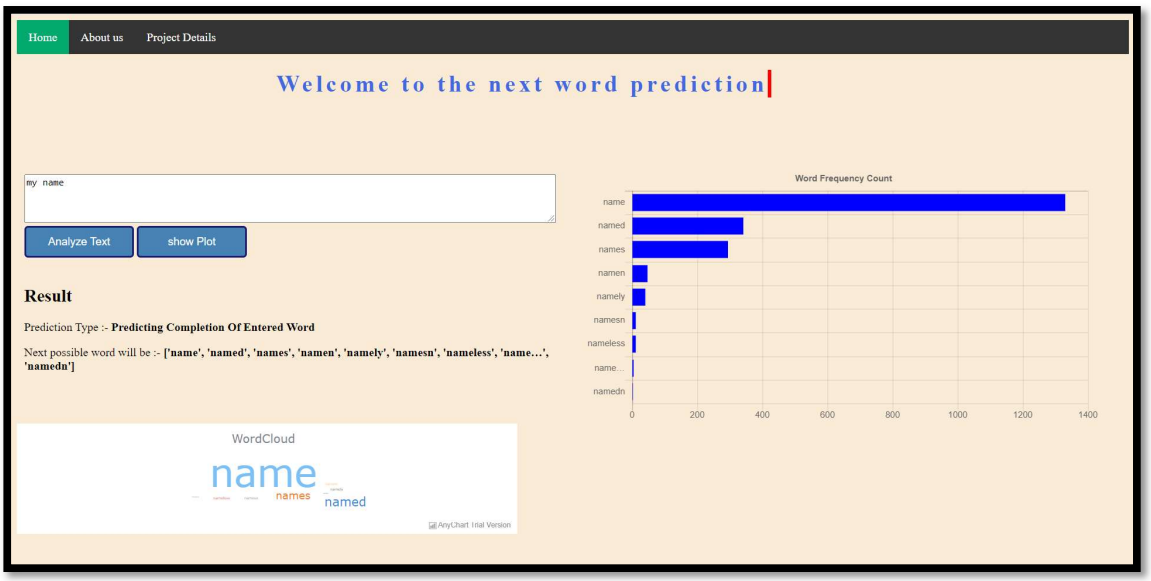



## III. Project detail page

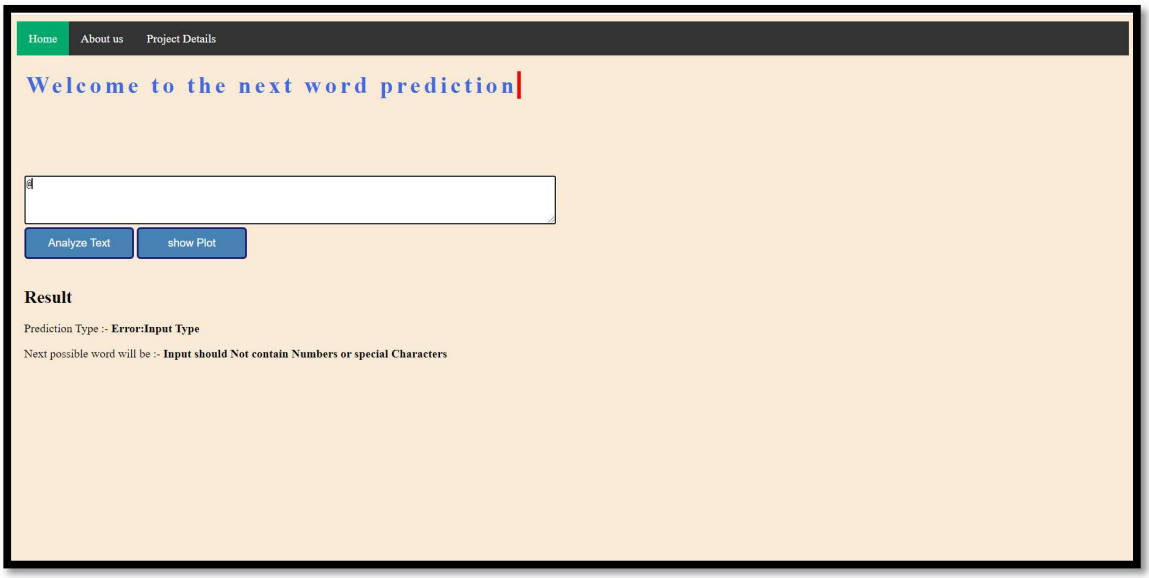
## IV. Predicting the half word




## V. Predicting the sentence.



VI. Checking with special chararcters.




VII. Checking with numbers.