**Machine Learning Project**

**Project – AutoCorrect Keyboard**

**Dataset Source – Novels from Kaggle**

**Team Members**

* Sharan – AM.EN.U4CSE20265
* Emeric Alex – AM.EN. U4CSE20223

**PROBLEM DEFINITION**

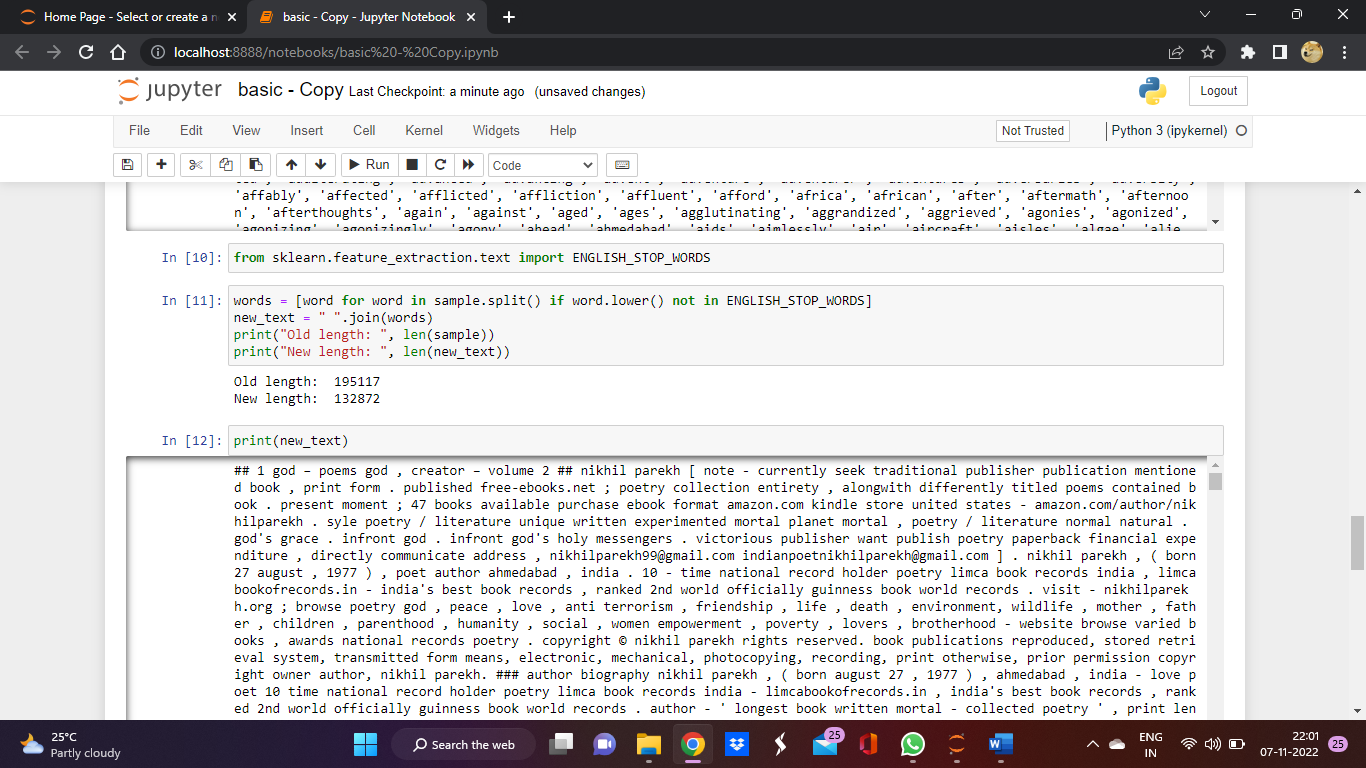
Getting frustrated of typing incorrectly and then editing all of the words after typing the entire document? Worry not! Because this tool can automatically replace words that have been typed improperly as we type without us having to go through the process of proofreading and replacing.

This not only increases our efficiency, but it also reduces the amount of effort and time we spend typing and editing.

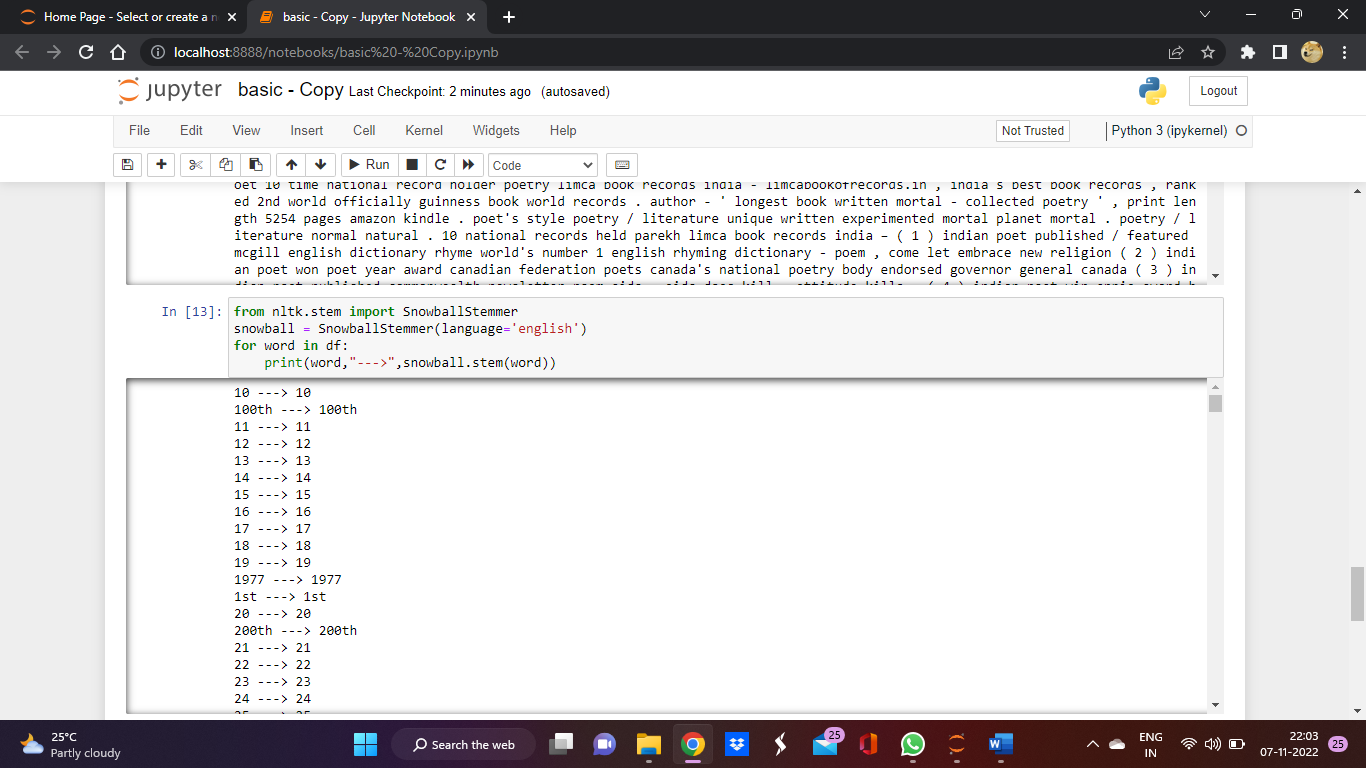
**DATASETS**

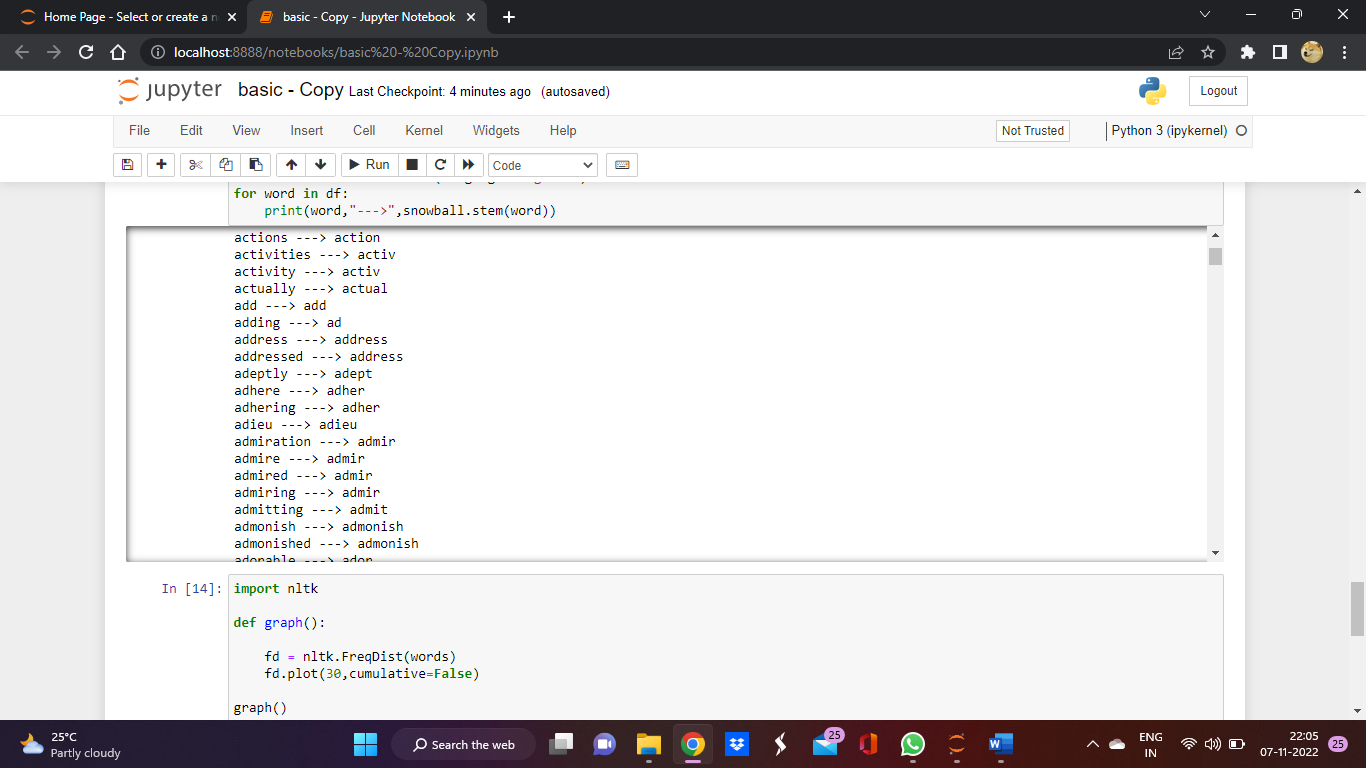
The dataset requirement for the following project is text, hence we have decided to use various novels collected from source Kaggle.com. The text datasets have been pre-processed. The pre-processing steps includes: -

* Stop-word removal

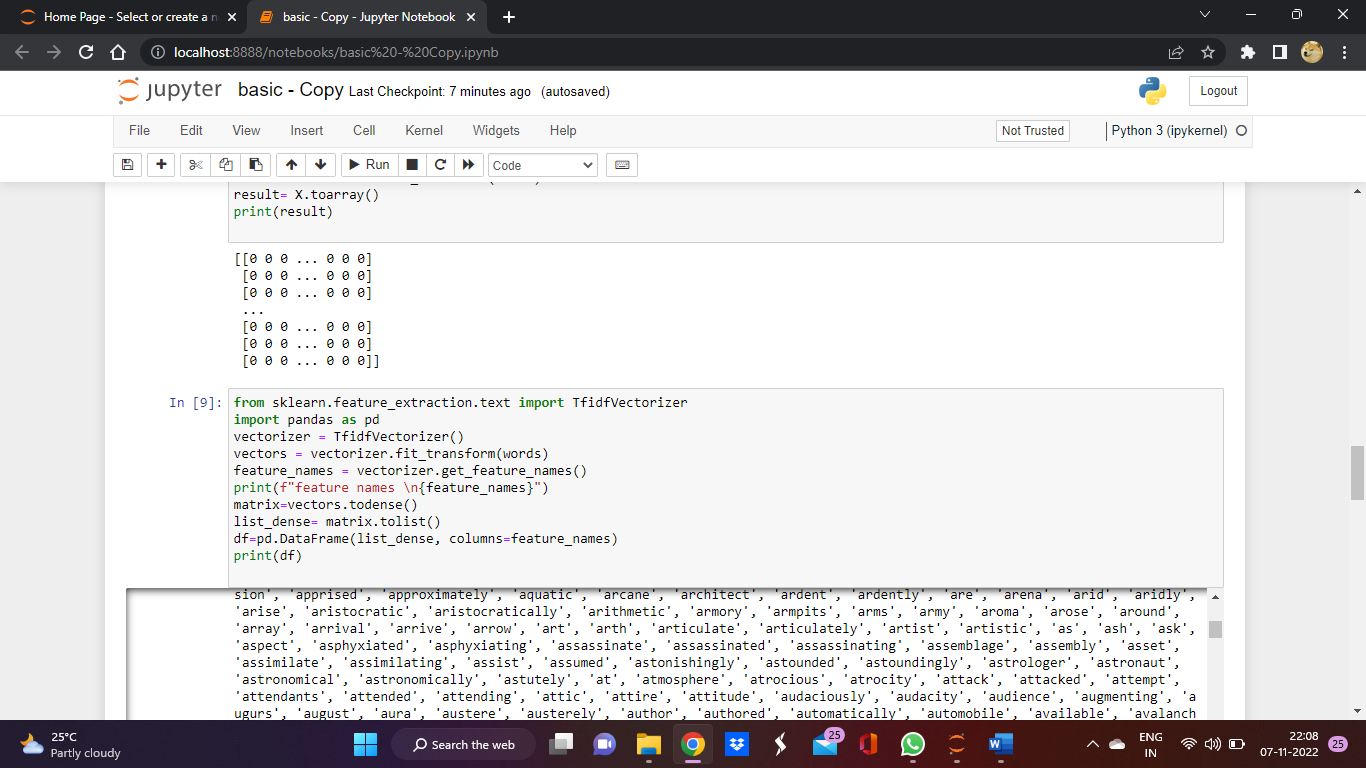


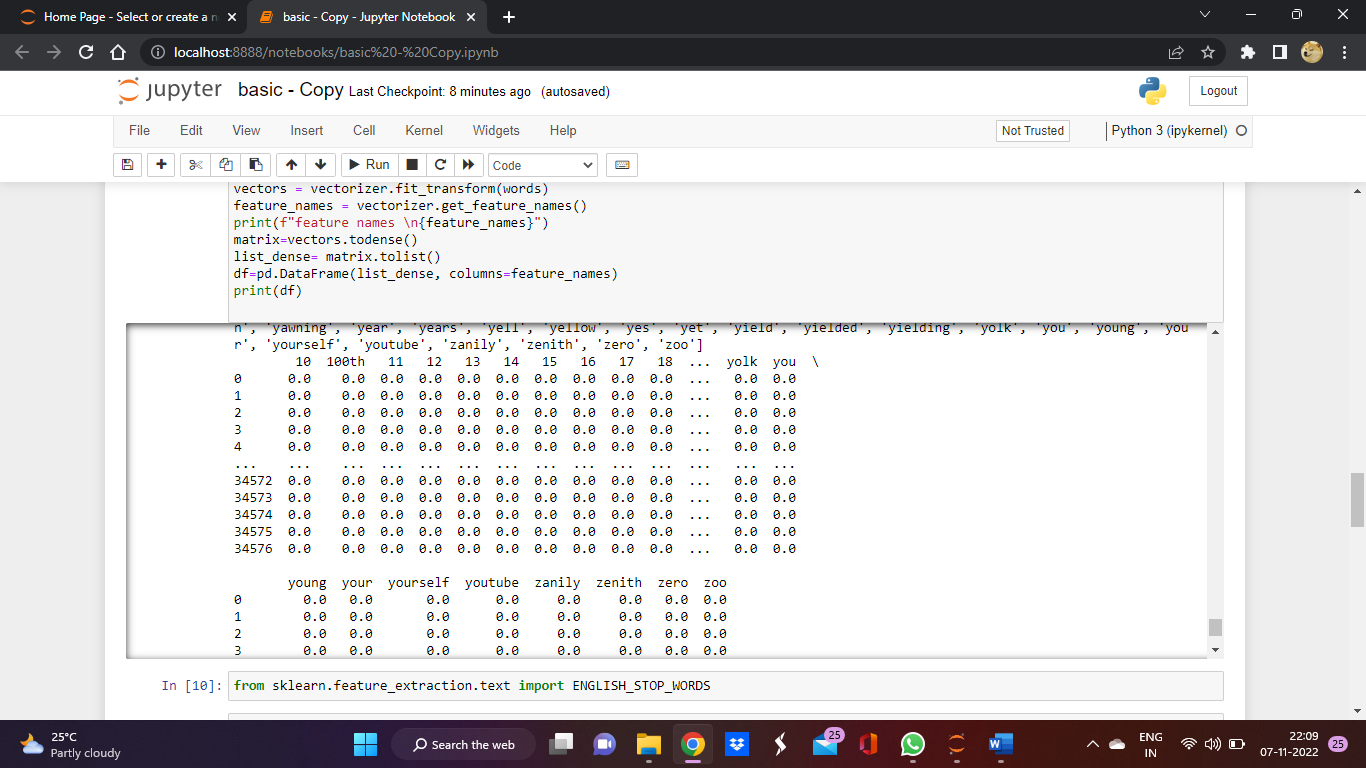
* Word stemming



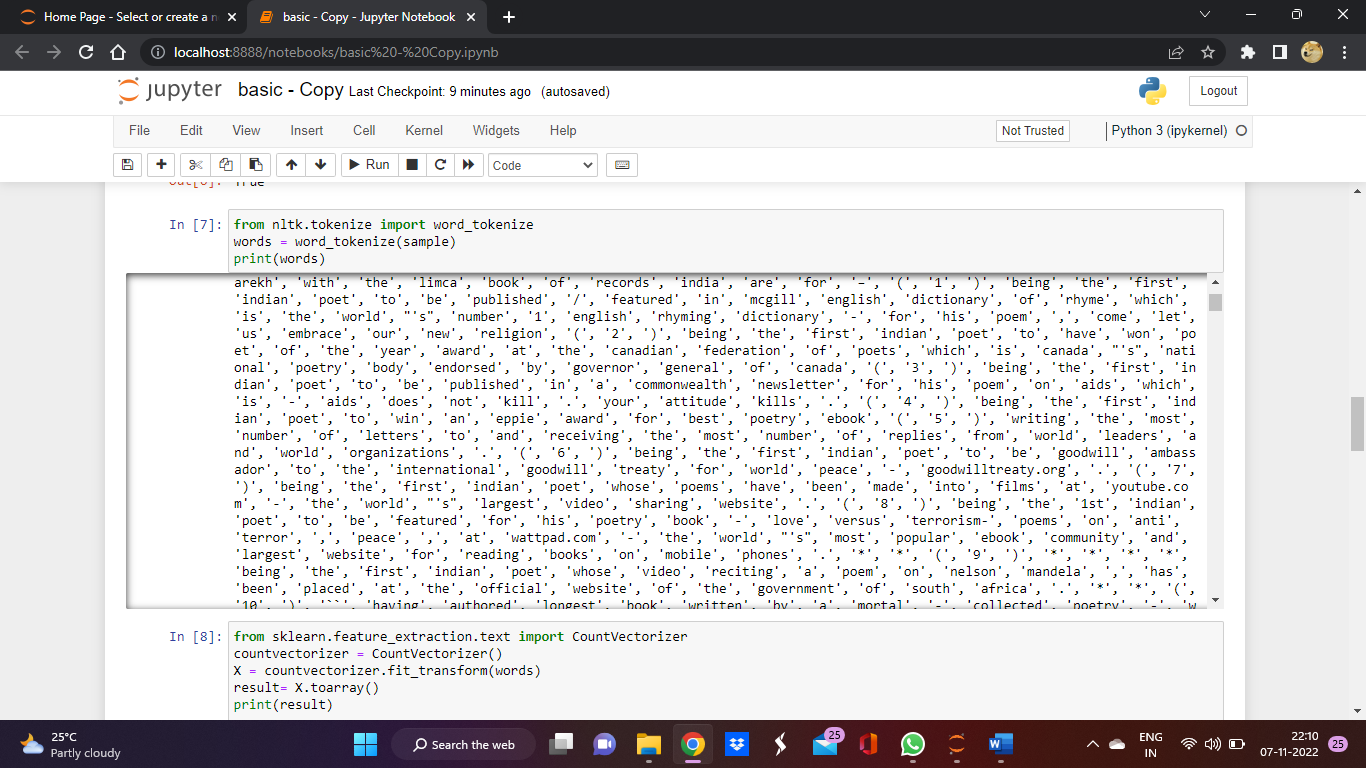


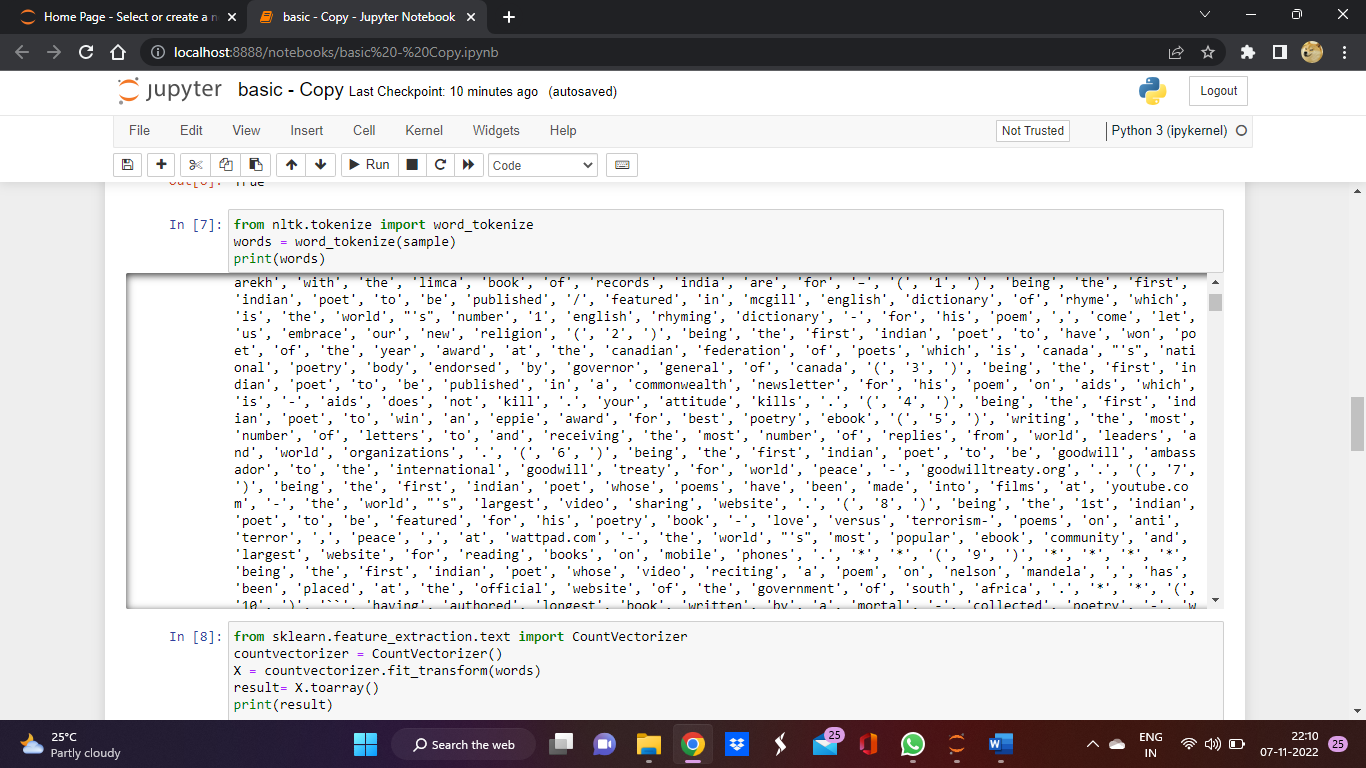
* Conversion of text to numeric (tf-id vectorization)





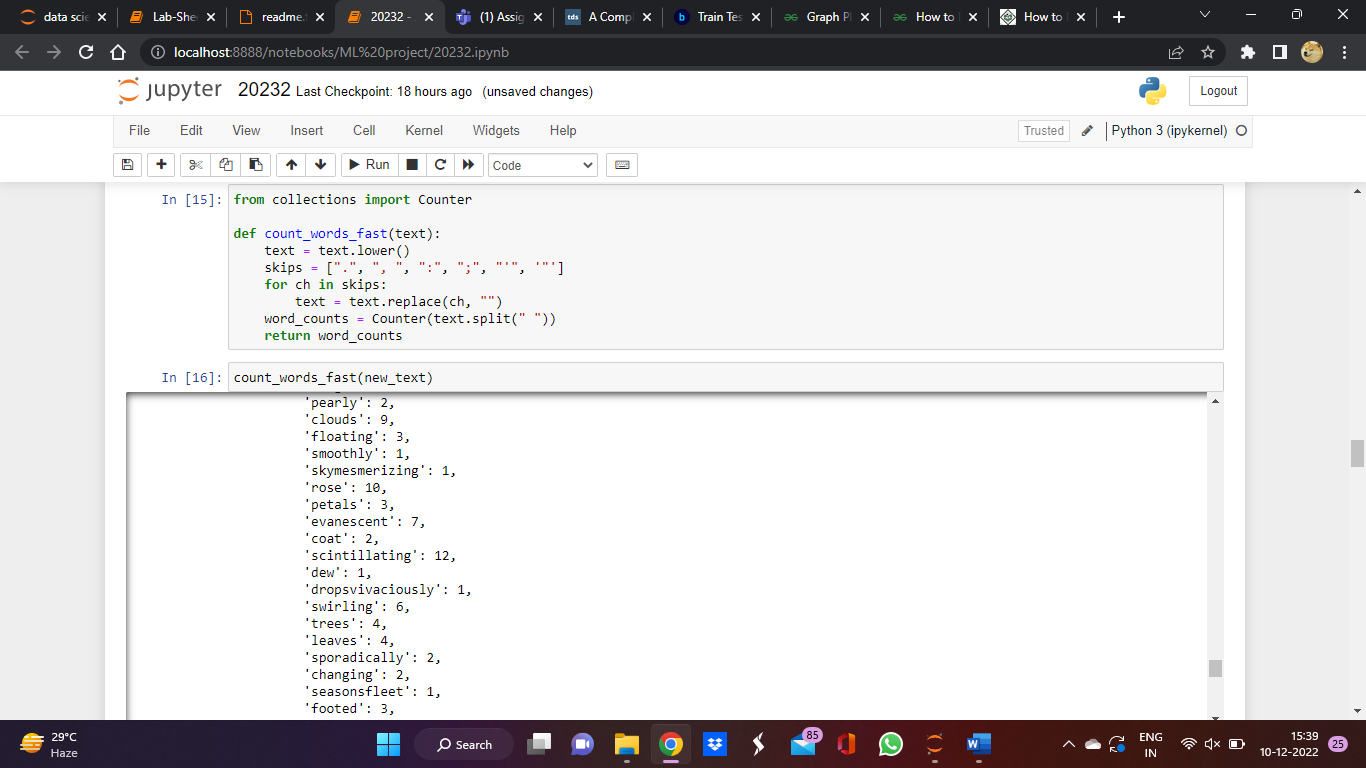
* Tokenization



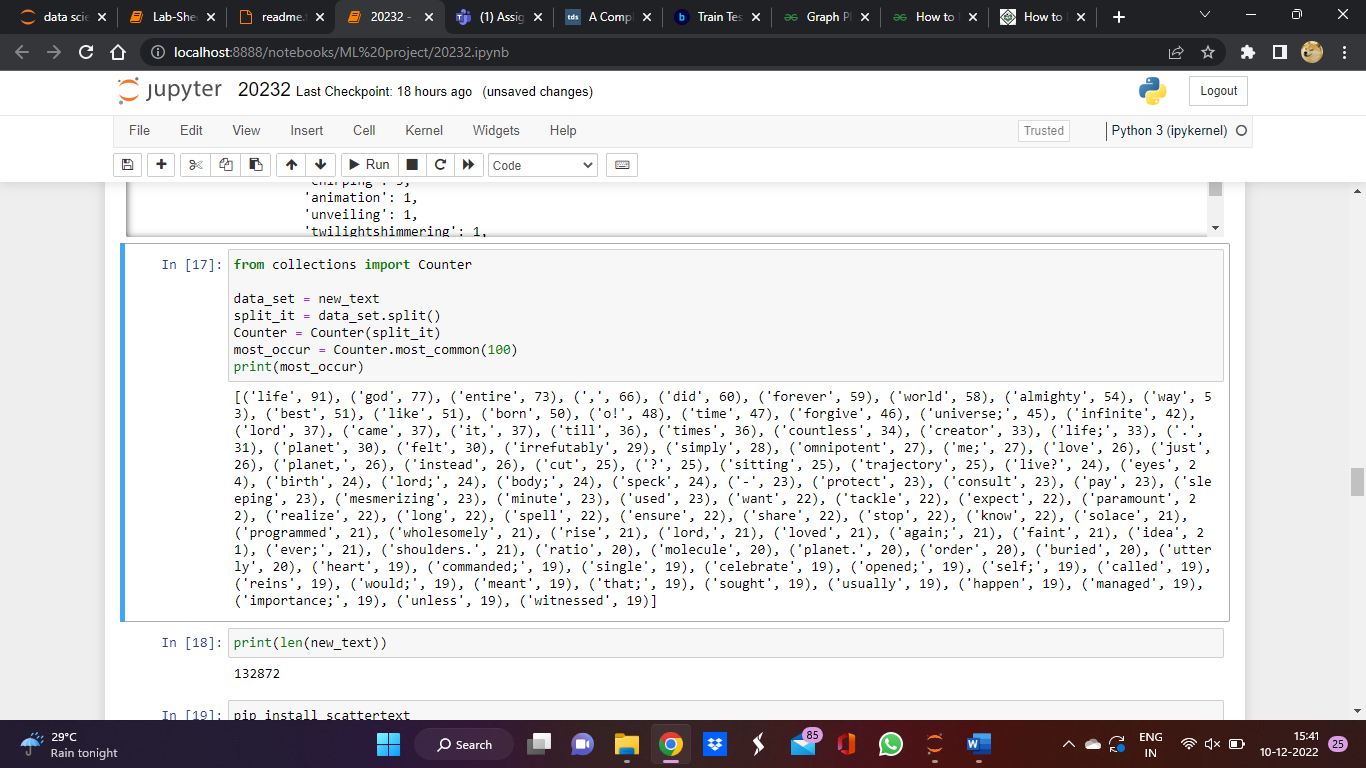


**DATA VISUALIZATION**

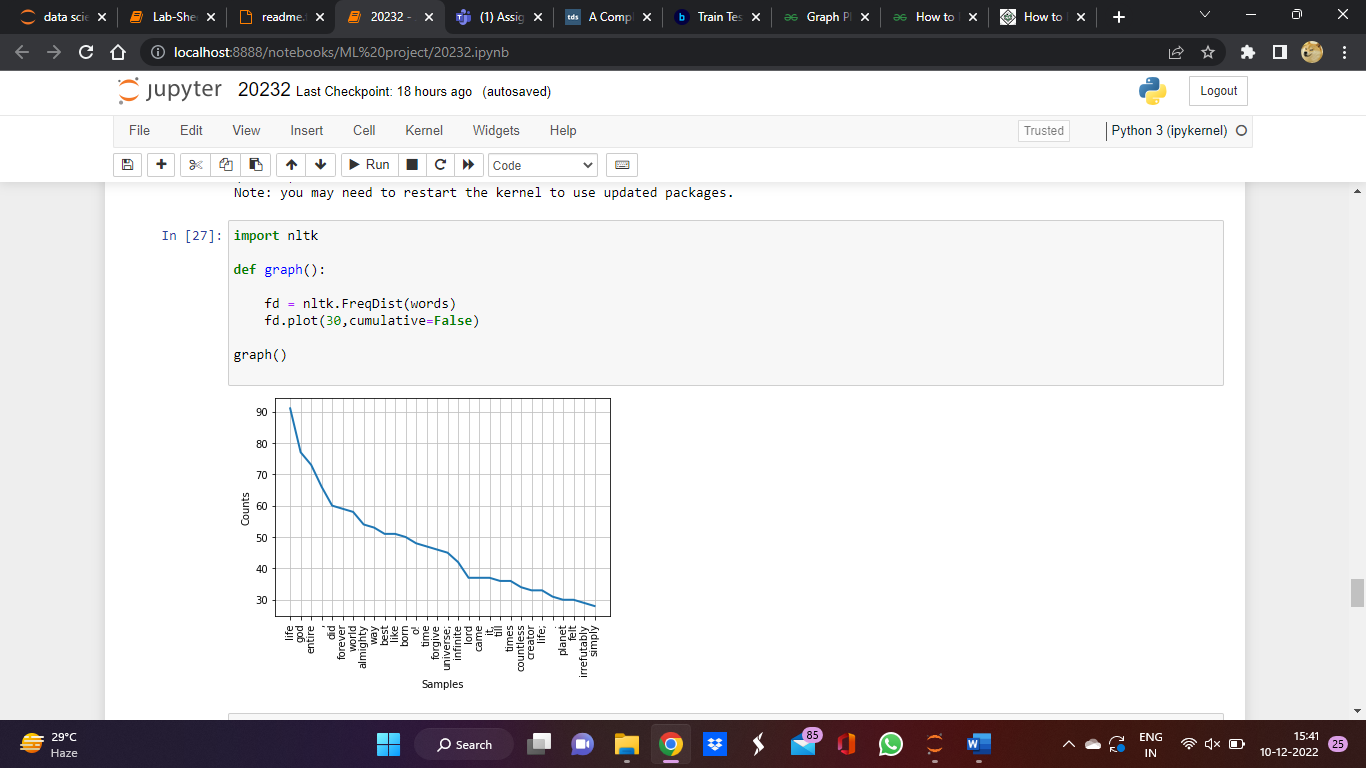
Count the number of occurrences of each word:



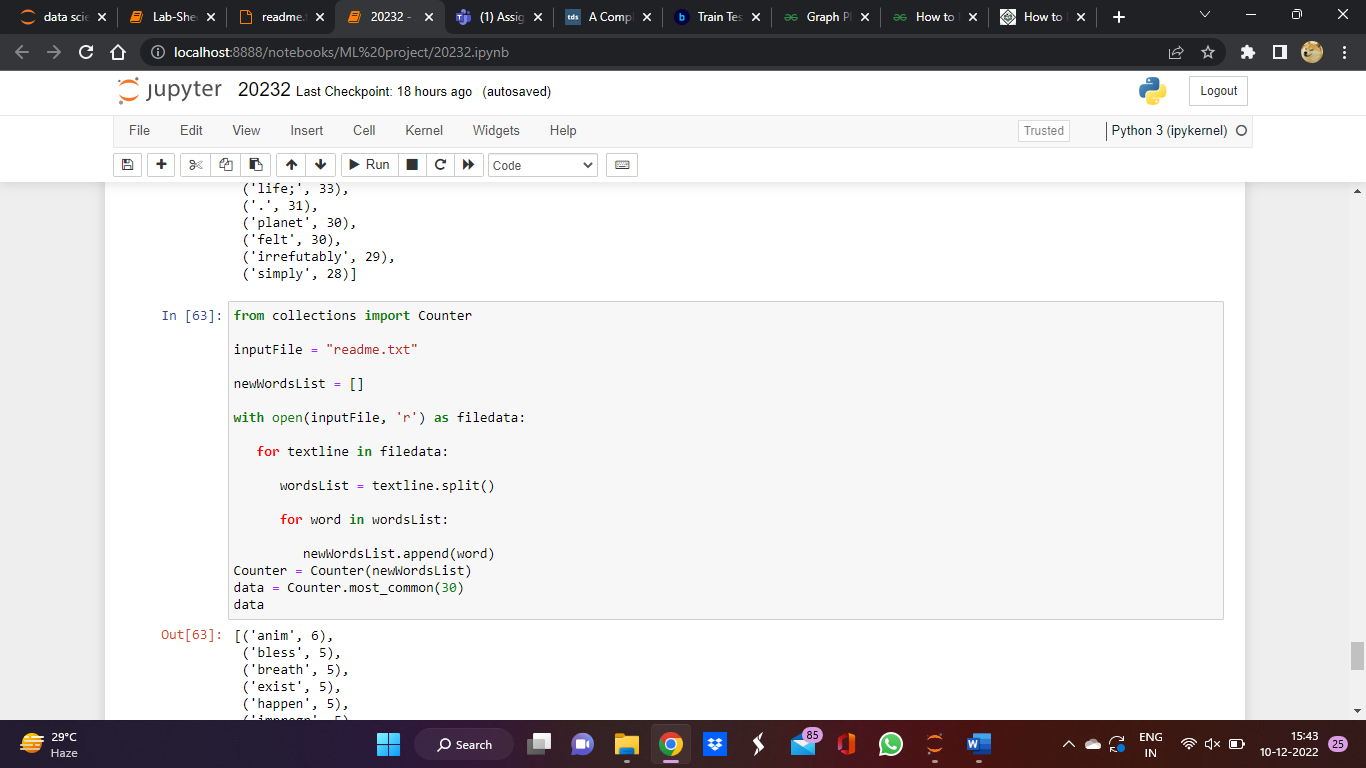
Count the most common words from the text file

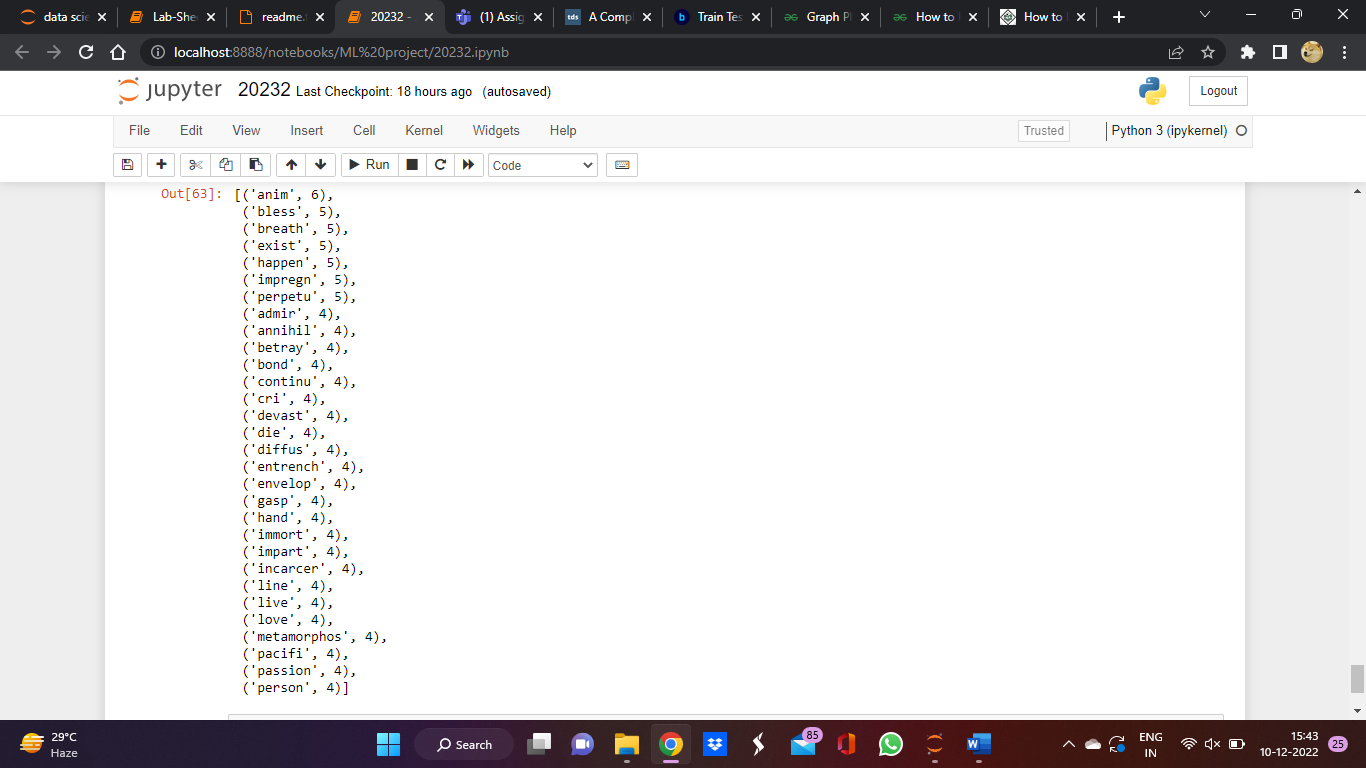


Plot the most common words



The most common stemmed words





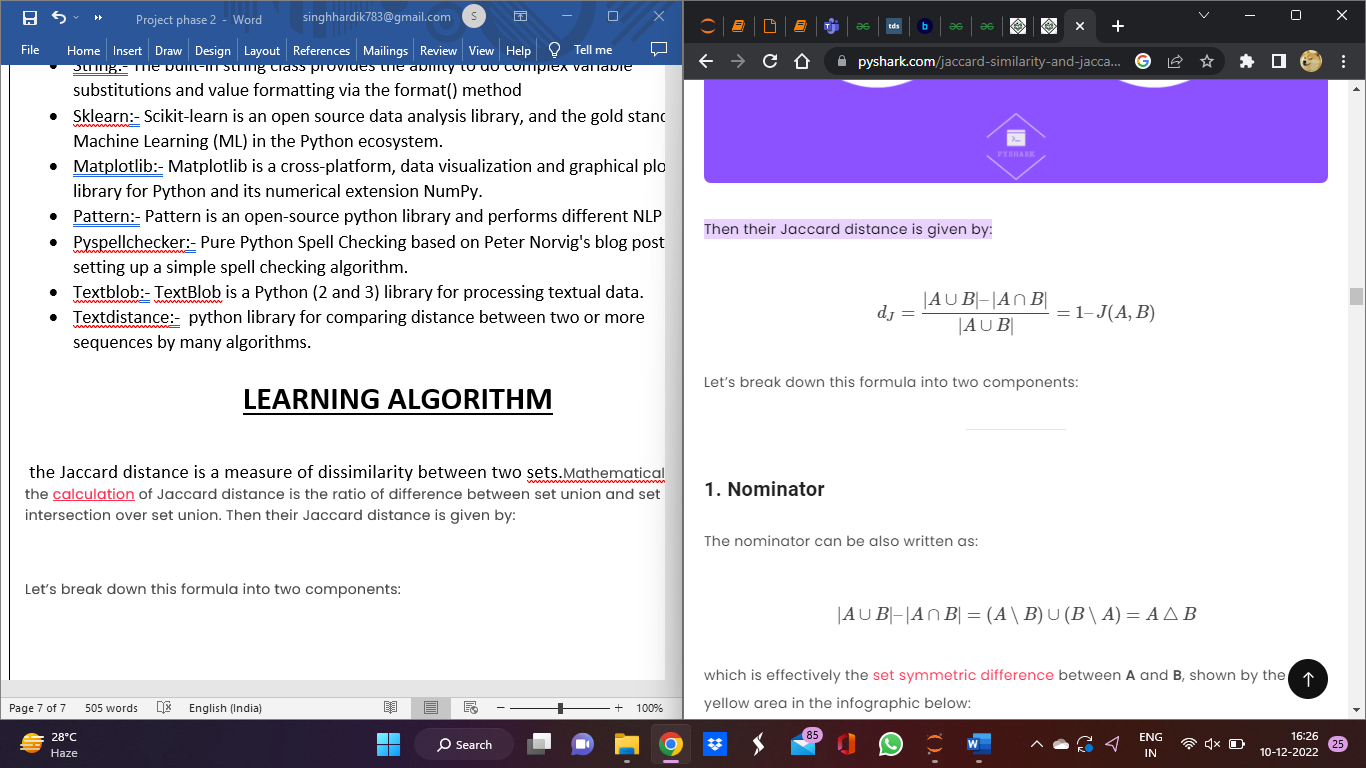
**PYTHON PACKAGES**

* Regular expression(re): - A RegEx, or Regular Expression, is a sequence of characters that forms a search pattern.
* Collections :- The collection Module in Python provides different types of containers. A Container is an object that is used to store different objects and provide a way to access the contained objects and iterate over them.
* Numpy :- NumPy is a Python library used for working with arrays.
* Pandas :- pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive.
* Natural language toolkit(nltk):- NLTK, or Natural Language Toolkit, is a Python package that you can use for NLP.
* Contractions:- A Python library for expanding and creating common English contractions in text.
* Inflect:- Correctly generate plurals, singular nouns, ordinals, indefinite articles; convert numbers to words.
* Bs4(beautiful soup 4):- Beautiful Soup is a Python library for pulling data out of HTML and XML files.
* String:- The built-in string class provides the ability to do complex variable substitutions and value formatting via the format() method
* Sklearn:- Scikit-learn is an open source data analysis library, and the gold standard for Machine Learning (ML) in the Python ecosystem.
* Matplotlib:- Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy.
* Pattern:- Pattern is an open-source python library and performs different NLP tasks.
* Pyspellchecker:- Pure Python Spell Checking based on Peter Norvig's blog post on setting up a simple spell checking algorithm.
* Textblob:- TextBlob is a Python (2 and 3) library for processing textual data.
* Textdistance:-  python library for comparing distance between two or more sequences by many algorithms.

**LEARNING ALGORITHM**

1)JACCARD SIMILARITY/DISTANCE

the Jaccard distance is a measure of dissimilarity between two [sets](https://pyshark.com/everything-about-python-set-data-structure/).Mathematically, the [calculation](https://amzn.to/3vnkzN2) of Jaccard distance is the ratio of difference between set union and set intersection over set union. Then their Jaccard distance is given by:



The [Jaccard similarity index](https://www.statology.org/jaccard-similarity/) measures the similarity between two sets of data. It can range from 0 to 1. The higher the number, the more similar the two sets of data.

The Jaccard similarity index is calculated as:

Jaccard Similarity = (number of observations in both sets) / (number in either set)

Or, written in notation form:

J(A, B) = |A∩B| / |A∪B|

This tutorial explains how to calculate Jaccard Similarity for two sets of data in Python.

2)[**Edit Distance**](https://en.wikipedia.org/wiki/Edit_distance)**, AKA**[**Levenshtein distance**](https://en.wikipedia.org/wiki/Levenshtein_distance)**.**

It’s a method is evaluating how dissimilar two strings are based on the **minimum number of operations required to transform one string into another.** Similar to before, this function will by default take in the same list of words as in recommender 1.

The function will iteratively compare entries with the list of correct words and get their respective edit distances. Then, the word with the lowest distance will be deemed as the best matching word, appended to outcomes, and returned by the function.

**Natural Language Processing (NLP)**

Natural Language Processing (NLP) is a branch of Artificial Intelligence that allows computers to understand and process natural human language.

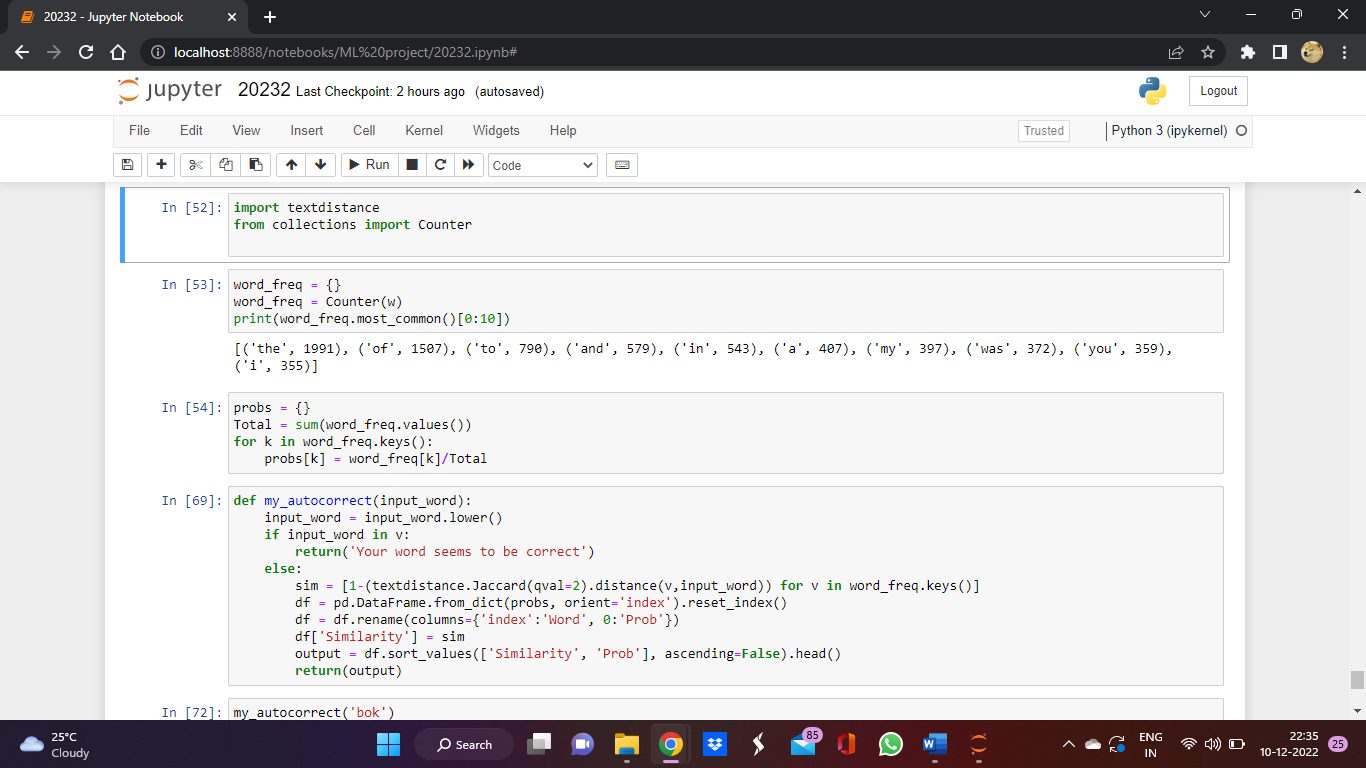
NLP uses a programming language that enables computers to evaluate and interpret large volumes of natural language data.

It paves the door for more interactivity and productivity in a variety of fields, like:

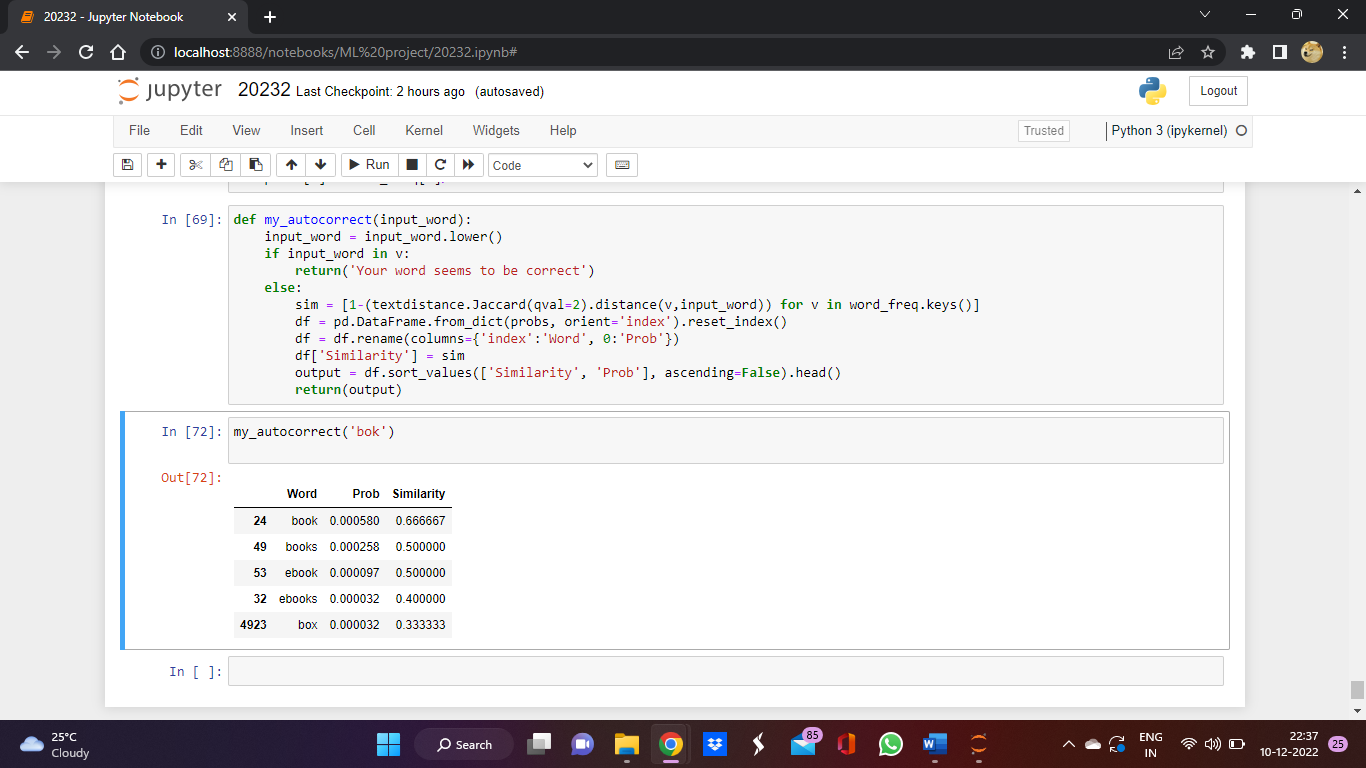
* Search autocorrect and autocomplete.
* Language translation and grammar checkers.
* Chatbots and social media monitoring.
* Email filtering and voice assistants.

We’ll look at how it’s employed in auto-correction systems in this project.

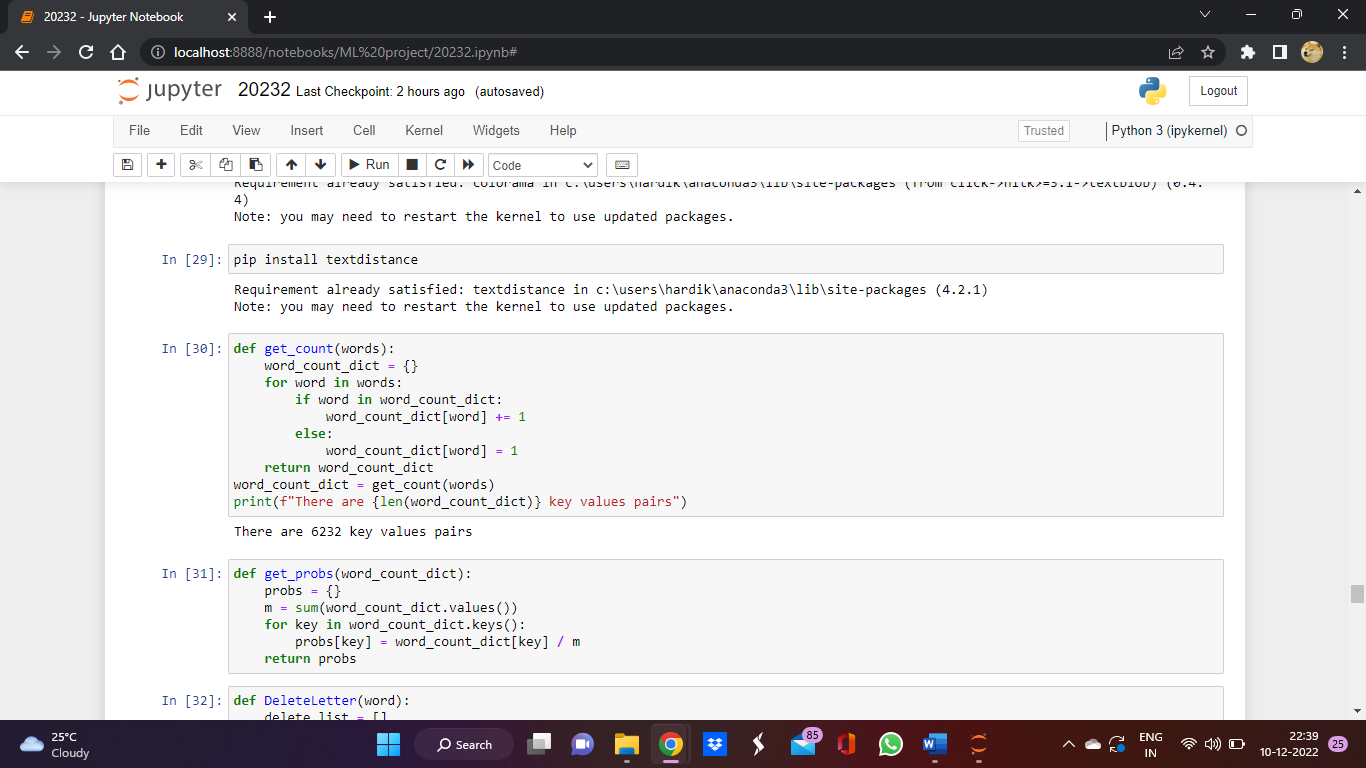
**JACCARD MODEL**

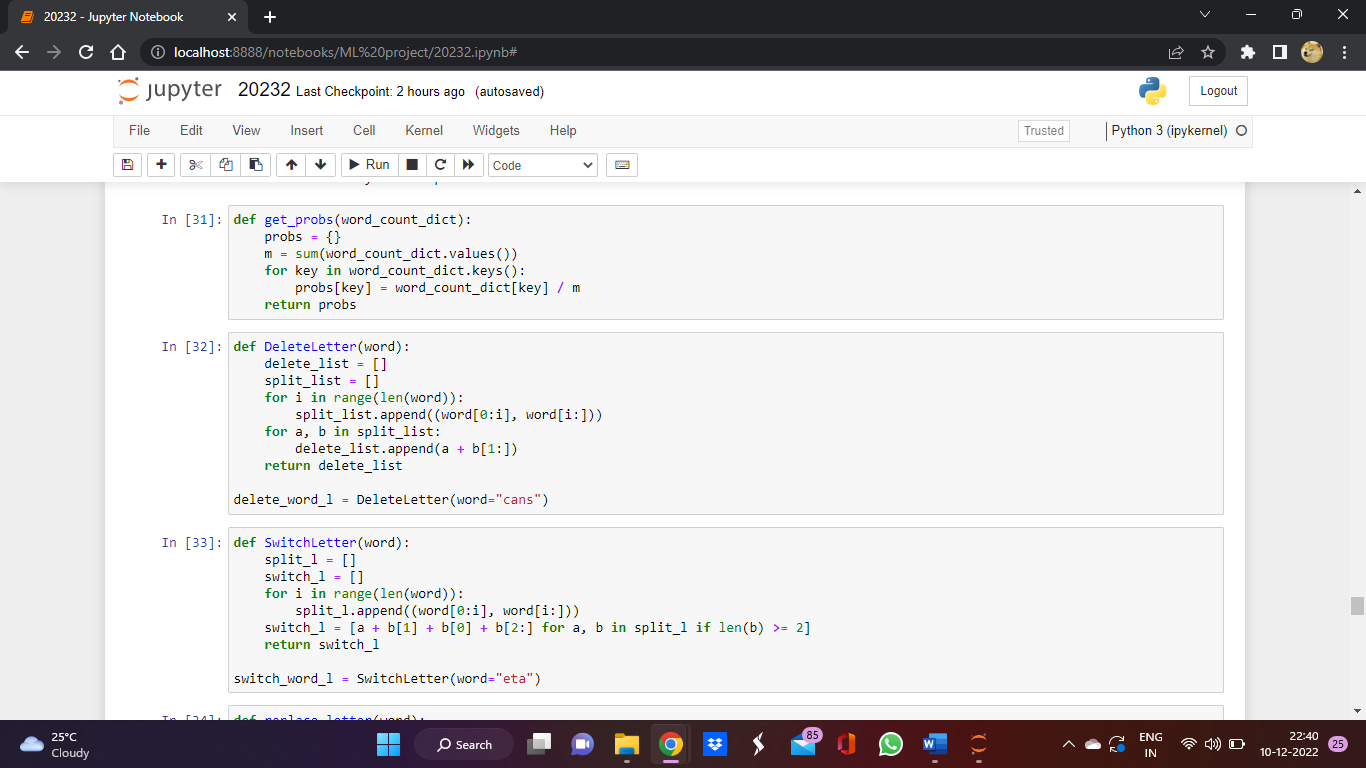


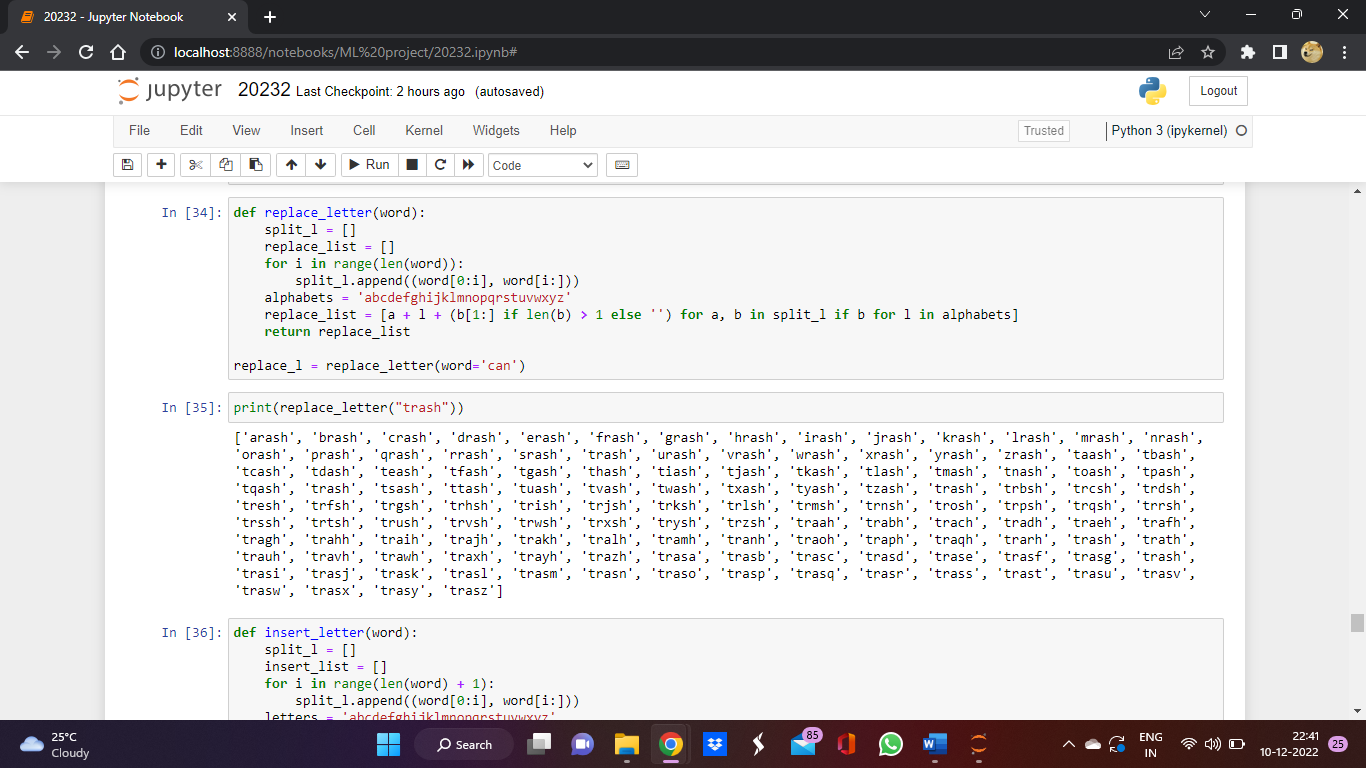
**OUTPUT**

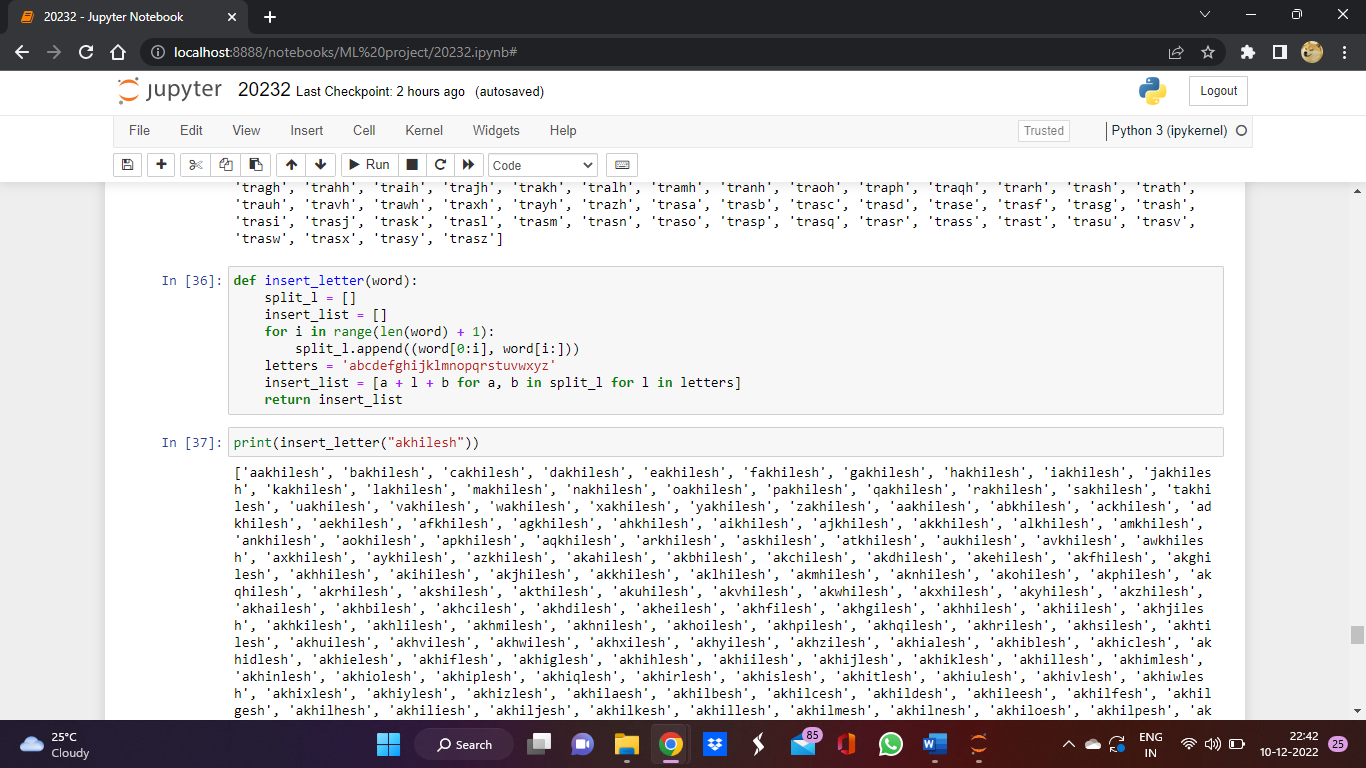


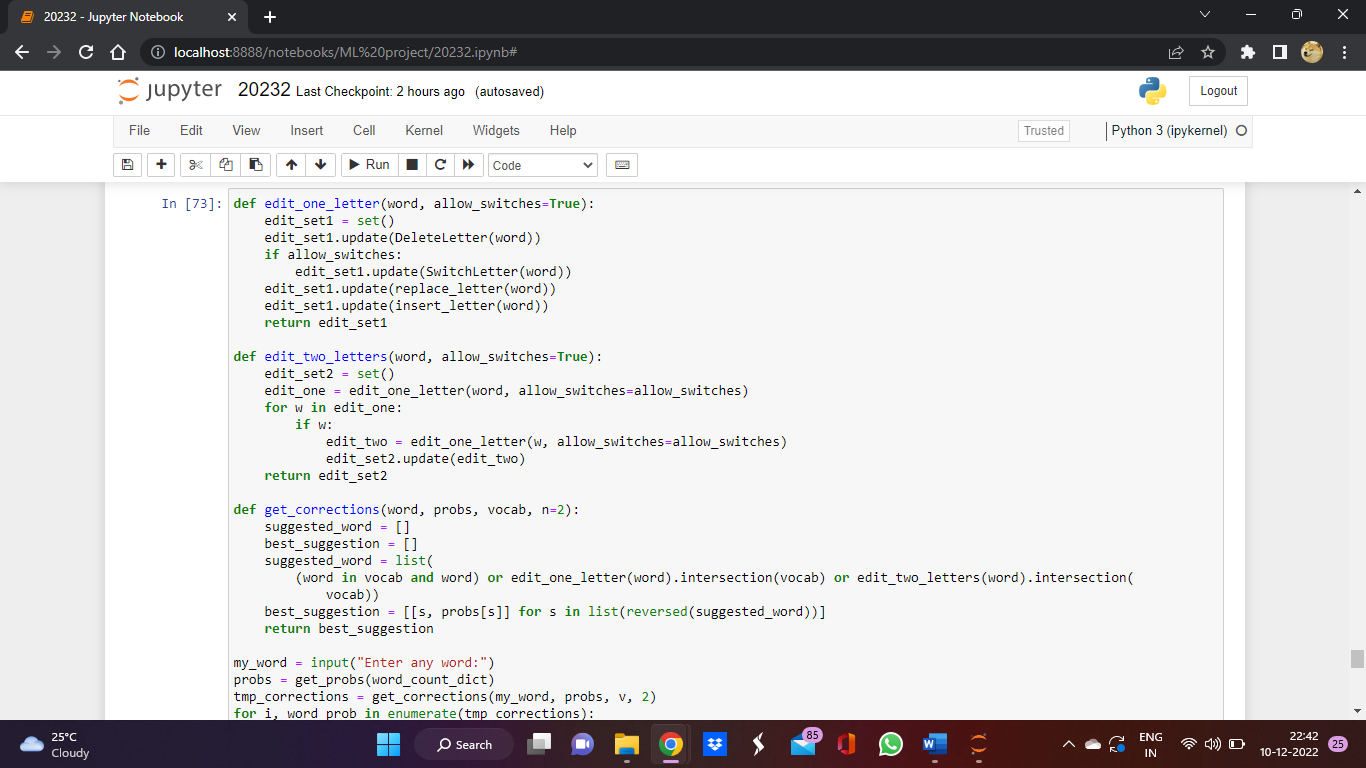
**LEVENSHTEIN MODEL**

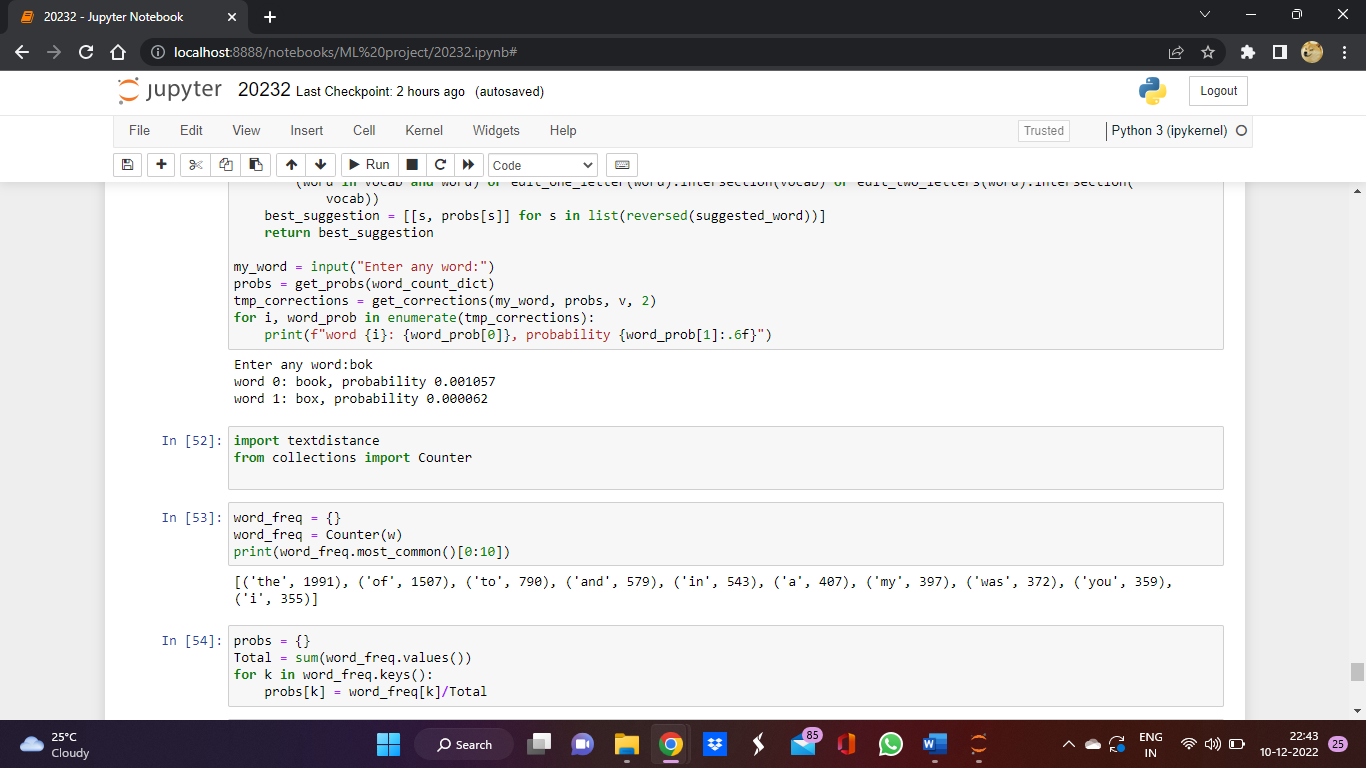




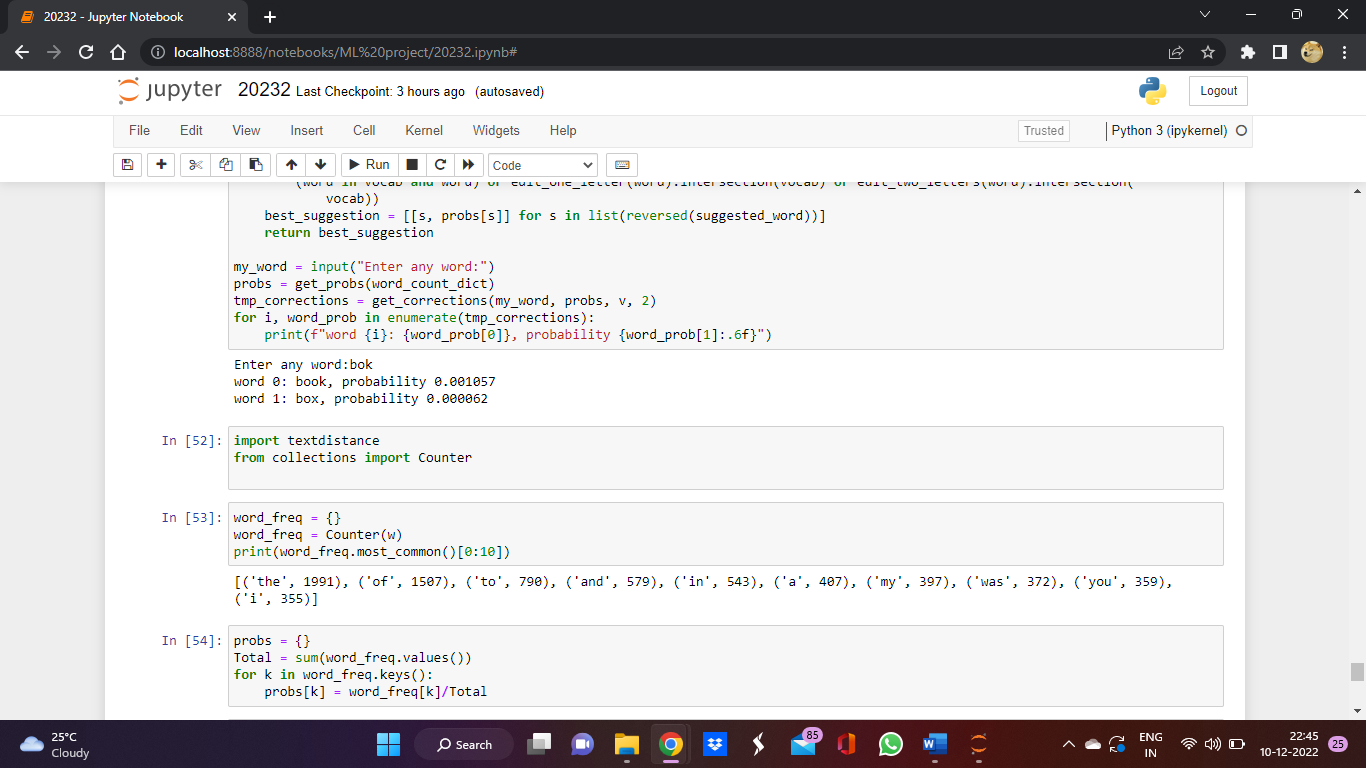








OUTPUT



**Measure of accuracy can be judged by the Measure of probability in predictions**