**Steam Review Analysis**

Abstract:

Sentiment analysis is one of the major topics in natural language processing. The following process will provide the complete process of sentiment analysis of steam using the Natural Language Toolkit (nltk). The process consists of preparing the data followed by removing useless data like punctuations and symbols along with the removal of stopwords. The processed data is then tokenized where the data is converted into tokens which allows us to compare different words in the reviews. Stemming is done after tokenization to merge similar words in the data. At last, we calculate TF, IDF, and TFIDF for the following data. All these steps allow us to analyze the data to identify whether the game has overall positive or negative feedback.

Data Collection:

The dataset which is used in the project is ‘Steam Game Review Dataset’. Which is opensource

dataset which can be download from the Kaggle. The dataset contains reviews of different games taken in steam. Steps To Follow Before Data Preprocessing and Feature Extraction

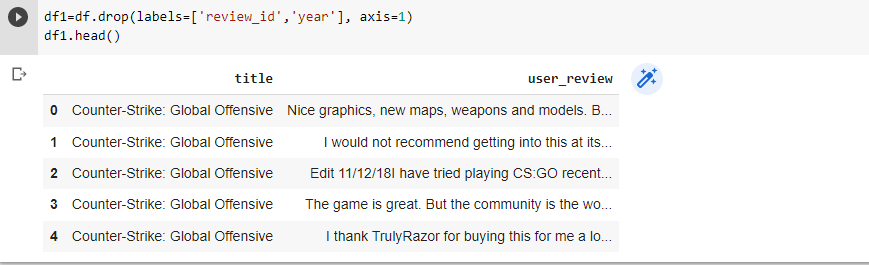
1.Signup for Kaggle to get access to dataset

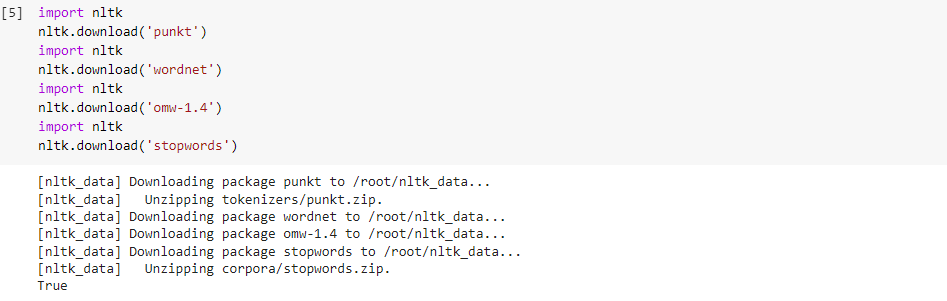
2.Follow the below link to download the dataset:

https://www.kaggle.com/datasets/arashnic/game-review-dataset

3.After downloading the dataset extract it and upload it to drive.

4.Mount the drive and import dataset for further process.





**DATA PREPROCESSING AND FEATURE EXTRACTION**

The below Functions are used to remove stopwords and symbols. The function is called bellow the definition.

**Tokenization and stemming**

After Removal of stop words, we have processed the text by tokenization the sentences and then

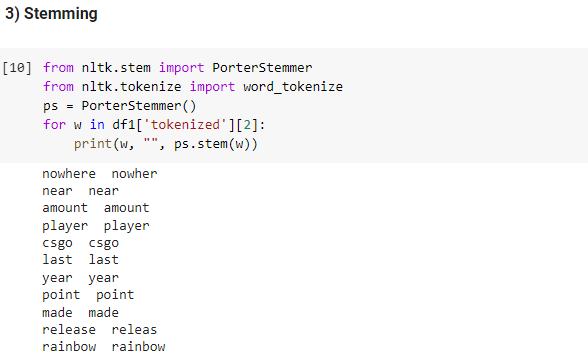
used stemming process for further Pre-processing.

Tokenization is the process of dividing text into a set of meaningful pieces. These pieces are called

tokens. Stemming is a process where words are reduced to a root by removing inflection through

dropping unnecessary characters, usually a suffix





**FEATURE EXTRACTION TF AND IDF**

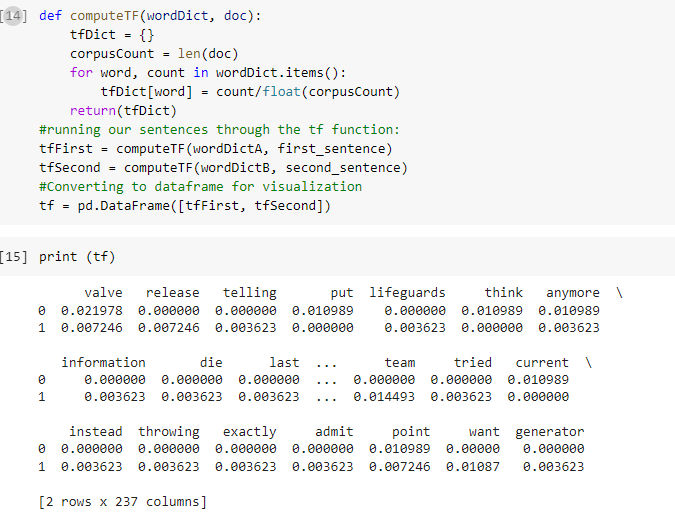
TF-IDF stands for Term Frequency Inverse Document Frequency of records. It can be defined as the

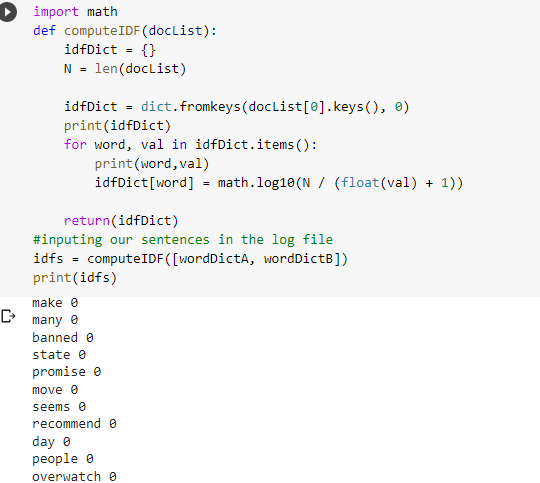
calculation of how relevant a word in a series or corpus is to a text. The meaning increases

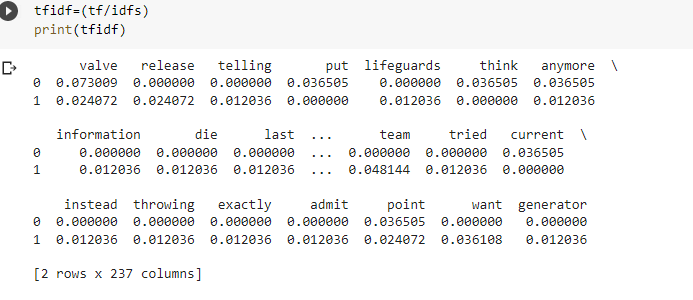
proportionally to the number of times in the text a word appears but is compensated by the word

frequency in the corpus.









Architecture:

