**CS5542 Big Data Analytics and Apps**

**2019 Spring**

**Lab Assignment 1**

**By:**

**Pavankumar Manchala**

**Class Id: 16**

**Objective:**

The objective of the Lab Assignment 1 is to complete following task:

1. Download the Dataset related to project theme form list of Project Datasets in the Google Spreadsheet.
2. Performing the NLP operations (Tokenization and Lemmatization) on the data.
3. Reporting the image statistics from the caption dataset extracted from downloaded dataset.
4. Applying SIFT algorithm on the images extracted to extract features.

**Platform used:**

PyCharm IDE

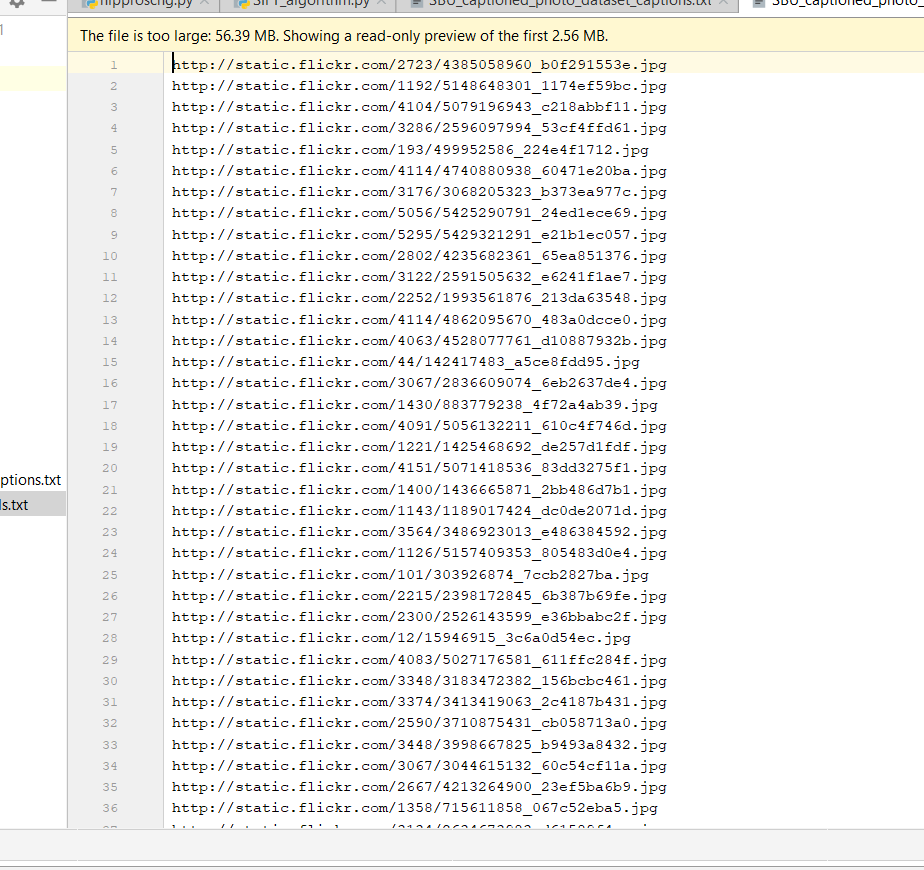
**Packages installed:**

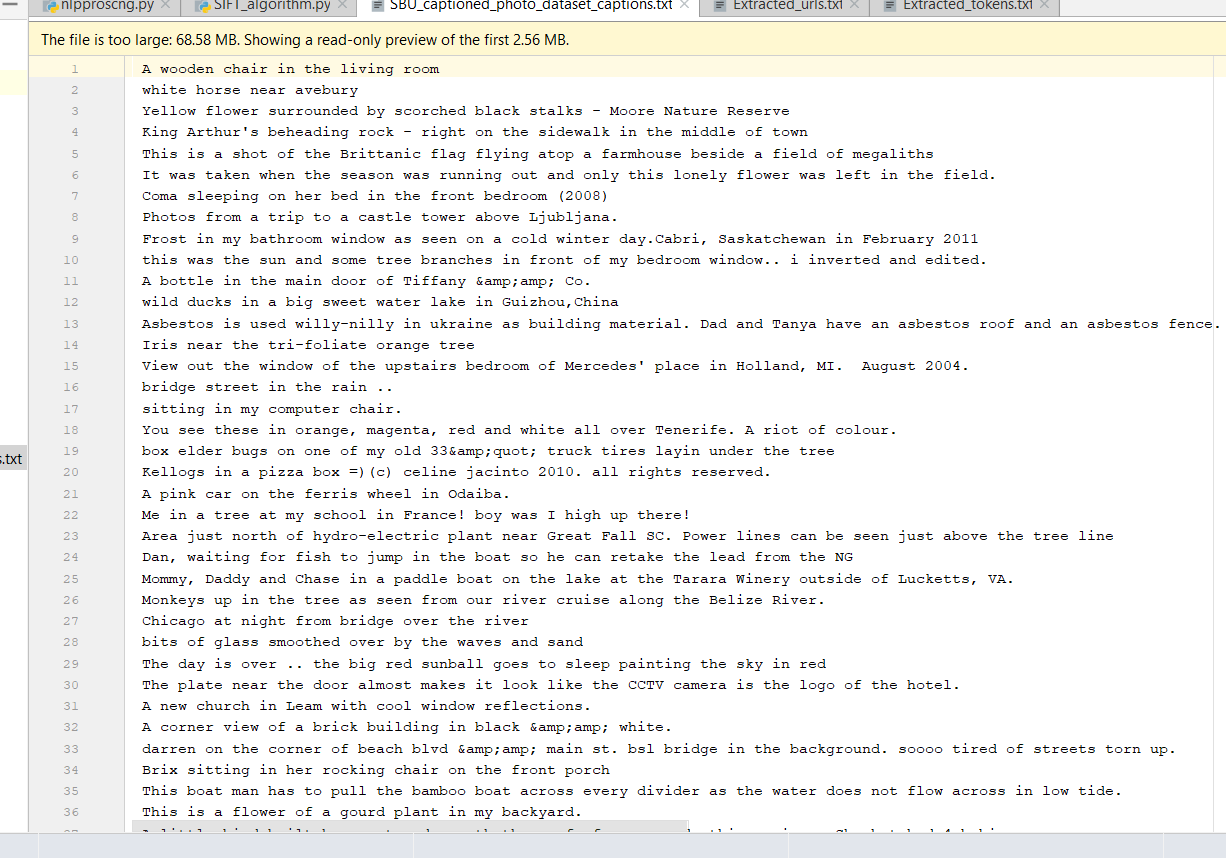
1. nltk
2. numpy
3. matplotlib
4. opencv-python
5. Tensorflow

**Dataset:**

There are different datasets available online and the dataset used is SBU Image Captions dataset which consists of image URLs and Captions for them.

The dataset image URLs and Captions are extracted in txt files.

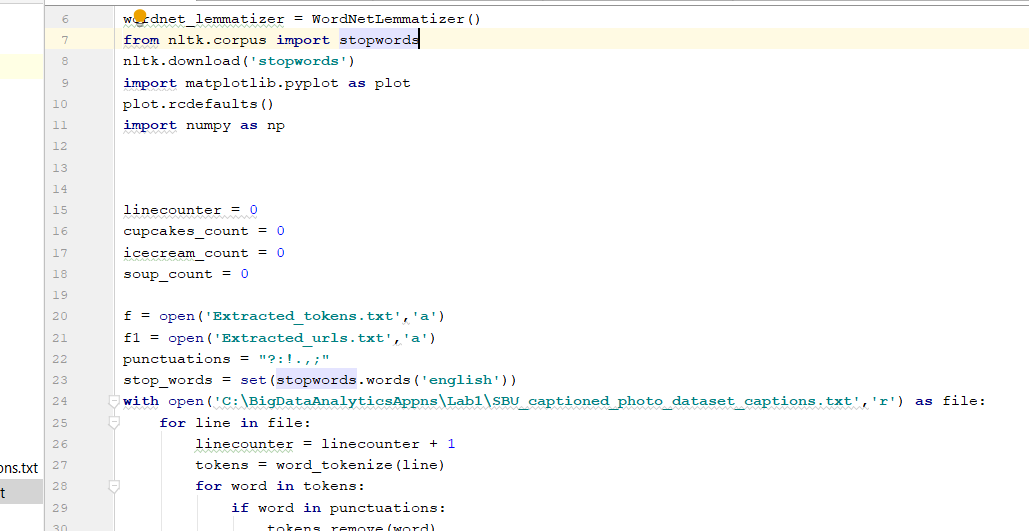
****

****

**NLP** **operations**:

1. **Tokenization:**

* Tokenization is performed over the Captions data to make either sentence separation or word separation on each image caption.
* The Sentence separation is done with sentence\_tokenize and word separation by word\_tokenize.
* We used word\_tokenize for extracting different keywords for extracting different datasets.
* The tokenization makes to extract captions related to project and stores in new text file by using line cache. Both the image URLs corresponding to the Captions are saved in respective files.

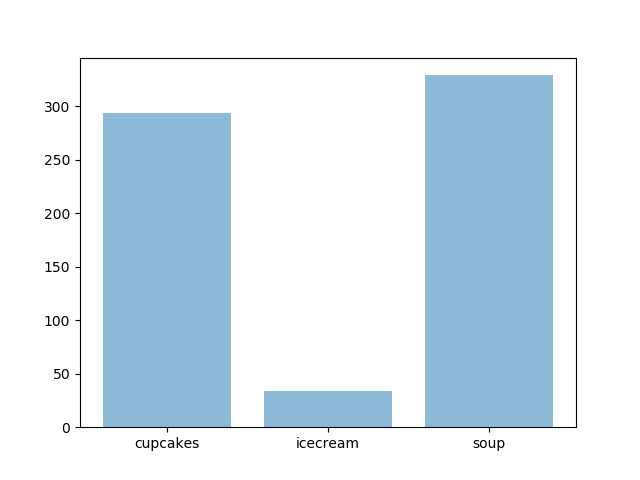


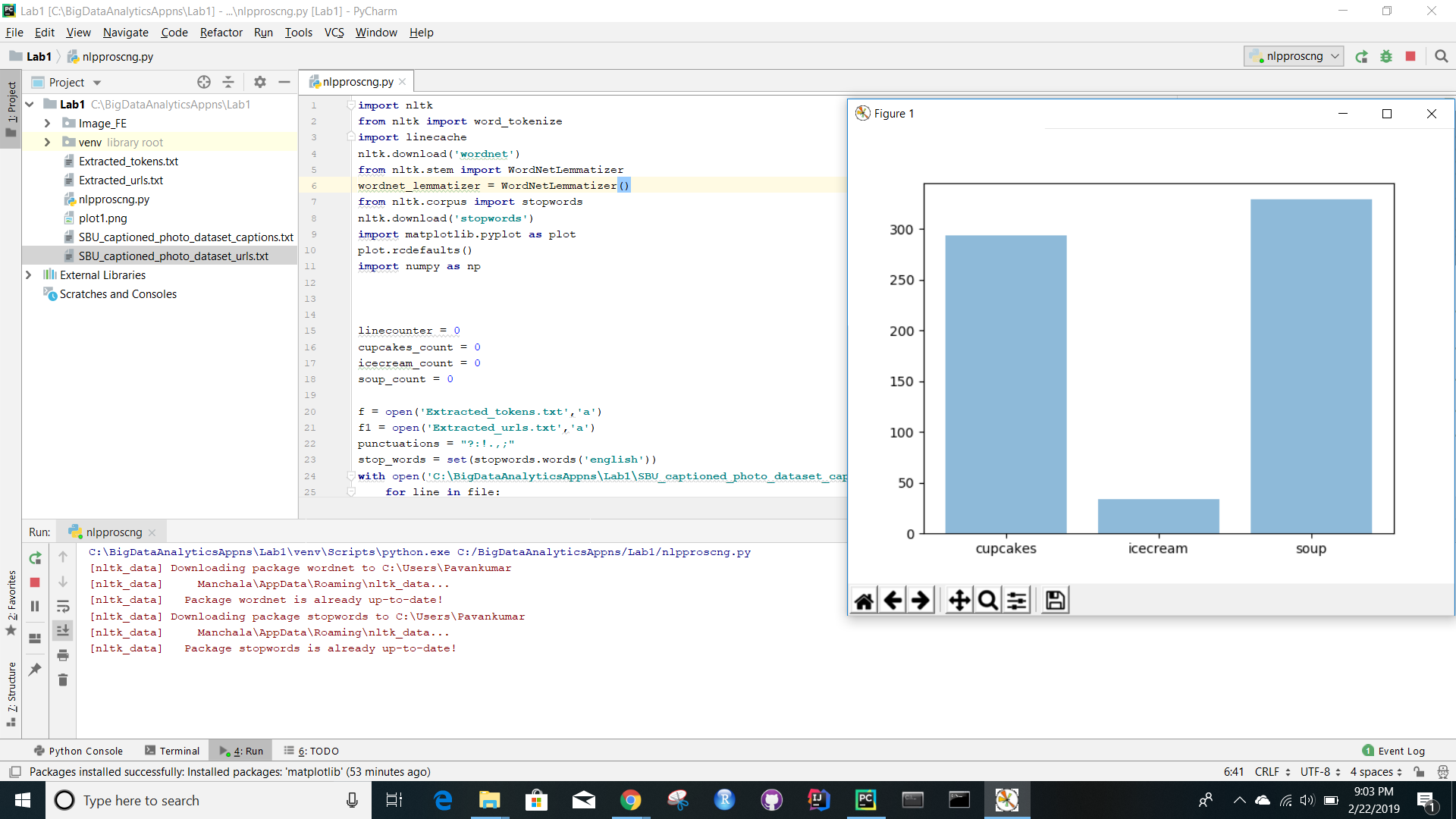
1. **Lemmatization:**

* Lemmatization is performed to get the root word from the given keyword helps in pre-processing the data.
* The stop words are eliminated after performing the Lemmatization.



A word count used to track the count of keywords and used for plotting them.





**SIFT Algorithm:**

The SIFT algorithm is used to extract the image features.

The output of the SIFT algorithm on given input images is shown.

**Input Images:**

**Output Images:**

