

## **CS5542 – Bigdata Analytics and Applications**

### **Lab report – 1**

#### **Submitted by:**

Saitejaswi Koppuravuri – 13

Team – 5

#### **Objectives:**

The objectives of the Lab assignment 1:

- Downloading the dataset according to the idea of the project proposal.
- Performing tokenization and lemmatization on the caption or text data.
- Reporting the image statistics from the caption data extracted.
- Perform feature extraction using SIFT algorithm on the image data extracted.

#### **Technologies:**

Pycharm – IDE for executing the python files

#### **Packages used:**

- nltk
- opencv-python
- numpy
- matplotlib
- Tensorflow

#### **Dataset :**

There are many datasets available out of which an SBU dataset is chosen, which gives the data in the form of two separate files “URL’s” and the “captions” respectively.

The following screenshots describe the dataset URL’s and their corresponding captions.

```
test [C:\Users\Koppu\PycharmProjects\test] - \SBU_captioned_photo_dataset_captions.txt [test] - PyCharm
File Edit View Navigate Code Refactor Run Tools VCS Window Help
test SBU_captioned_photo_dataset_captions.txt SIFT_algorithm.py
Project: test C:\Users\Koppu\PycharmProjects\test
venv library root
building_1.jpg
building_2.JPG
extract.py
Extracted_tokens.txt
Extracted_urls.txt
filteredtext.txt
google.ico
sample.ico
SBU_captioned_photo_c
SBU_captioned_photo_c
SIFT_algorithm.py
test1.py
test_lemmit.py
word.py
External Libraries
Scratches and Consoles
Run: SIFT_algorithm.py
C:\Users\Koppu\PycharmProjects\test\venv\Scripts\python.exe C:\Users\Koppu\PycharmProjects\test\SIFT_algorithm.py
[ INFO:0] Initialize OpenCL runtime...
Process finished with exit code 0
Packages installed successfully. Installed packages: 'opencv-contrib-python==3.4.0.12' (today 9:15 PM)
41:55 CRLF UTF-8 4 spaces 2/22/2019
```

```
test [C:\Users\Koppu\PycharmProjects\test] - \SBU_captioned_photo_dataset_urls.txt [test] - PyCharm
File Edit View Navigate Code Refactor Run Tools VCS Window Help
test SBU_captioned_photo_dataset_urls.txt SIFT_algorithm.py
Project: test C:\Users\Koppu\PycharmProjects\test
venv library root
building_1.jpg
building_2.JPG
extract.py
Extracted_tokens.txt
Extracted_urls.txt
filteredtext.txt
google.ico
sample.ico
SBU_captioned_photo_c
SBU_captioned_photo_c
SIFT_algorithm.py
test1.py
test_lemmit.py
word.py
External Libraries
Scratches and Consoles
Run: SIFT_algorithm.py
C:\Users\Koppu\PycharmProjects\test\venv\Scripts\python.exe C:\Users\Koppu\PycharmProjects\test\SIFT_algorithm.py
[ INFO:0] Initialize OpenCL runtime...
Process finished with exit code 0
Packages installed successfully. Installed packages: 'opencv-contrib-python==3.4.0.12' (today 9:15 PM)
12:56 CRLF UTF-8 4 spaces 2/22/2019
```

## Tokenization and Lemmatization:

- Tokenization generally splits the sentence or a paragraph and splits them into words and store them in the form of list.
- It can be done in two ways namely word tokenization(word\_tokenize) and sentence tokenization(sent\_tokenize).

- We have used word tokenization in order to extract the different keywords required to extract the data from the entire dataset.
- By tokenizing the data the extracted captions are stored in the new file and by using the line cache the corresponding urls are retrieved and stored in another file.

```

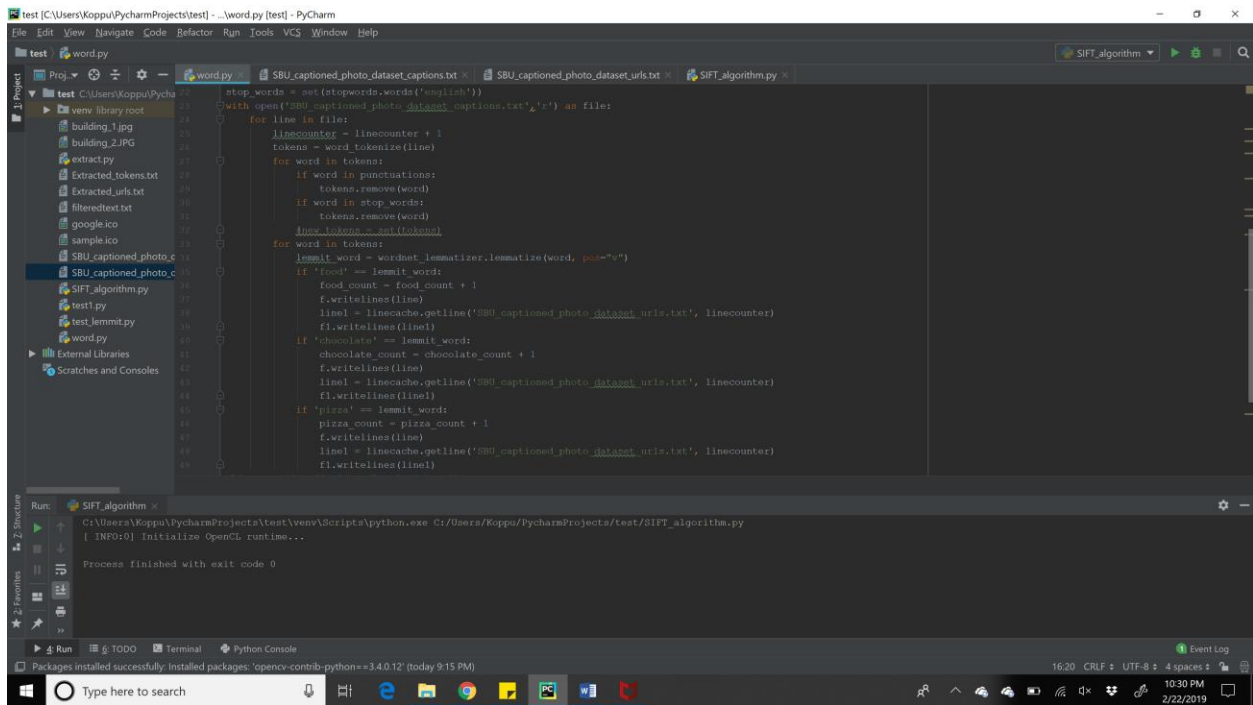
16 chocolate_count = 0
17 food_count = 0
18
19 f = open('Extracted_tokens.txt','a')
20 fl = open('Extracted_urls.txt','a')
21 punctuations = "?!.,;"
22 stop_words = set(stopwords.words('english'))
23 with open('SBU_captioned_photo_dataset_captions.txt','r') as file:
24     for line in file:
25         linecounter = linecounter + 1
26         tokens = word_tokenize(line)
27         for word in tokens:
28             if word in punctuations:
29                 tokens.remove(word)
30             if word in stop_words:
31                 tokens.remove(word)
32             new_tokens.append(word)
33         for word in tokens:
34             lemmat_word = wordnet.Lemmatizer().lemmatize(word, pos='n')
35             if 'food' == lemmat_word:
36                 food_count = food_count + 1
37             fl.writelines(line)
38             line1 = linecache.getline('SBU_captioned_photo_dataset_urls.txt', linecounter)
39             fl.writelines(line1)
40             if 'chocolate' == lemmat_word:
41                 chocolate_count = chocolate_count + 1
42             f.writelines(line)
43             line1 = linecache.getline('SBU_captioned_photo_dataset_urls.txt', linecounter)

```

Run: SIFT\_algorithm ×  
 C:\Users\Koppu\PycharmProjects\test\venv\Scripts\python.exe C:/Users/Koppu/PycharmProjects/test/SIFT\_algorithm.py  
 [ INFO:0] Initialize OpenCL runtime...  
 Process finished with exit code 0

## Lemmatization:

- Lemmatization is the process of getting a root word to the given keyword which is very helpful in order to pre-process the data.
- For example, consider a word running which is present in the caption. Its root word is run and when one predicts the output on 'run' may lose certain data.
- We eliminate stopwords when lemmatization is done as those words are not important.



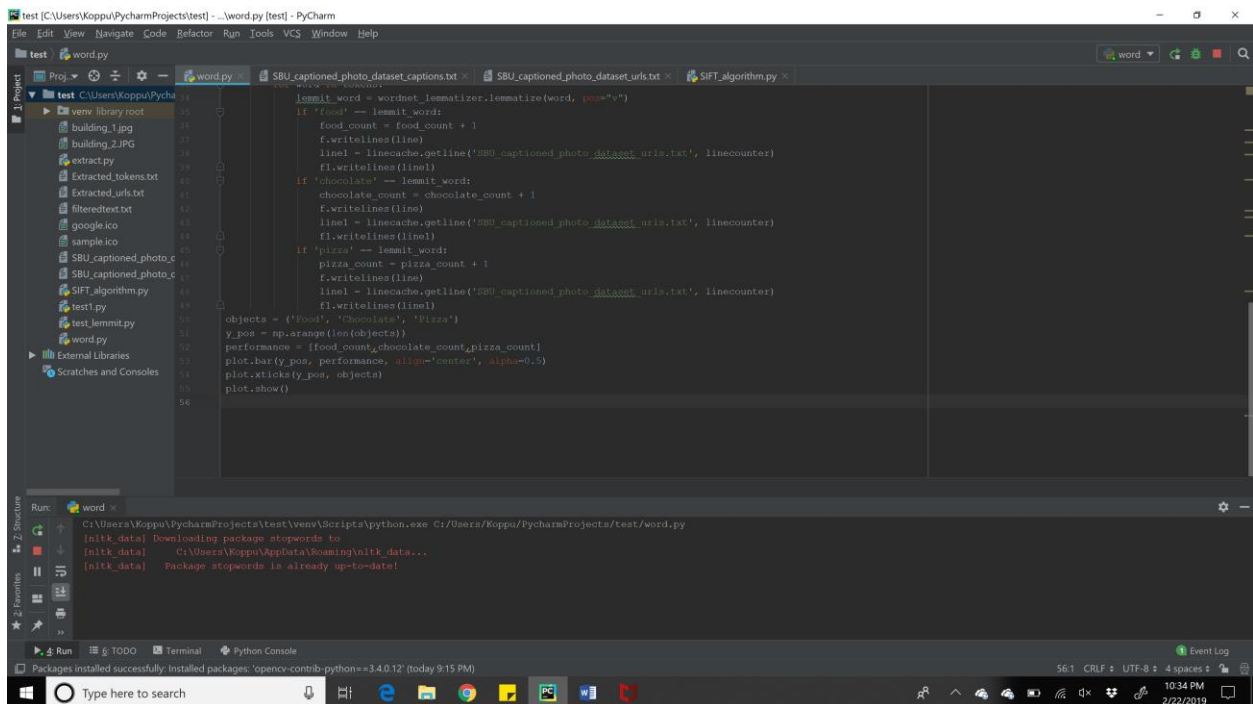
The image shows a PyCharm IDE window with the file `SIFT_algorithm.py` open. The code processes a dataset of captions and URLs, tokenizing and lemmatizing the text to count specific words like 'food', 'chocolate', and 'pizza'. The Run window shows the execution output, indicating that the OpenCL runtime was initialized successfully and the process finished with exit code 0.

```
stop_words = set(stopwords.words('english'))
with open('SBU_captioned_photo_dataset_captions.txt','r') as file:
    for line in file:
        linecounter = linecounter + 1
        tokens = word_tokenize(line)
        for word in tokens:
            if word in punctuations:
                tokens.remove(word)
            if word in stop_words:
                tokens.remove(word)
            lemmat_word = wordnet_lemmatizer.lemmatize(word, pos='n')
            if 'food' == lemmat_word:
                food_count = food_count + 1
            f.writelines(line)
            line1 = linecache.getline('SBU_captioned_photo_dataset_urls.txt', linecounter)
            f1.writelines(line1)
            if 'chocolate' == lemmat_word:
                chocolate_count = chocolate_count + 1
            f1.writelines(line)
            line1 = linecache.getline('SBU_captioned_photo_dataset_urls.txt', linecounter)
            f1.writelines(line1)
            if 'pizza' == lemmat_word:
                pizza_count = pizza_count + 1
            f1.writelines(line)
            line1 = linecache.getline('SBU_captioned_photo_dataset_urls.txt', linecounter)
            f1.writelines(line1)
```

Run: SIFT\_algorithm x  
C:\Users\Koppu\PycharmProjects\test\venv\Scripts\python.exe C:/Users/Koppu/PycharmProjects/test/SIFT\_algorithm.py  
[ INFO:0] Initialize OpenCL runtime...  
Process finished with exit code 0

## Image statistics:

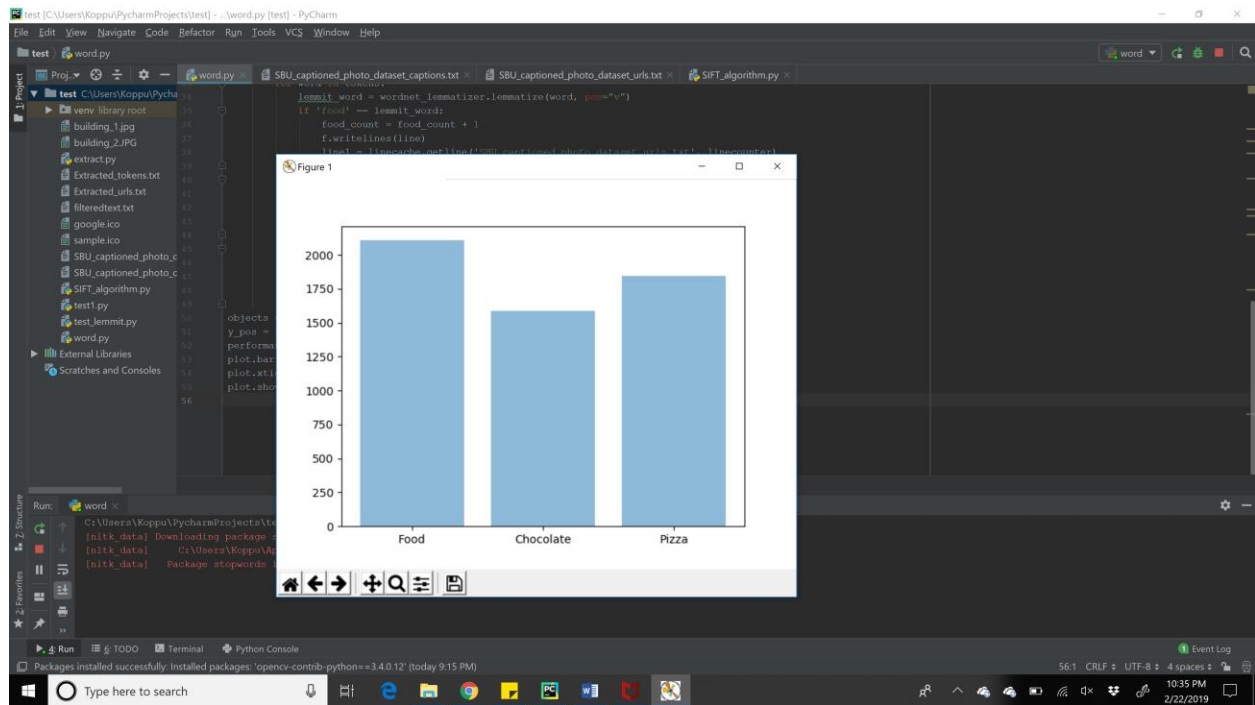
A word count has been performed and respected data has been plot using a matplotlib library.



The image shows a PyCharm IDE window with the file `word.py` open. The code uses the `word` module to count the words 'food', 'chocolate', and 'pizza' in the dataset. The Run window shows the execution output, indicating that the `word` package was successfully installed and the process finished with exit code 0.

```
objects = ('Food', 'Chocolate', 'Pizza')
y_pos = np.arange(len(objects))
performance = [food_count, chocolate_count, pizza_count]
plot.bar(y_pos, performance, align='center', alpha=0.5)
plot.xticks(y_pos, objects)
plot.show()
```

Run: word x  
C:\Users\Koppu\PycharmProjects\test\venv\Scripts\python.exe C:/Users/Koppu/PycharmProjects/test/word.py  
[nltk\_data] Downloading package stopwords to  
[nltk\_data] C:\Users\Koppu\AppData\Roaming\nltk\_data...  
[nltk\_data] Package stopwords is already up-to-date!



**SIFT Algorithm:**

