**Assignment No: 4**

**Prajakta Mane(BE/A/B1821016)**

**Aim**: Exploratory analysis on Twitter Text Data. Perform text preprocessing. Apply Zipf’s and Heaps law , Identify topics.

**Objectives**:

1. The student will be able to perform text preprocessing

2. The student will learn the concept of Zipf’s law.

**Software Requirements**: Windows with python and Jupyter notebook.

**Hardware Requirements**:

Pentium IV system with latest configuration

**Outcomes:** The students will be able to perform text preprocessing.And will learn the Zipf’s law concept.

**Theory:**

**Data Preprocessing :**

Data preprocessing is a technique which is used to transform the raw data in a useful and efficient format.

**Steps Involved in Data Preprocessing:**

**1. Data Cleaning:**

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

* **(a). Missing Data:**   
  This situation arises when some data is missing in the data. It can be handled in various ways.   
  Some of them are:
  1. **Ignore the tuples:**   
     This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
  2. **Fill the Missing values:**   
     There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.
* **(b). Noisy Data:**   
  Noisy data is meaningless data that can’t be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
  1. **Binning Method:**   
     This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segment is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  2. **Regression:**   
     Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
  3. **Clustering:**   
     This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

**2. Data Transformation:**

This step is taken in order to transform the data in appropriate forms suitable for the mining process. This involves following ways:

1. **Normalization:**   
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
2. **Attribute Selection:**   
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
3. **Discretization:**   
   This is done to replace the raw values of numeric attributes by interval levels or conceptual levels.
4. **Concept Hierarchy Generation:**   
   Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

**3. Data Reduction:**

Since data mining is a technique that is used to handle huge amounts of data. While working with a huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction techniques. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

1. **Data Cube Aggregation:**   
   Aggregation operation is applied to data for the construction of the data cube.
2. **Attribute Subset Selection:**   
   The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use the level of significance and p- value of the attribute.The attribute having p-value greater than significance level can be discarded.
3. **Numerosity Reduction:**   
   This enables us to store the model of data instead of whole data, for example: Regression Models.
4. **Dimensionality Reduction:**   
   This reduces the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction is called lossless reduction, else it is called lossy reduction. The two effective methods of dimensionality reduction are:Wavelet transforms and PCA (Principal Component Analysis).

**Zipf’s Law:-**

Zipf's law is a law about the frequency distribution of words in a language (or in a collection that is large enough so that it is representative of the language). To illustrate Zipf's law let us suppose we have a collection and let there be V unique words in the collection (the vocabulary). For each word in the collection we need to compute the freq(word) = how many times a word occurs in the collection. Then we rank the words descending by their frequency (most frequent words have rank 1, next frequent word has rank 2, ...).**Zipf’s law**, in [probability](https://www.britannica.com/science/probability), assertion that the frequencies *f* of certain events are inversely proportional to their rank *r*. The law was originally proposed by American linguist [George Kingsley Zipf](https://www.britannica.com/biography/George-Zipf) (1902–50) for the frequency of usage of different words in the English language; this frequency is given approximately by *f*(*r*) ≅ 0.1/*r*. Thus, the most common word (rank 1) in English, which is *the*, occurs about one-tenth of the time in a typical text; the next most common word (rank 2), which is *of*, occurs about one-twentieth of the time; and so forth. Another way of looking at this is that a rank *r* word occurs 1/*r* times as often as the most frequent word, so the rank 2 word occurs half as often as the rank 1 word, the rank 3 word one-third as often, the rank 4 word one-fourth as often, and so forth. Beyond about rank 1,000, the law completely breaks down.

Zipf’s law purportedly has been observed for many other statistics that follow an exponential distribution. For example, in 1949 Zipf claimed that the largest city in a country is about twice the size of the next largest, three times the size of the third largest, and so forth. While the fit is not perfect for languages, populations, or any other data, the basic idea of Zipf’s law is useful in schemes for [data compression](https://www.britannica.com/technology/data-compression) and in [allocation of resources](https://www.britannica.com/topic/allocation-of-resources) by urban planners.

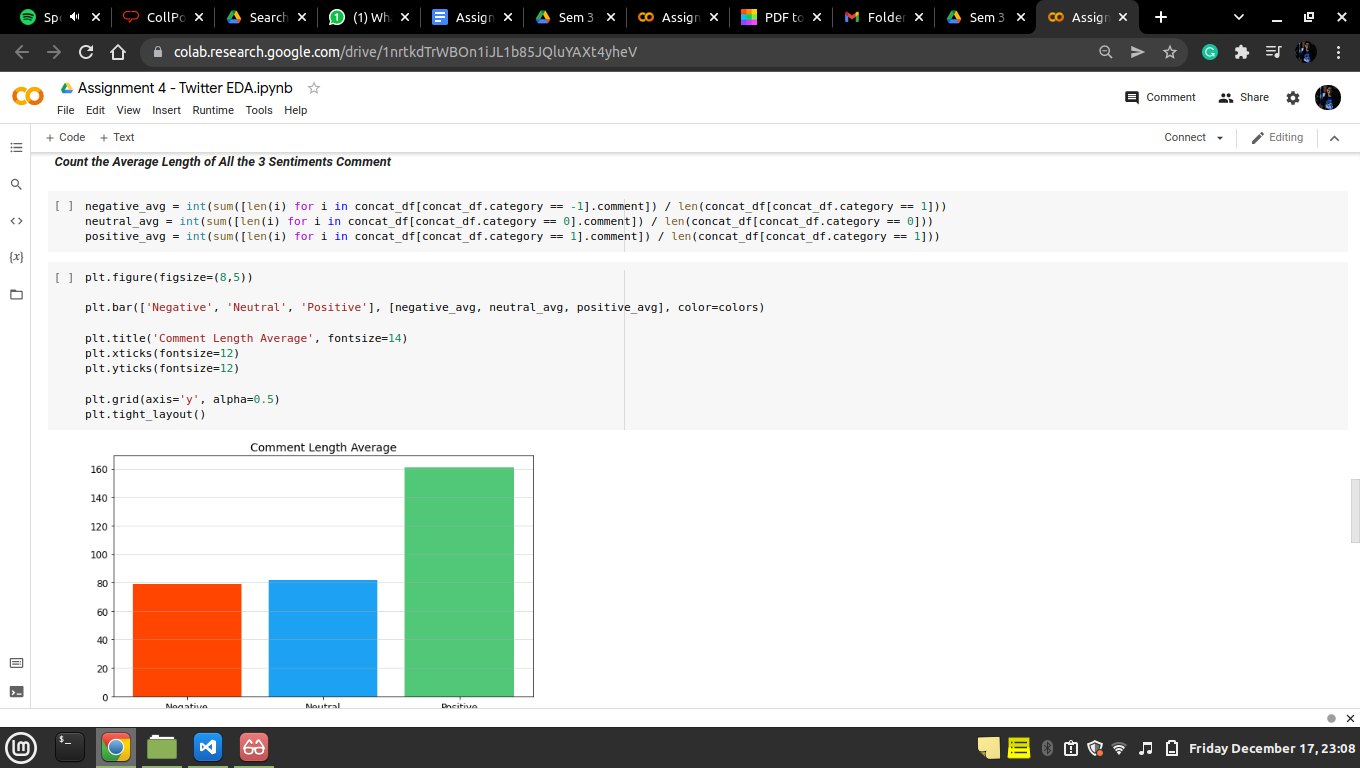
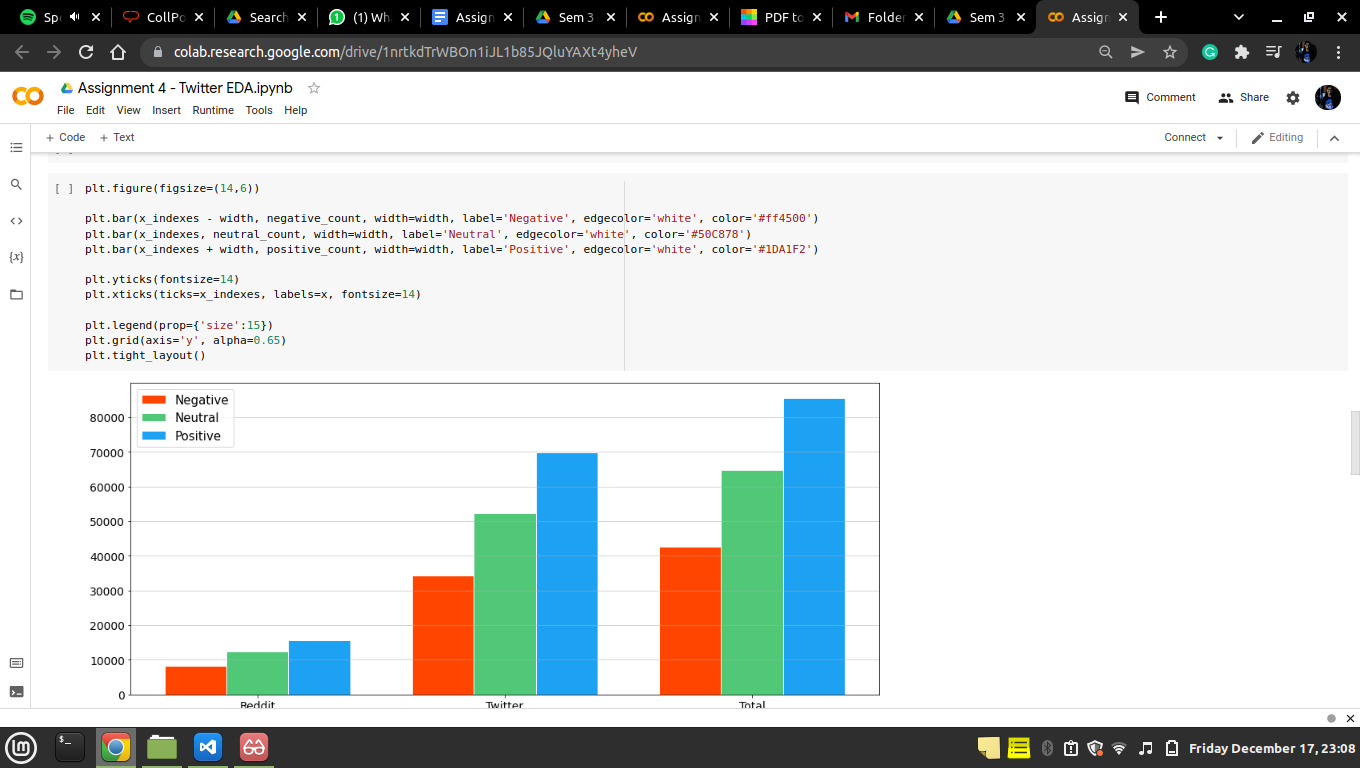
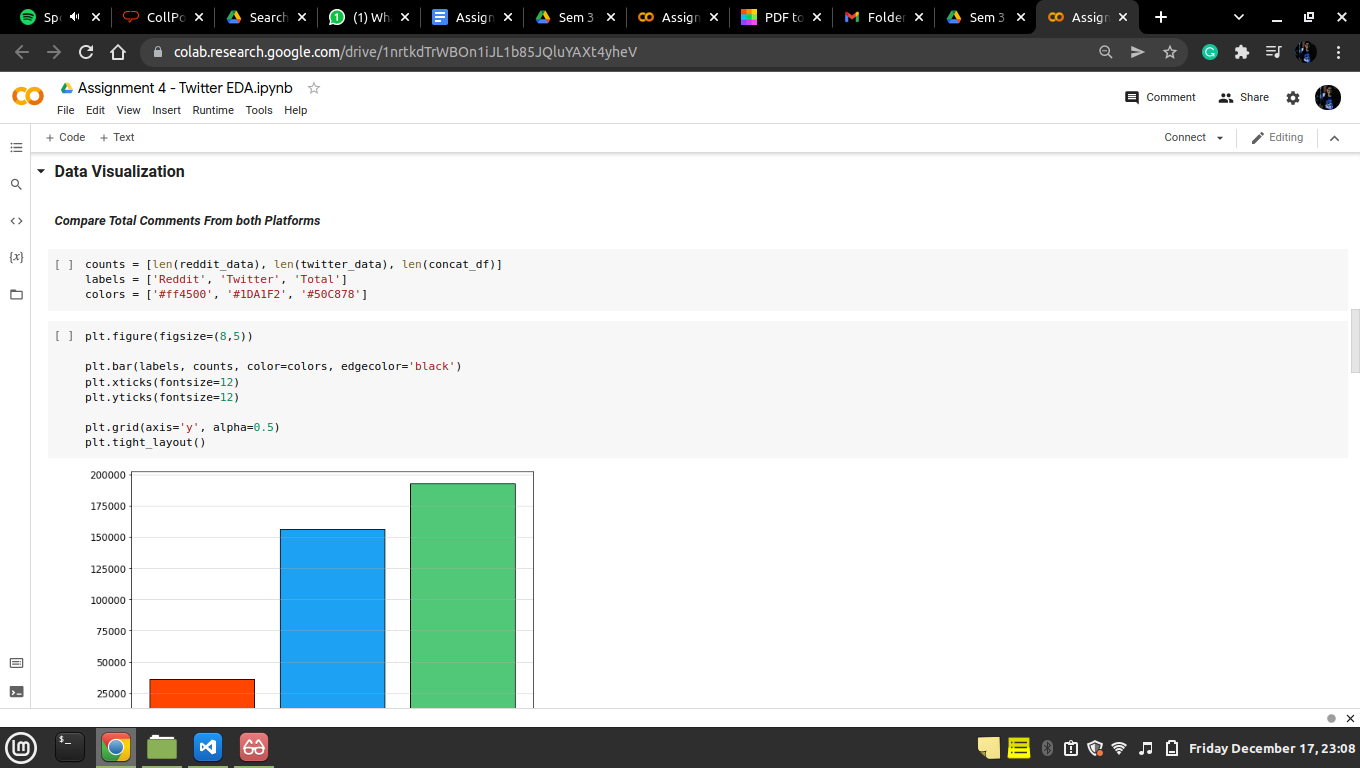
**Heaps' law:-**

**Heaps' law** (also called **Herdan's law**) is an [empirical law](https://en.wikipedia.org/wiki/Empirical_law) which describes the number of distinct words in a document (or set of documents) as a function of the document length

**Code Link :**

[**https://drive.google.com/file/d/1nrtkdTrWBOn1iJL1b85JQluYAXt4yheV/view?usp=sharing**](https://drive.google.com/file/d/1nrtkdTrWBOn1iJL1b85JQluYAXt4yheV/view?usp=sharing)

**Implementation( Output Screenshot)**



**Conclusion:** Learnt text preprocessing technique and Zipf’s law.