**DATA COLLECTION AND PREPROCESSING:**

**The following document contains the information regarding the data collection like source of data and number of observations in the data etc. also various steps involved in data preprocessing.**

**Code :** <https://github.com/avataarapp/Data-Preprocessing/blob/master/DataPreprocessor.py>

**Section 1 : IMDB data set**

|  |  |
| --- | --- |
| **Positive Reviews** | **Negative Reviews** |
| Observations : 25000 | Observations:25000 |
| E:\Avatar internship (heuro)\4 th june 2020\positive2.png | E:\Avatar internship (heuro)\4 th june 2020\positive2.png |
| Fig : Count plot of Positive Sentiment | Fig : Count plot og Negative sentiment |
| E:\Avatar internship (heuro)\4 th june 2020\positive1.png | E:\Avatar internship (heuro)\4 th june 2020\negative.png |
| Fig : Hist plot of Positive Sentiment | Fig: Hist plot of Negative Sentiment |

**Source** :<https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews>

**Section 2: Mturk Experiment ( Formality and Informality data set )**

|  |  |
| --- | --- |
| **Formality** | **Informality** |
| Observations : 3470 | Observations : 11573 |
| E:\Avatar internship (heuro)\4 th june 2020\count_formal.png | E:\Avatar internship (heuro)\4 th june 2020\count_informal.png |
| Fig : Count plot of Formality | Fig : Count plot of Informality |
| E:\Avatar internship (heuro)\4 th june 2020\hist_formal.png | E:\Avatar internship (heuro)\4 th june 2020\hist_informal.png |
| Fig : Hist plot of Formality | Fig : Hist plot of Informality |

**Source :** <https://drive.google.com/file/d/0B2Mzhc7popBgdXZmRlg2RUdqdDA/view>

**Section 3: Authors Data set**

|  |  |  |
| --- | --- | --- |
| **Authors Data set** | | |
| **Total 5635 Observations** | | |
| **Author 1:EAP:Edgar Allan Poe** | **Author 2 : HPL:HP Lovercraft** | **Author 3 : MWS : Mary Wollstonecraft** |
| **E:\Avatar internship (heuro)\4 th june 2020\count_EAP.png** | **E:\Avatar internship (heuro)\4 th june 2020\count_HPL.png** | **E:\Avatar internship (heuro)\4 th june 2020\count_MWS.png** |
| Fig : Count plot of EAP author | Fig :Count plot of HPL author | Fig : Count plot of MWS author |
| **E:\Avatar internship (heuro)\4 th june 2020\hist_EAP.png** | **E:\Avatar internship (heuro)\4 th june 2020\hist_HPL.png** | **E:\Avatar internship (heuro)\4 th june 2020\hist_MWS.png** |
| Fig : Hist plot of EAP author | Fig : Hist plot of HPL author | Fig : Hist plot of MWS author |

**Source :**<https://www.kaggle.com/c/spooky-author-identification/data>

**Section 4: Simple English [Taken Answers variable from this data set]**

|  |  |
| --- | --- |
| **Simple English** | |
| Total Observations :29208 | |
| E:\Avatar internship (heuro)\4 th june 2020\count_eng.png | E:\Avatar internship (heuro)\4 th june 2020\hist_eng.png |
| **Fig :** Count plot of simple English | Fig : Hist plot of Simple English |

**Source** :<https://www.microsoft.com/en-us/download/details.aspx?id=52419>

**Pre Processing steps Followed :**

|  |
| --- |
| 1.Remove Brackets ( ) |
| 2.Remove double slashes ( \\,//) |
| 3.Remove the numbers ,where the numbers begin with sentence . |
| 4.Remove special characters |
| 5.Remove special symbols |
| 6 Remved HTML tags |
| 7 Mapped Contractions |