Assessment-03
User Manual

Student Name: Pragy Parashar
Student ID: 31940757

Objective: To visualize posting trends from HardwareRecs (2015-2019).
Libraries Imported: re, pandas, matplotlib
Input Data Files: data.xml

The dataset is known as HardwareRecs [https://hardwarerecs.stackexchange.com] which is a Q&A site for people seeking specific hardware recommendations. The Q&A site is a platform for users to exchange knowledge by asking and answering questions such as Quora, Zhihu, and Stack Overflow. Within HardwareRecs, users can ask questions about hardware recommendations, while other users can also answer those questions with corresponding suggestions.

The data is written in XML (Extensible Markup Language) format.

```
1 <?xml version="1.0" encoding="utf-8"?>
2 <posts>
3   <row Id="2481" PostTypeId="1" CreationDate="2016-04-07T18:11:33.793" Body="&lt;p&gt;In $200 price range,
    should I be looking at cards from AMD or Nvidia?&lt;/p&gt;&#xA;" />
4   <row Id="7588" PostTypeId="2" CreationDate="2017-06-18T13:27:28.750" Body="&lt;p&gt;Sparkfun will be more
    louder as it can provide you with about 85dbA&lt;/p&gt;&#xA;" />
5   <row Id="8412" PostTypeId="1" CreationDate="2017-11-19T17:39:35.250" Body="&lt;p&gt;If you can't name just
    one, name a few. In this case, price does not matter. Thanks.&lt;/p&gt;&#xA;" />
6   <row Id="9752" PostTypeId="2" CreationDate="2018-09-16T17:37:36.710" Body="&lt;p&gt;I have made my decision,
    after some time. I got the laptop with the GL702VS&lt;/p&gt;&#xA;" />
7 </posts>
```

preprocessData_31940757.py

This file is used for data pre-processing and cleaning. It contains two method:

- **preprocessLine**: This method takes in the input string and returns a dictionary that will contain the Post Id of the text as the key and the clean body as the value.
- **splitFile:** This method takes in the data file path, question file path and answer file path as the parameters. The function segregates the data into question and answers based on the post type id into two text file question and answers.

parser_31940757.py

This is a class file for Parser which will contain the class attributes and the methods
Parser class has the following methods:

- *_init__(self, inputString)*
  creates an instance object
- *__str__(self)*
  print the Parser object in a formatted string
- *getID(self)*
  searches for the post id and returns it
- *getPostType(self)*
  searches for the post type and returns it
- *getDateQuarter(self)*
  searches for the year and quarter when the post got posted online
- *getCleanedBody(self)*
  returns preprocessed data; invokes the preprocessLine() function
- *getVocabularySize(self)*
  counts and returns vocabulary size of the post; before counting, it tackles some further processing of data, like removing punctuations.

dataVisualization_31940757.py

This file is used to perform the data analysis over the processed data to generate insights about the data trend.
The two output files generated will be the following:
vocabularySizeDistribution.png
postNumberTrend.png

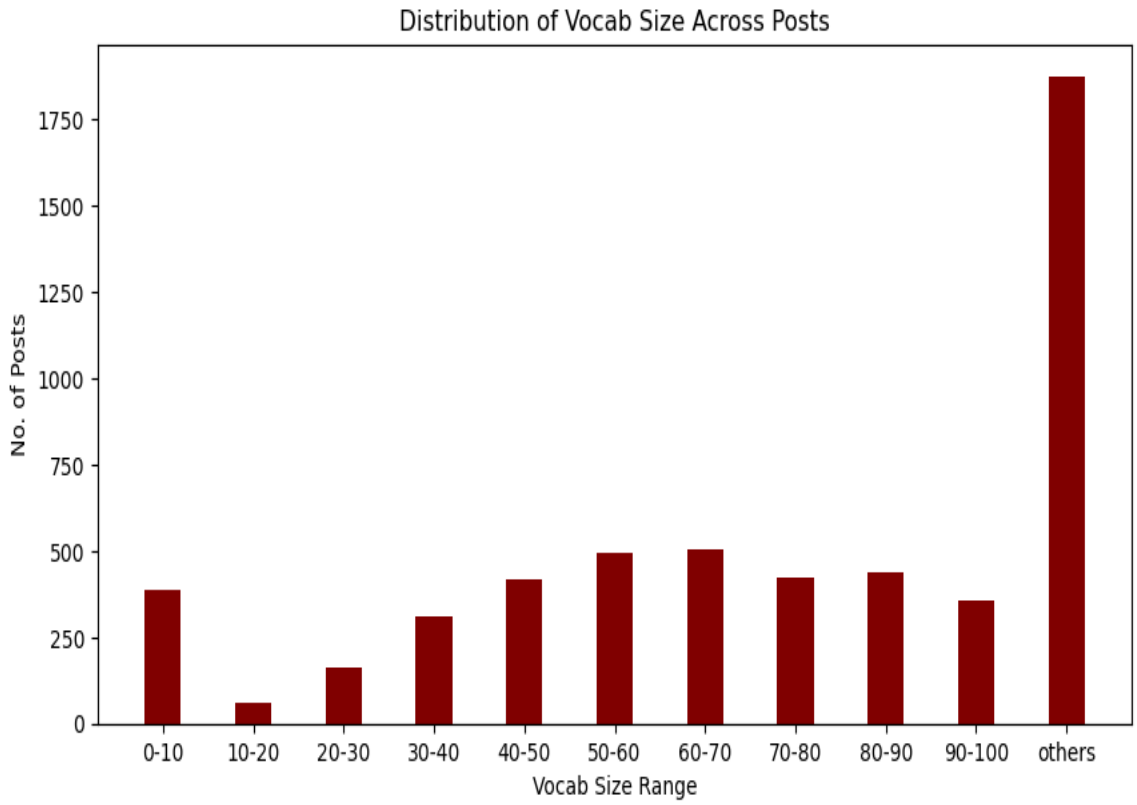
This file contains 2 functions:

- visualizeWordDistribution(inputFile, outputImage)

  This method is used to create a bar plot of the word distribution trend. The data is grouped into categories according to the word count. For the word count greater than 100 the group is labelled as others.
  .

- visualizePostNumberTrend(inputFile, outputImage)
  This method is used to analyze the trend for question and answers posted to over the years and quarters

vocabularySizeDistribution.png



The image above depicts the trend of the unique word range in the grouped data set. Here it can be observed that most of the unique words belongs to the other category.

postNumberTrend.png

The image above represents the line graph generated from the second method. This graph analyzes the trend of the questions and answers posted over the website over the year. The data is analyzed for each quarter over the years. It is interesting to note that the answers posted over the websites have significantly reduced during time. This trend also suggest the decreasing interest of people in the forum over the years.