USER MANUAL

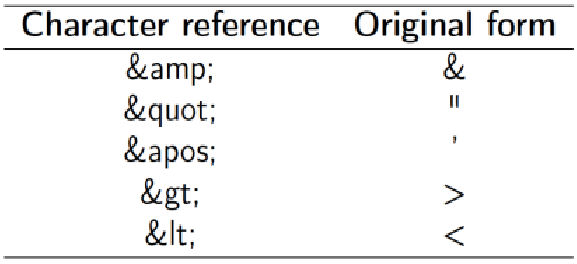
By Rohan Singh
31339646

Start of Application	The file must be a xml format and have the following format: ● ID: unique identifier to represent each post ● PostTypeID: where ○ 1 = Question ○ 2 = Answer ○ 3 to 8 = Others ● CreationDate: The creation date and time of post where the format is yyyy-mm-ddThh:mm:ss ● Body: The content of the post
Pre processing file	● The Xml file should be named data.xml and should be present in the same folder path as the other three project files. ● After this the preprocessingdata file should be run. Nothing need to be input. ● Upon completion of the pre processing tasks the contents of the post will be saved as two individual files and an output confirmation will be printed. ● The code begins by reading in all of the posts of the given dataset. ● Then we will conduct a number of pre processing tasks to clean the post content that is the body of the post which we will use for further analysis. ● For each post, we first extract the body of the string embedded within “Body:"..."” in each row of the XML file. After this we carry out the following steps:

Pre processing file

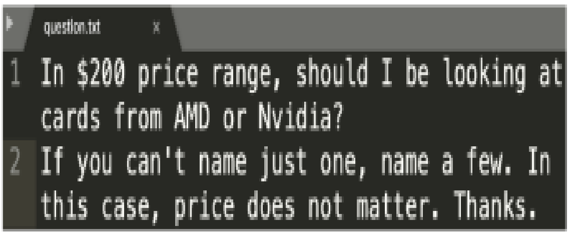
- We convert special character references back to its original form by following the rules in Table 1.

Table 1: Character Reference Transformation.



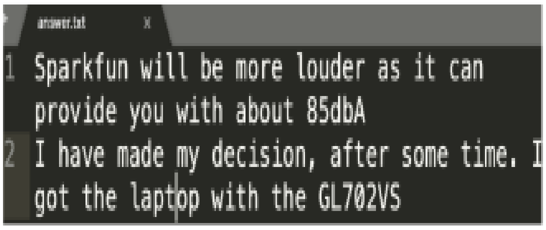
Character reference	Original form
&amp;	&
&quot;	"
&apos;	'
&gt;	>
&lt;	<

- We replace special characters including “ ”, “ ” by a single empty space
- We remove all HTML tags.
- Example:
  - Before filtering: <p>In $200 price range, should I be looking at cards from AMD or Nvidia?</p>
  - After filtering: In $200 price range, should I be looking at cards from AMD or Nvidia?
- Then we identify if the post is a question or answer. We then save the data into two different files “question.txt” and “answer.txt” according to the post type shown in the data. The cleaned body/content for each post is then saved in one line in the output file. Examples can be seen in the figure below.



```
In $200 price range, should I be looking at cards from AMD or Nvidia?
If you can't name just one, name a few. In this case, price does not matter. Thanks.
```

(a) question.txt



```
Sparkfun will be more louder as it can provide you with about 85dbA
I have made my decision, after some time. I got the laptop with the GL702VS
```

(b) answer.txt

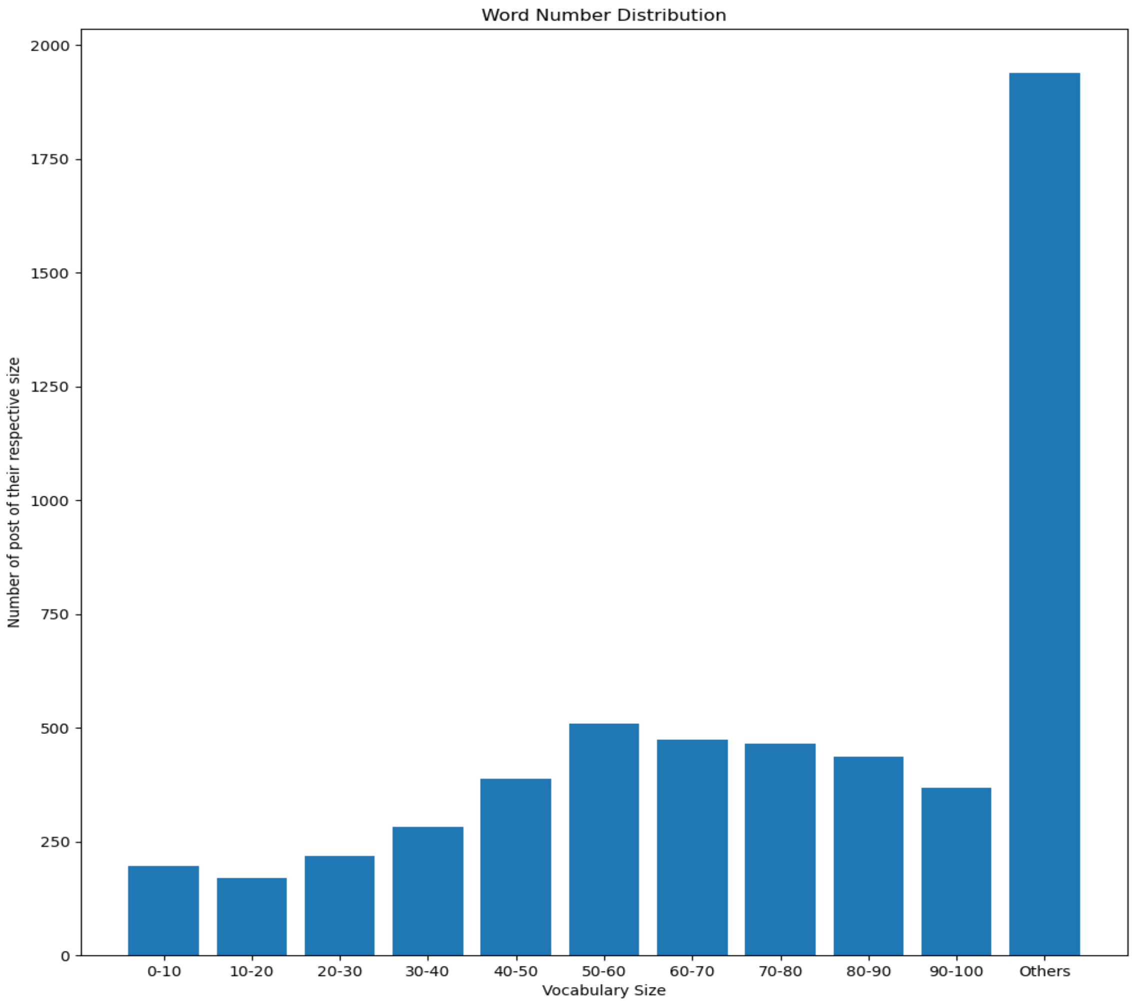
Data Analysis or Parser class	● The main function of this file is to further parse the given row of the data in XML format. ● This function is just used to create a class so no input or output is generated here. It is a vital function used in the datavisualisation file and does not need to be run as it is automatically run in the datavisualisation file. ● This function contains the following classes: ○ ● __init__(self, inputString): This is the constructor required for creating instances of this class. The inputString will be the row of data from the XML file. ○ __str__(self): Returns a formatted string. The order of output is “ID, post type, creation date quarter, the cleaned content”. ○ getID(self): Get Id of the post (indicated by “Id” attribute) ○ getPostType(self) Get the post type of the post (indicated by “PostTypeId” attribute) with 1 as the question, 2 as the answer, and 3-8 as others. ● getDateQuarter(self) Get the date quarter of the creation date (indicated by the “CreationDate” attribute). One year has four quarters including Q1 (Jan to Mar), Q2 (Apr to Jun), Q3 (Jul to Sep) and Q4 (Oct to Dec). For example, given “2016-04-07T18:11:33.793” as the CreationDate, the program returns a string named “2016Q2”. ○ getCleanedBody(self) Get the cleaned body of the posts (indicated by “Body” attribute) which is the

Data analysis or Parser class	extracted cleaned body as that of the first file. We import the function preprocessLine() to reuse the pre-processing functionality. But we do not split the question/answers or save to the file. ○ getVocabularySize(self) Get the number of unique words in the cleaned body converted in the lower case. Note that we do not count space or punctuation as the word. For example, given the sentence "Although I use Mac, I do not like Mac.", there are 7 unique words including {"although", "i", "use", "mac", "do", "not", "like"}.
Data Visualization Methods file	● In this file we implement two functions to visualise the statistics in the form of a bar graph and line plot. ● We just need to run the file to generate the two data visualisation images. ● This is carried out using two functions: ○ VisualizeVocabularySizeDistribution(inputFile, outputImage): We count the vocabulary size for each post and draw a bar chart to visualise the distribution of the vocabulary size of all posts. The graph of your visualisation figure is converted into a png file and saved as "vocabularySizeDistribution.png" in the same root folder as the above files. Once the files have been created and saved you will see a confirmation message on your screen. ○ visualizePostNumberTrend(inputFile, outputImag): This

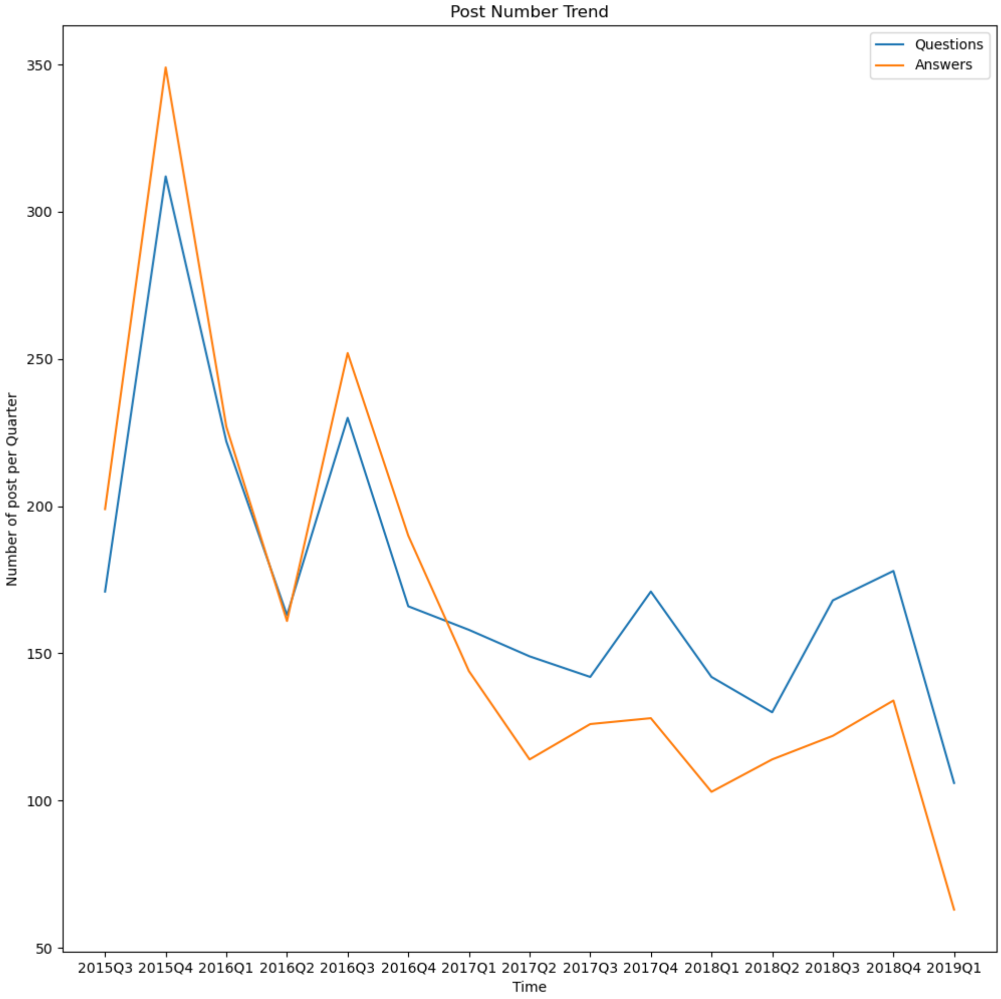
Data Visualization Methods file	function displays the trend of the post number in the Q&A site. Here we first get the number of questions and answers in each quarter. Then following the time order, a line chart is drawn to annotate the number of posts in each quarter. The graph of your visualisation figure is converted into a png file and saved as “postNumberTrend.png” in the same root folder as the above files. Once the files have been created and saved you will see a confirmation message on your screen.

Explanation to Describe the Graphs

The graph shown on the next page is a bar chart to visualise the distribution of the vocabulary size of all posts. The x-axis is the vocabulary size, and the y-axis represents the number of posts with a certain vocabulary size. In the x-axis, the vocabulary size interval is 10 and once the vocabulary size is larger than or equal to 100, it added into “others”, i.e., 0-10, 10-20, 20-30, 30-40, 40-50, 50-60, 60-70, 70-80, 80-90, 90-100, others (left inclusive).



The graph shown on the next page displays the trend of the post number in the Q&A site. For the input file “data.xml”, we first get the number of questions and answers in each quarter. Then following the time order, a line chart is plotted to annotate the number of posts in each quarter. A legend in the figure for more information about the lines. The x axis is time and the y axis represents the number of posts in the quarter.


Post Number Trend
Questions
Answers
Number of post per Quarter
350
300
250
200
150
100
50
2015Q3
2015Q4
2016Q1
2016Q2
2016Q3
2016Q4
2017Q1
2017Q2
2017Q3
2017Q4
2018Q1
2018Q2
2018Q3
2018Q4
2019Q1
Time