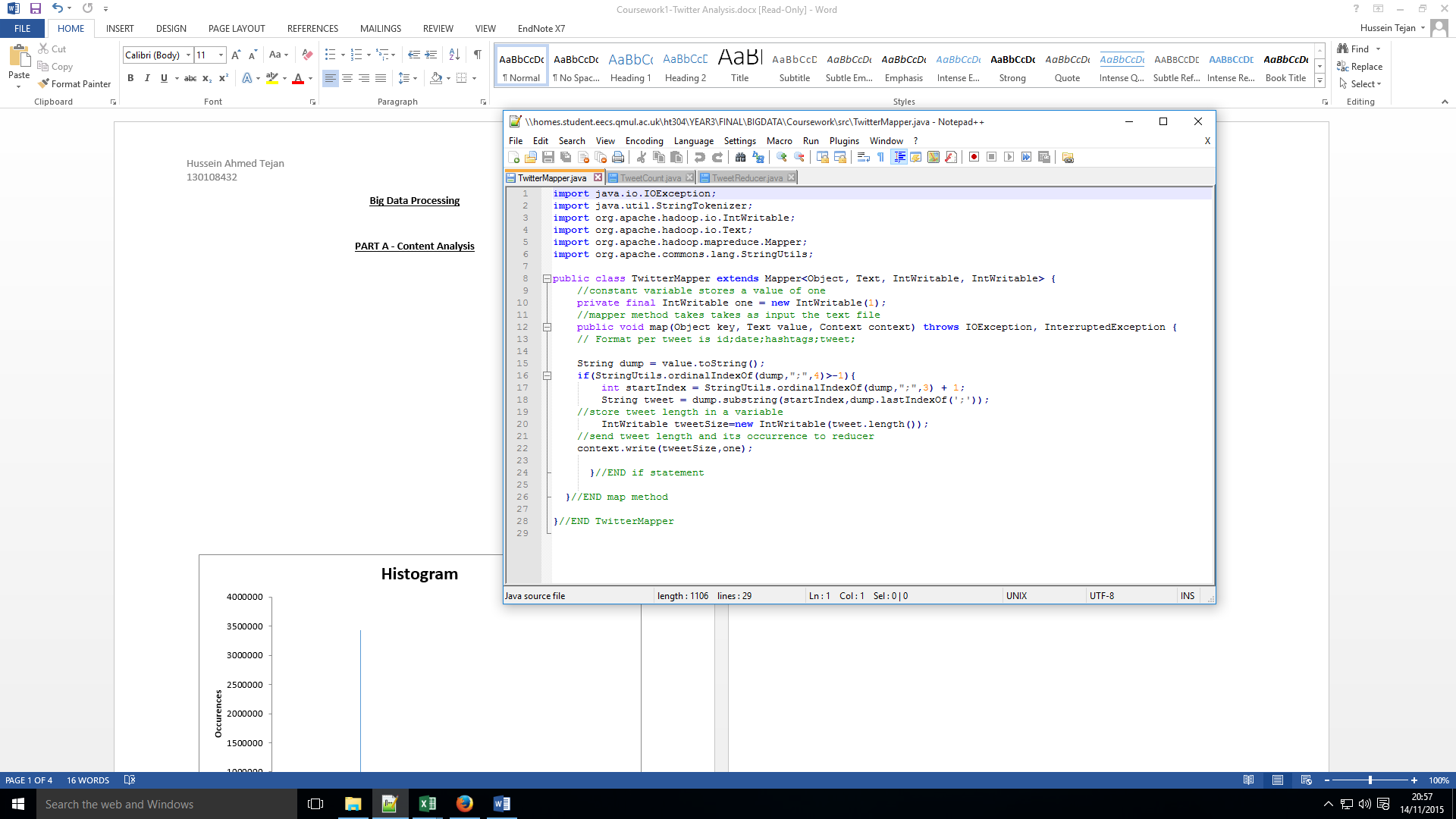
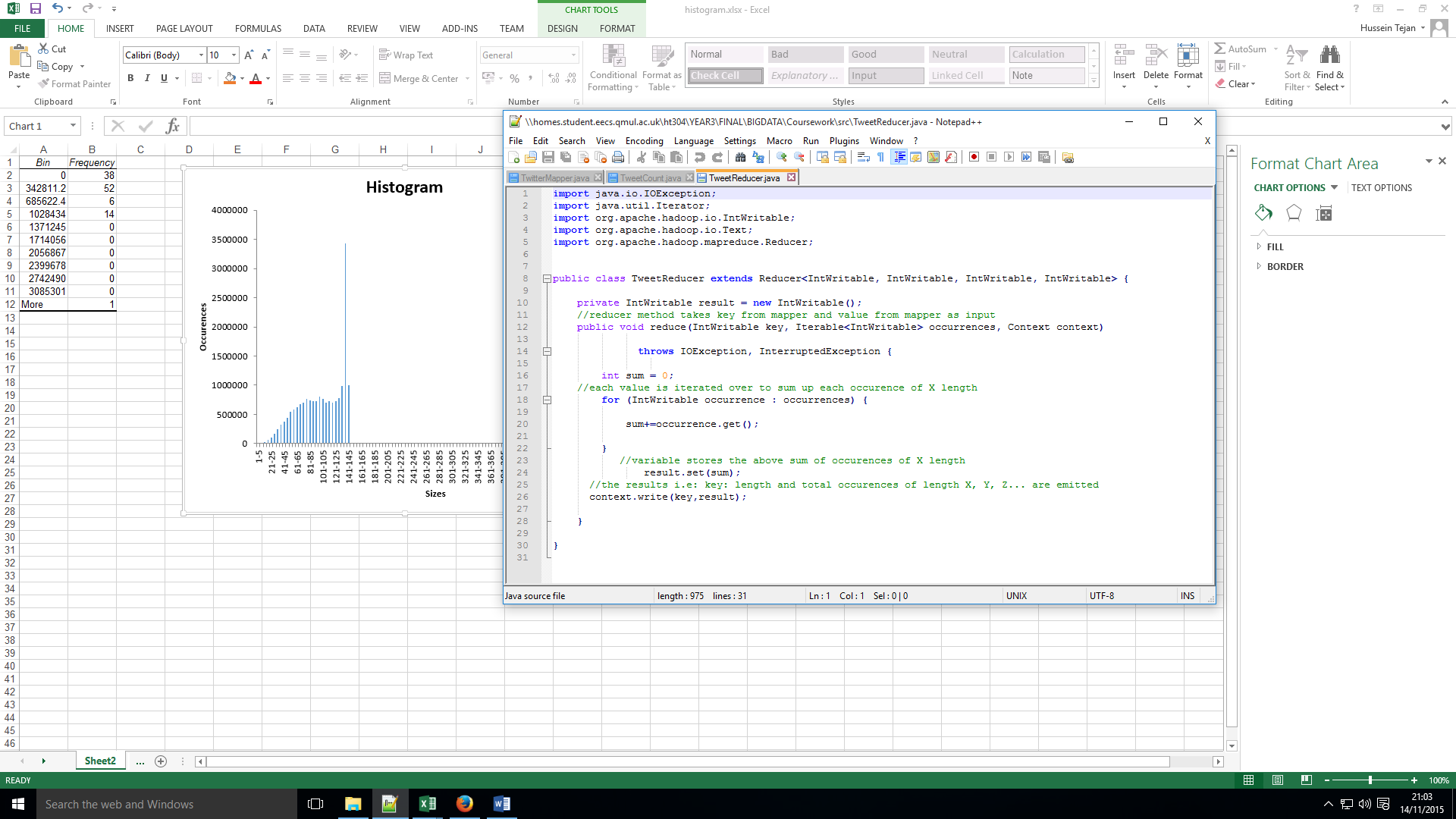
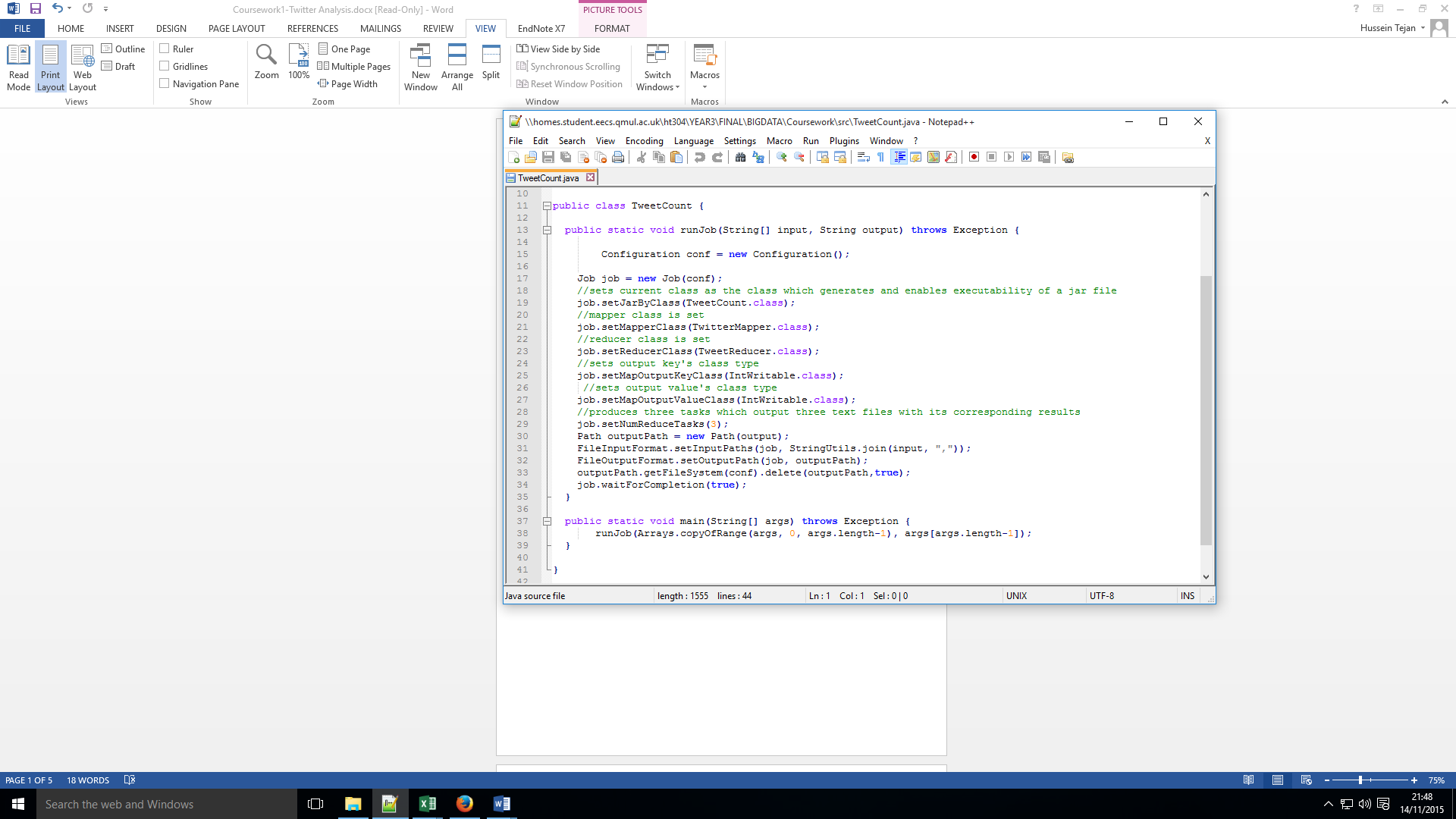
**Big Data Processing- Coursework1 – Twitter Analysis**

**PART A - CONTENT ANALYSIS**

**TWEET MAPPER CODE:**

**TWEET REDUCER CODE**:

**TWEET COUNT CODE:**



**EXPLANATION:**

The Map Reduce approach with the first part of (**A) Content Analysis** is where a counter is created to store 1 as a value which keeps track of a particular tweet length within the mapper and thus is sent (emitted) to the reducer to total up and count how many times a certain tweet length occurred within the data set (data/olympictweets). Henceforth, a map method’s would store the tweet length of every single tweet in a variable and emits it directly to the reducer along with its corresponding occurrence within the data set i.e.: length=300, occurrence=200. Which means for length 300 it would occur in the data set 200x.

In addition, a reducer method would take as input the key being a length and its occurrence being a value. It would contain an iteration which iterates through each occurrence and sums up the whole occurrence total for X length:

i.e.:

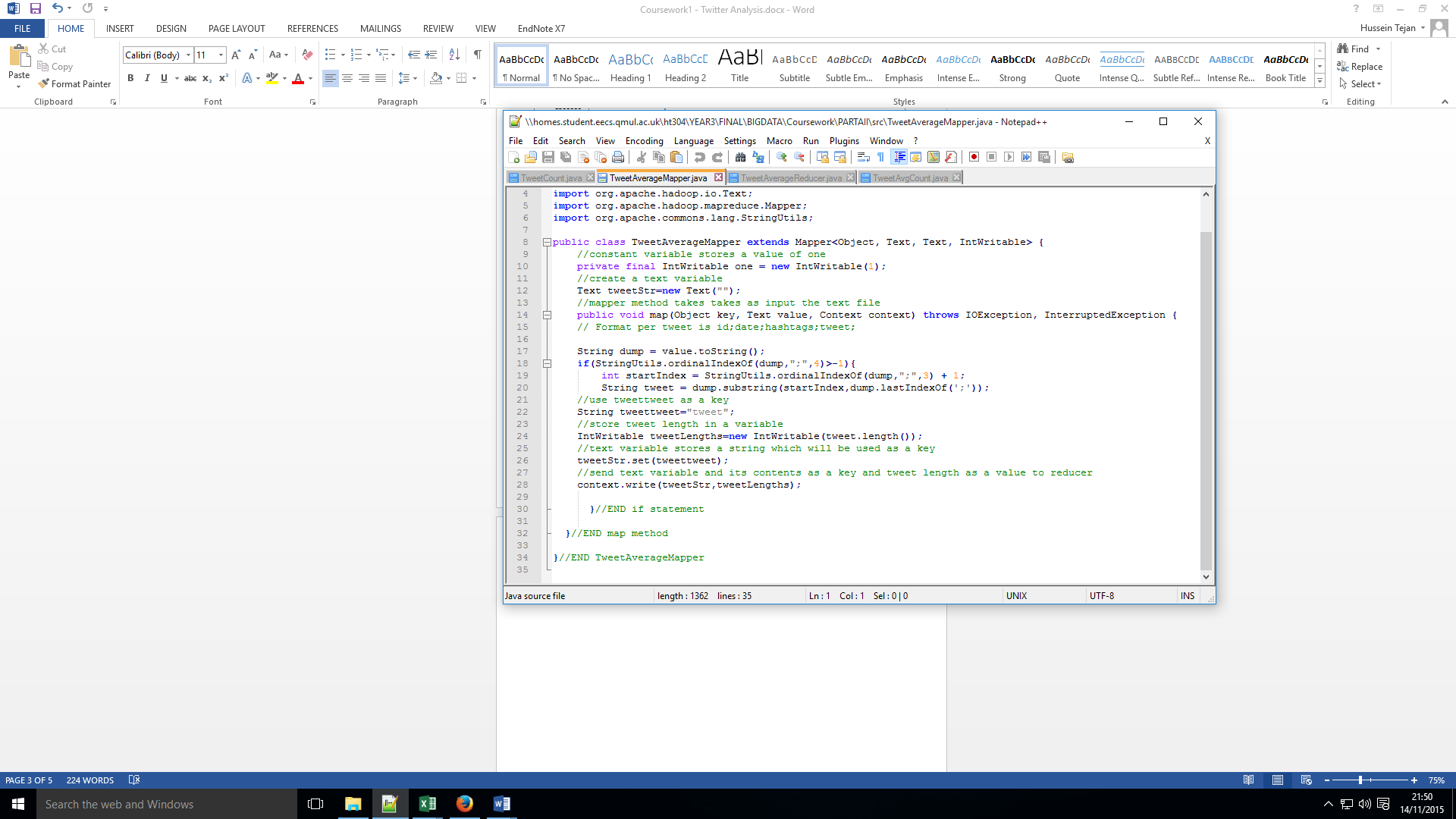
length=12, occurrence=1

length=12, occurrence=1

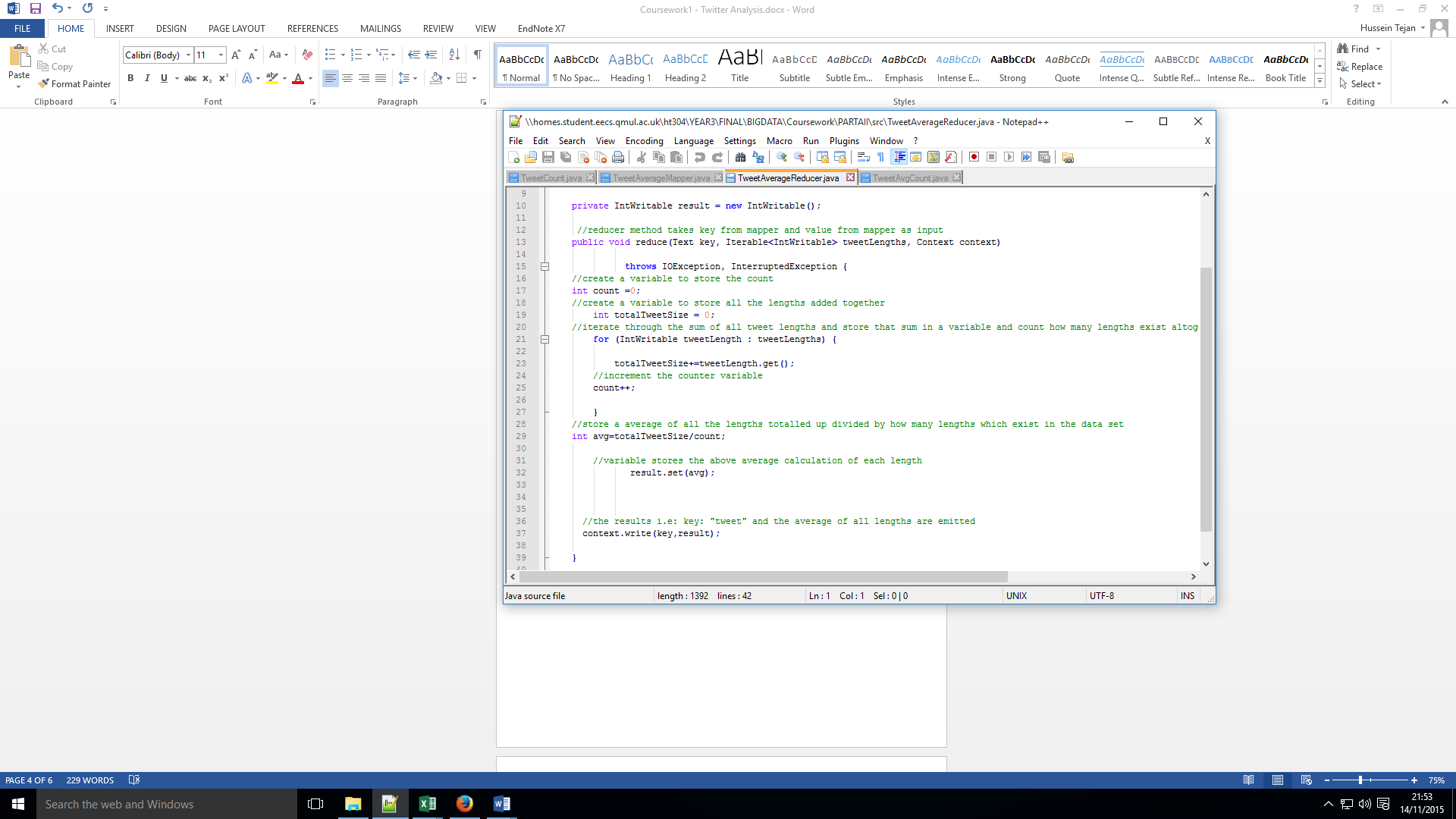
As a result, length=12, would result in two occurrences; this total of occurrences is set as an output value.

Furthermore, this enabled me to produce the histogram displayed above in Excel, which clearly denotes that the range of lengths 136-140 had the most lengths and that certain ranges had no lengths reported within in the dataset at all.

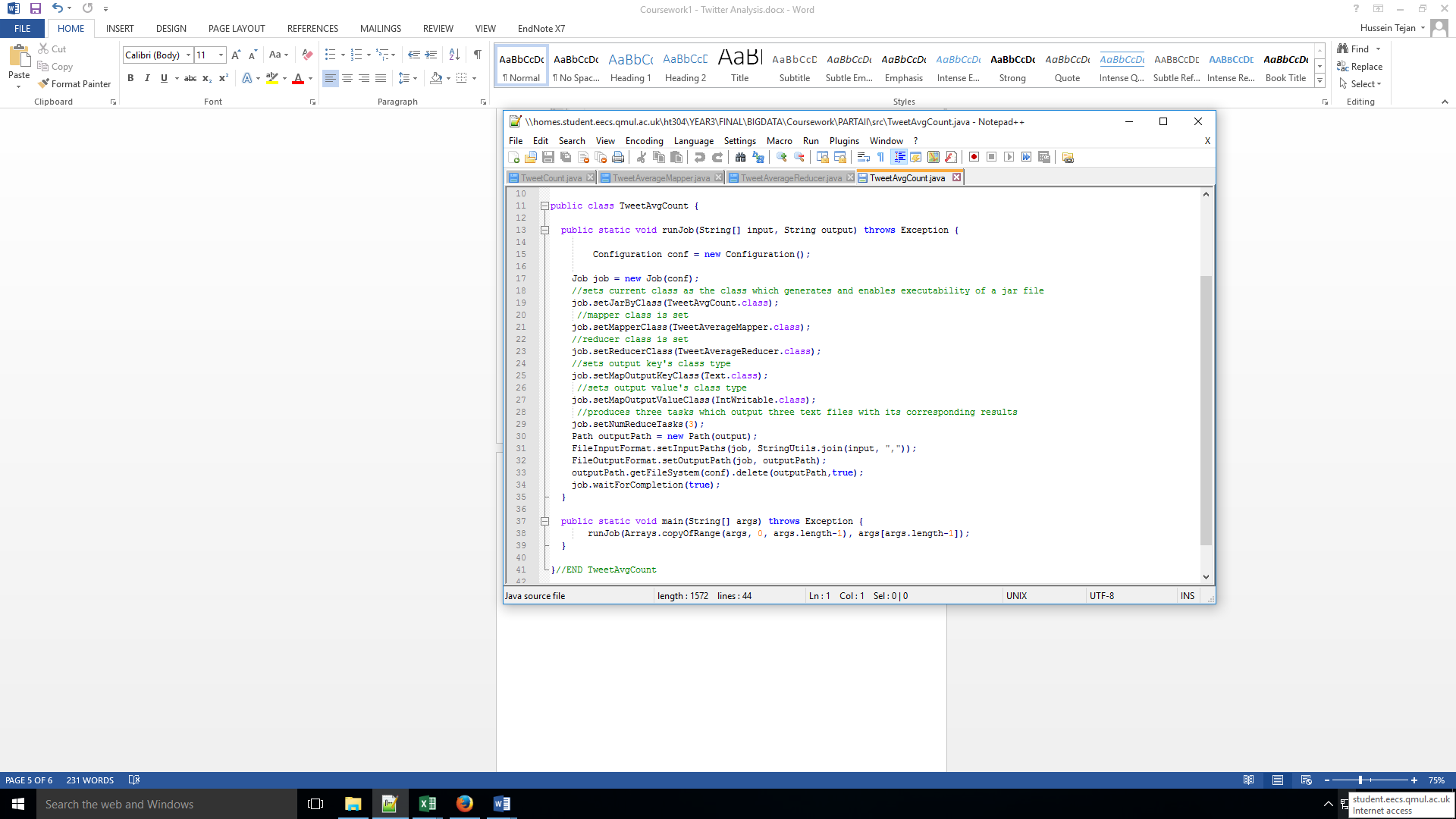
**(ii) AVERAGE LENGTH**

**TWEET AVERAGE MAPPER CODE:**

**TWEET AVERAGE REDUCER CODE:**



**TWEET AVERAGE COUNT CODE:**

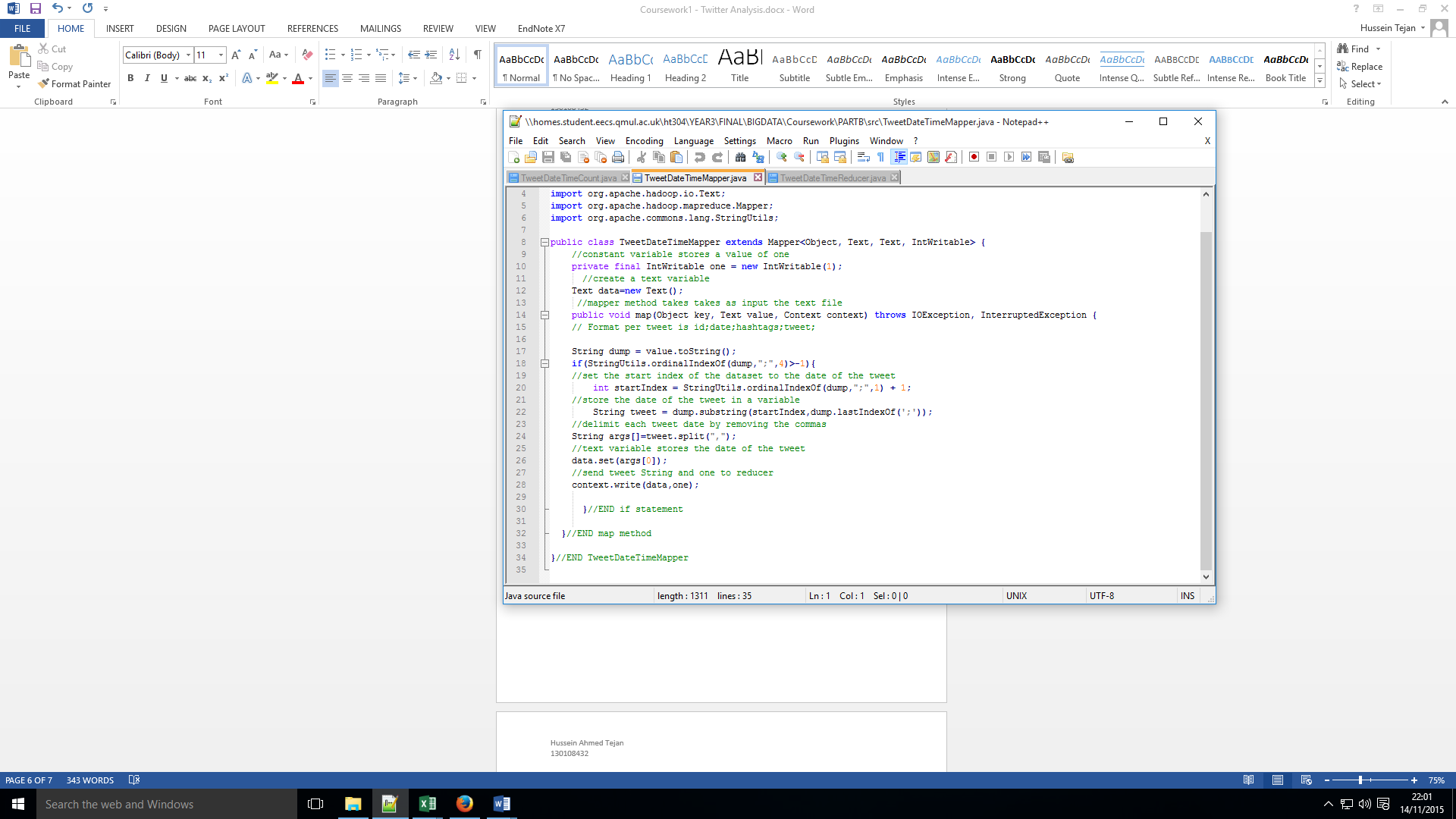


**EXPLANATION:**

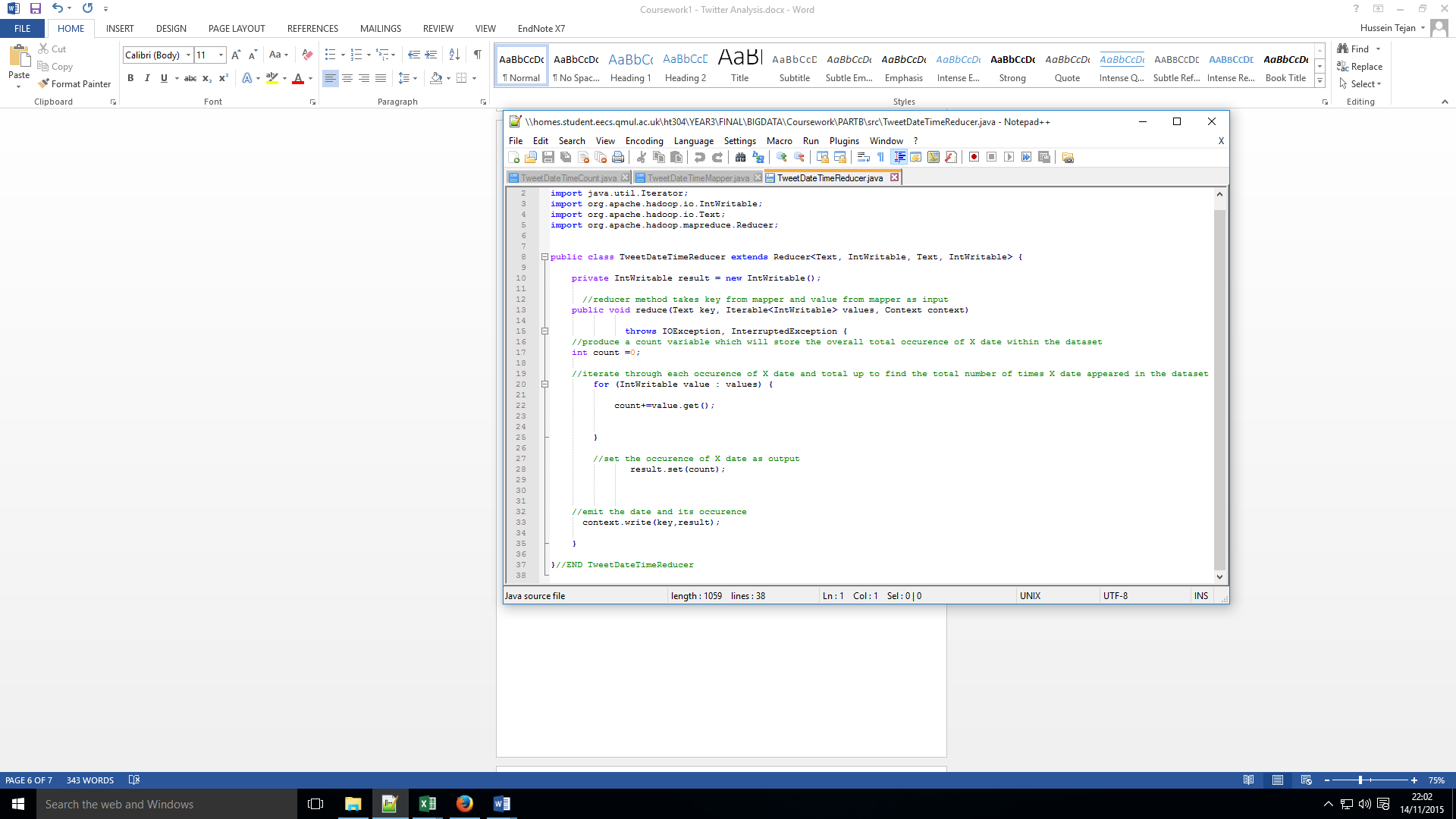
The Map Reduce approach with the second part of (**A) Content Analysis** enables you to retrieve the total number of tweet lengths and the total size of all tweet lengths totalled up. A key such as a string is used to denote the tweet and the tweet length is set as a value to be emitted to the reducer. Then you can find the average of tweet lengths by adding all the tweet lengths together and diving them by how many lengths exist within the dataset. The average is then emitted as a value along with “tweet” as a key.

**PART B- TIME ANALYSIS**

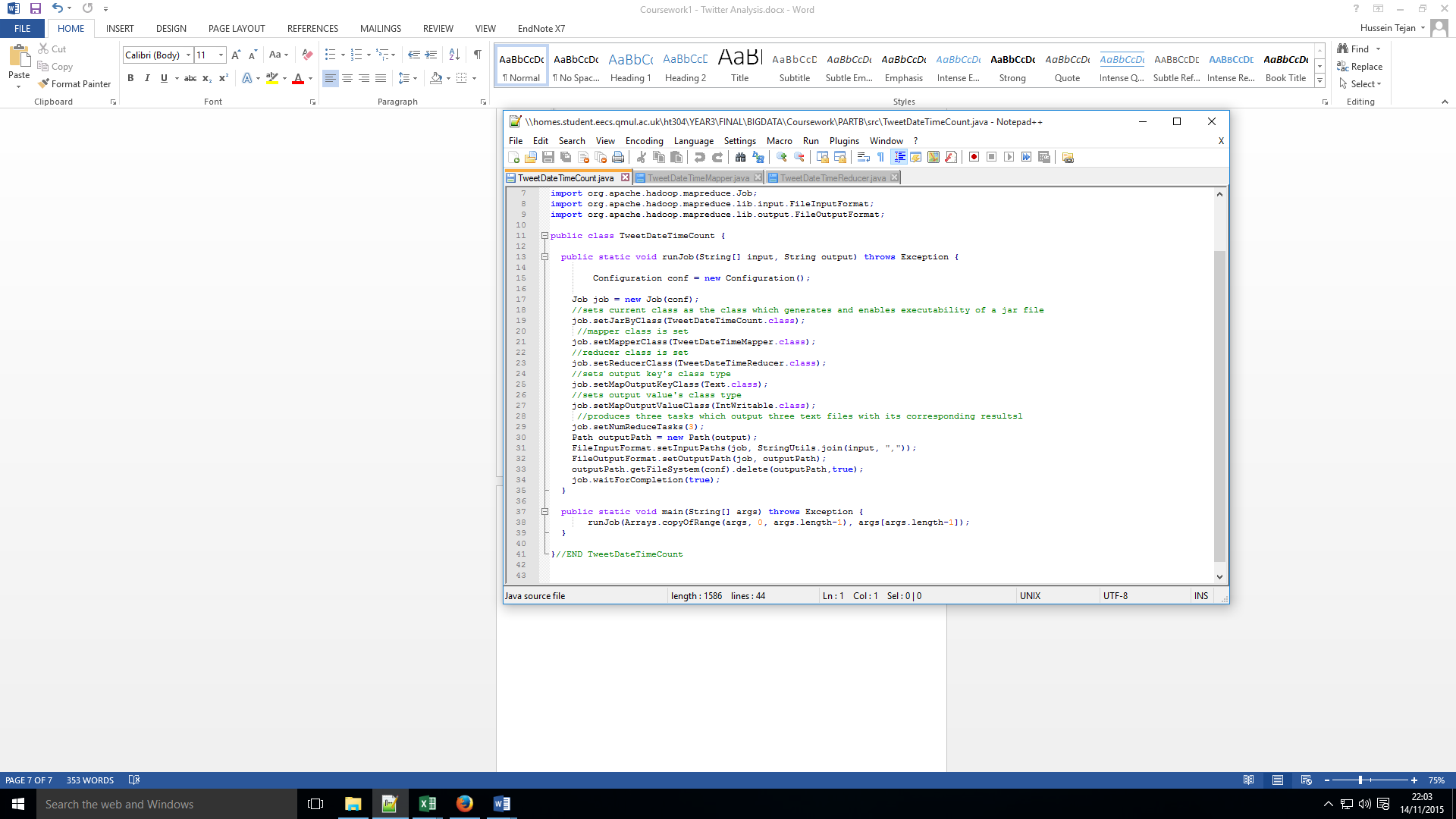
**TWEET DATE TIME MAPPER CODE:**



**TWEET DATE TIME REDUCER CODE:**



**TWEET DATE TIME COUNT CODE:**

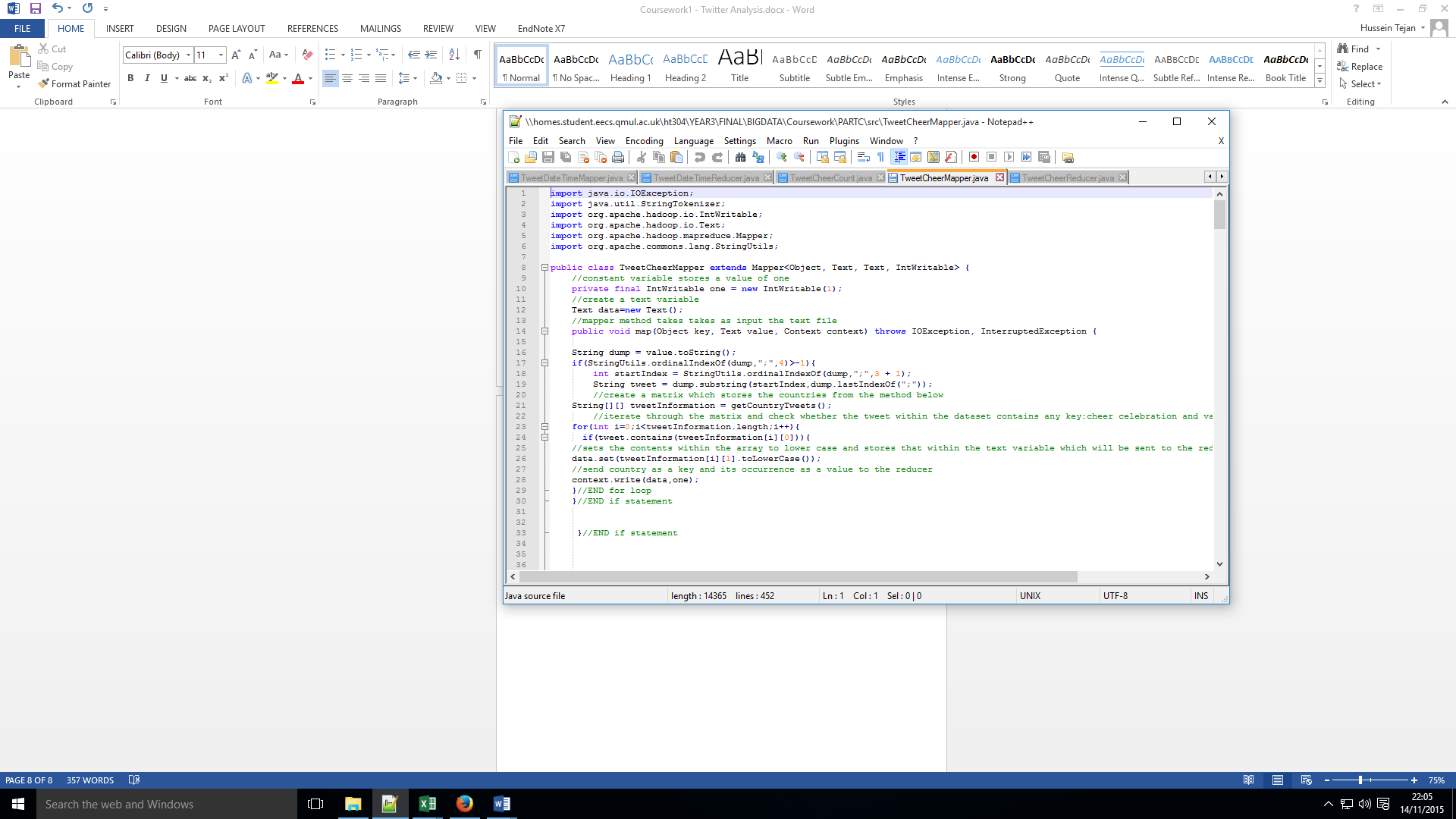


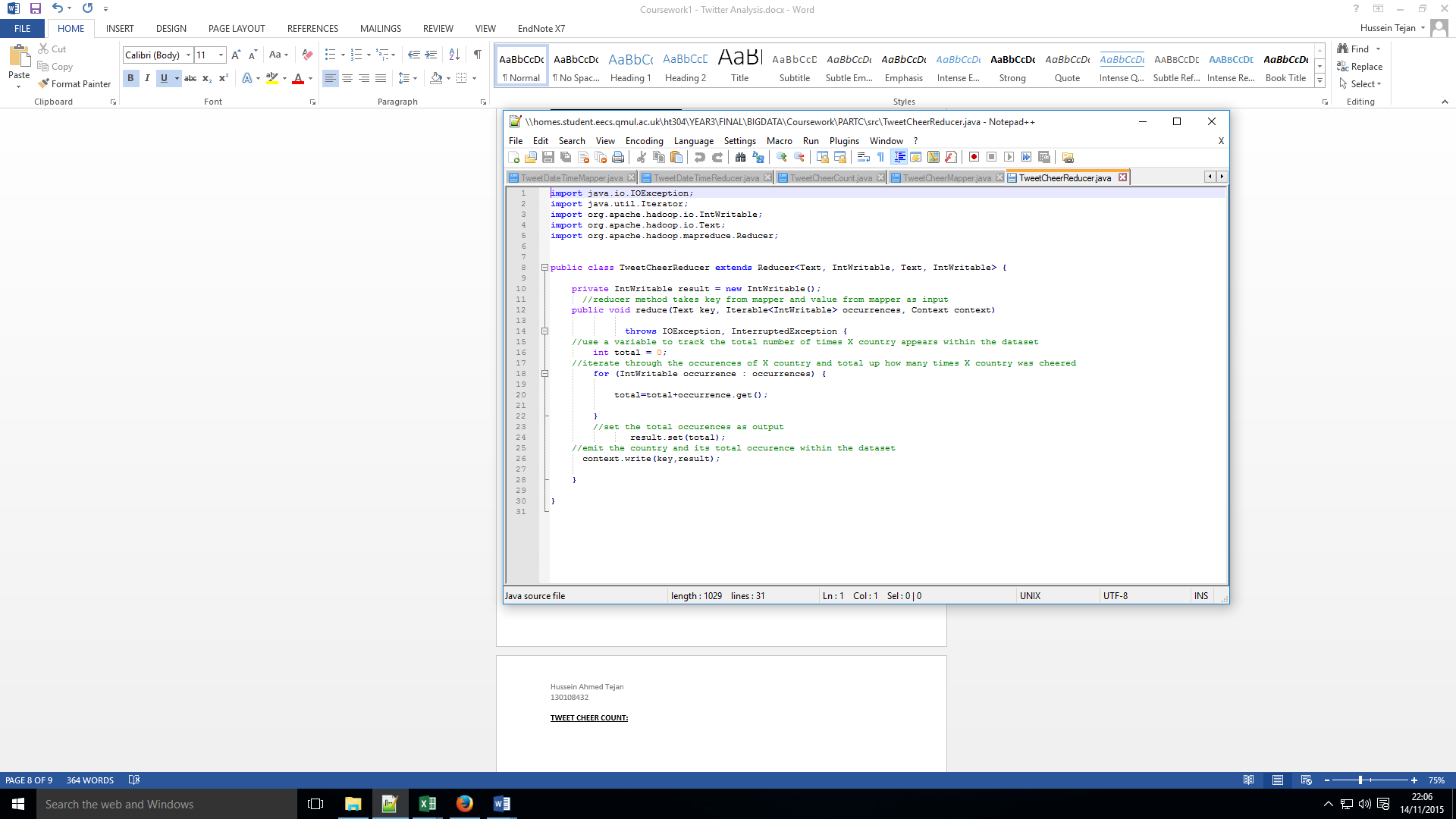
**EXPLANATION:**

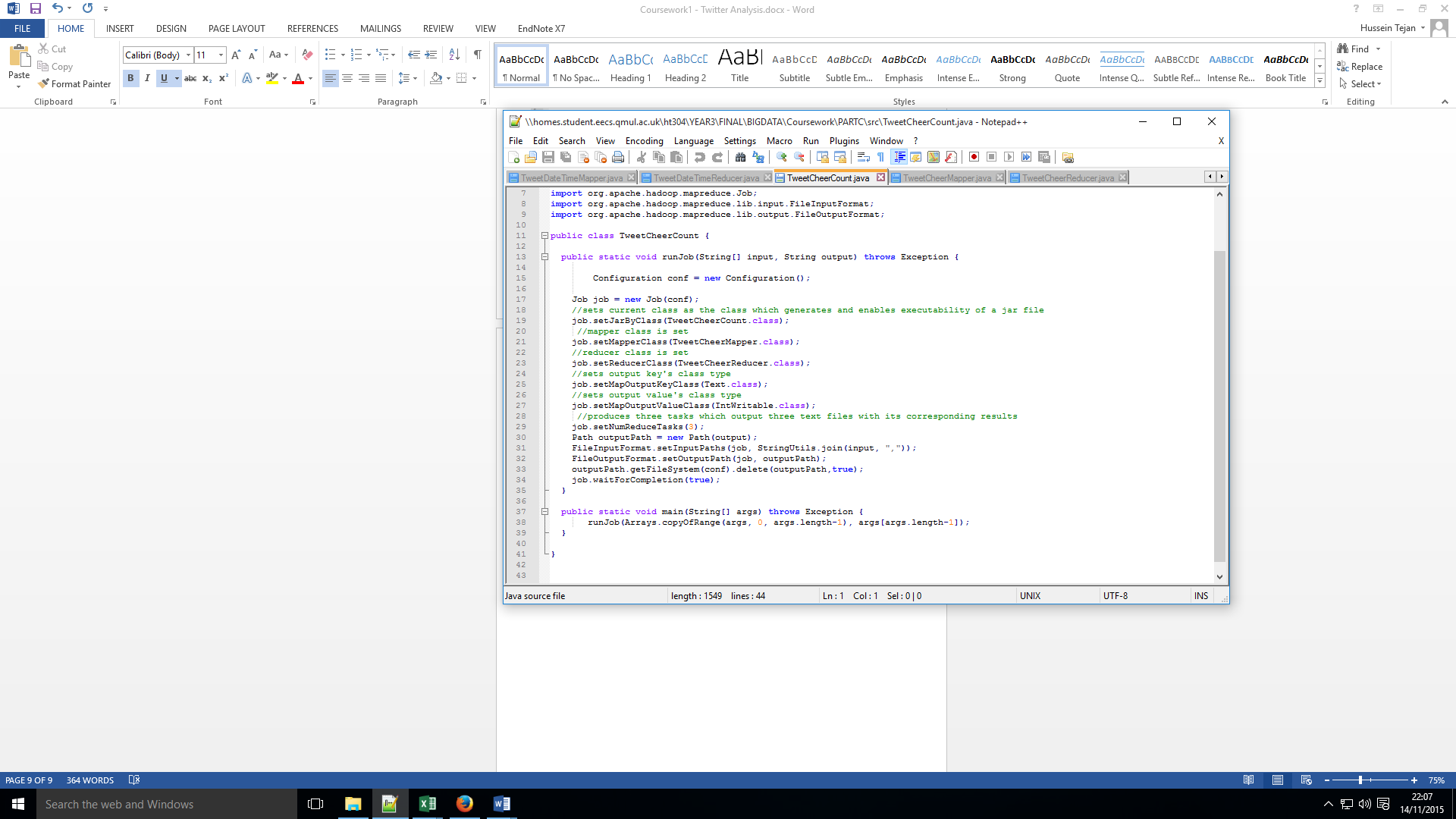
The Map Reduce approach with the second part of (**B) Time Analysis** enables you to retrieve the date that a tweet was posted and count each date and store an occurrence of each date within the dataset and thus, compare how many tweets were made on a particular date.

As the time series diagram show above, the number of tweets fluctuate varying day to day, sometimes increasing across a number of days and sometimes decreasing. Furthermore, the day which had the greatest number of tweet at a value of just over 180, 0000 tweets occurred 27 July 2012 - 29 July 2012.

**PART C- HASHTAG ANALYSIS**

**TWEET CHEER MAPPER COUNT CODE:**

**TWEET CHEER REDUCER CODE:**

**TWEET CHEER COUNT CODE:**

**EXPLANATION:**

The Map Reduce approach with the final part of (**C) Hash Tag Analysis** enables you to find the countries that had been cheered on via tweets that were posted, so each tweet would be compared against a matrix to see whether or not it had contained a cheer tweet of country X or country Y and if it had such content i.e: if gocountryX or teamcountryX was found, it would return country X along with the amount of times a tweet had been written with a cheer associated with it.

|  |  |
| --- | --- |
| **Countries** | **Tweet Count** |
| china | 2 |
| jamaica | 5 |
| kenya | 3 |
| niger | 2 |
| nigeria | 2 |
| sweden | 1 |
| bahamas | 1 |
| bulgaria | 1 |
| france | 2 |
| poland | 4 |
| canada | 9 |
| hungary | 2 |
| indonesia | 2 |

As you can see from the table displayed above, there were not many countries that had been found that had been cheered/celebrated i.e.: goCountryX or teamCountryX. However, of the countries that were found being cheered on in tweets were the countries above with Canada having the most amount of cheers.