

Capstone Project – Data Story

How does one go about creating a data story? You have some pointers from the material you've just gone through, but they’re probably a bit on the abstract side when you’re just getting started. Also, storytelling is an art, so you have to get your imagination going. Here are some pointers to get those creative juices flowing. In the following sections we will work step-by-step to create your first Data Story

* 1. Pick a dataset - ideally the dataset for your Capstone. If for some reason you want to do this on a different data set, you can find one on [Mode Analytics](http://blog.modeanalytics.com/five-public-dataset/) or [Google's public data sets directory](http://www.google.com/publicdata/directory), or pick another one you like from elsewhere.
  2. Get going by asking the following questions and looking for the answers with some code and plots:
     1. Can you count something interesting?
     2. Can you find some trends (high, low, increase, decrease, anomalies)?
     3. Can you make a bar plot or a histogram?
     4. Can you compare two related quantities?
     5. Can you make a scatterplot?
     6. Can you make a time-series plot?
  3. Having made these plots, what are some insights you get from them? Do you see any correlations? Is there a hypothesis you would like to investigate further? What other questions do they lead you to ask?
  4. By now you’ve asked a bunch of questions, and found some neat insights. Is there an interesting narrative, a way of presenting the insights using text and plots from the above, that tells a compelling story? As you work out this story, what are some other trends/relationships you think will make it more complete?

SMS spam dataset consists of text messages labelled as ham (legitimate messages) and spam. The first step to develop a spam filtering framework is to explore and analyze the ham and spam messages to find features capable of distinguishing between them. In this section we are determined to:

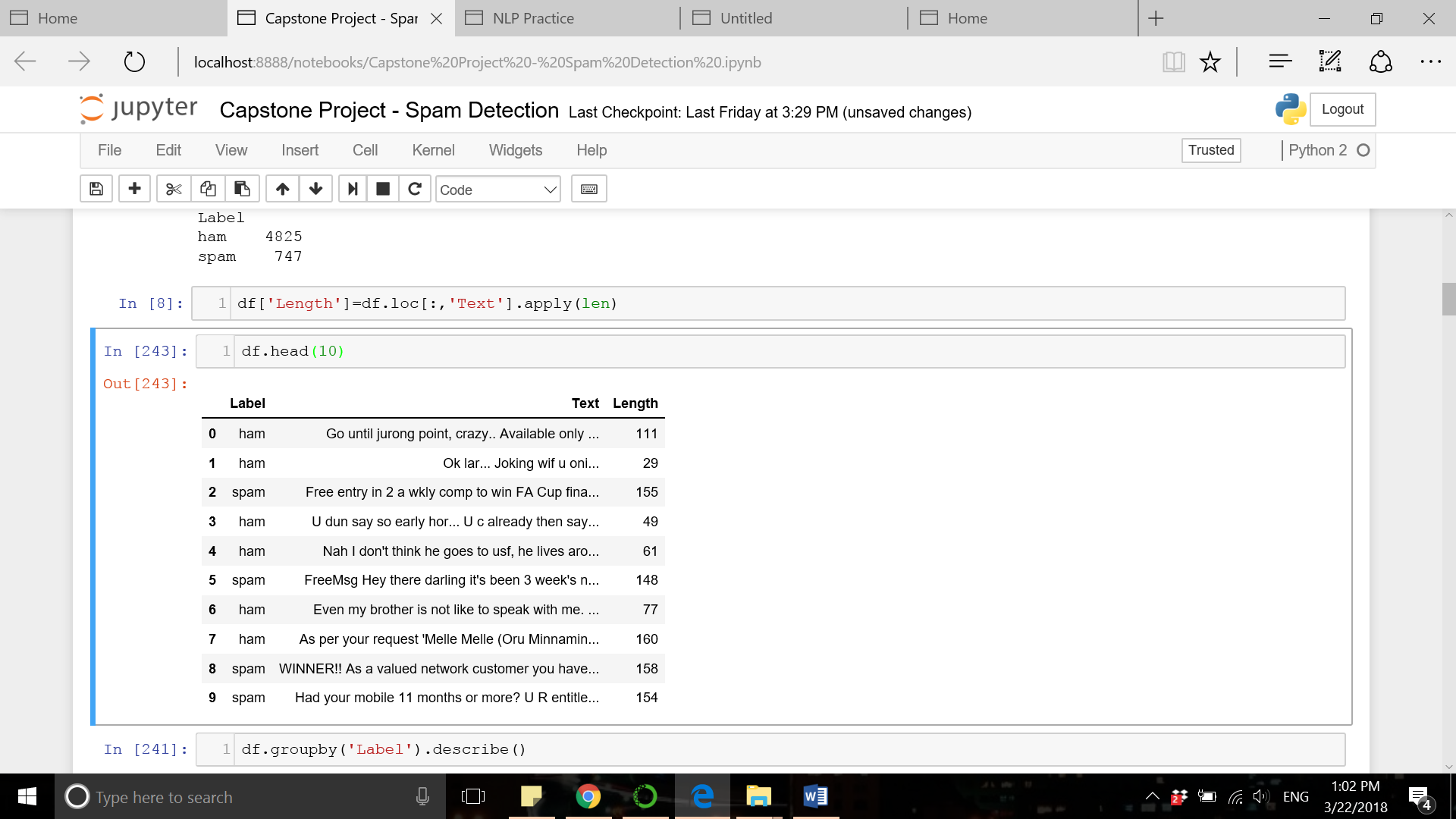
1) Explore the dataset and compare spam and ham messages based on the length of each message

2) Split the text messages into tokens (tokenization process) in an attempt to find useful features that could be used as an input to train a spam classifier framework

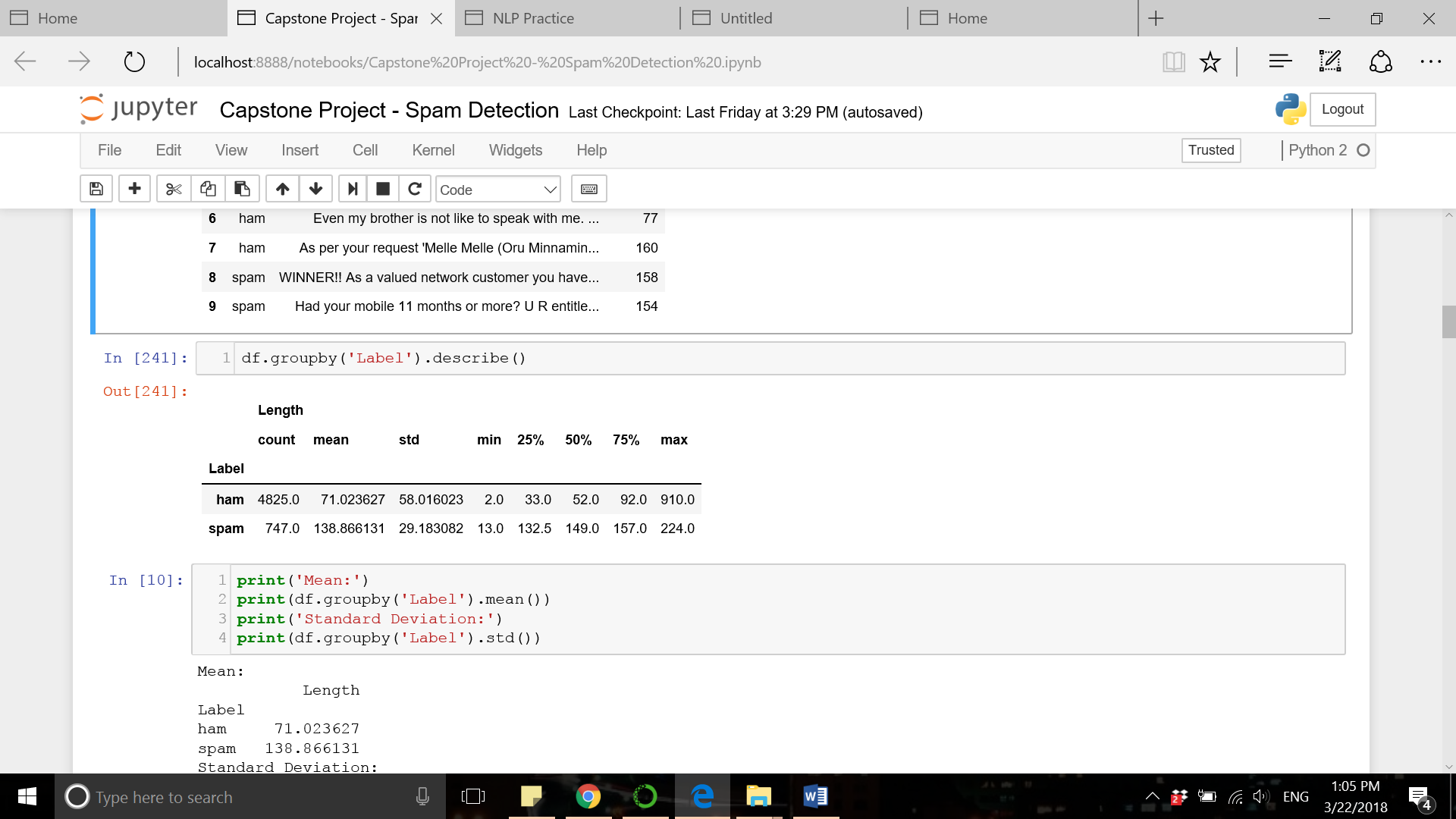
3) Extract the most frequent words occurring in the whole corpus.

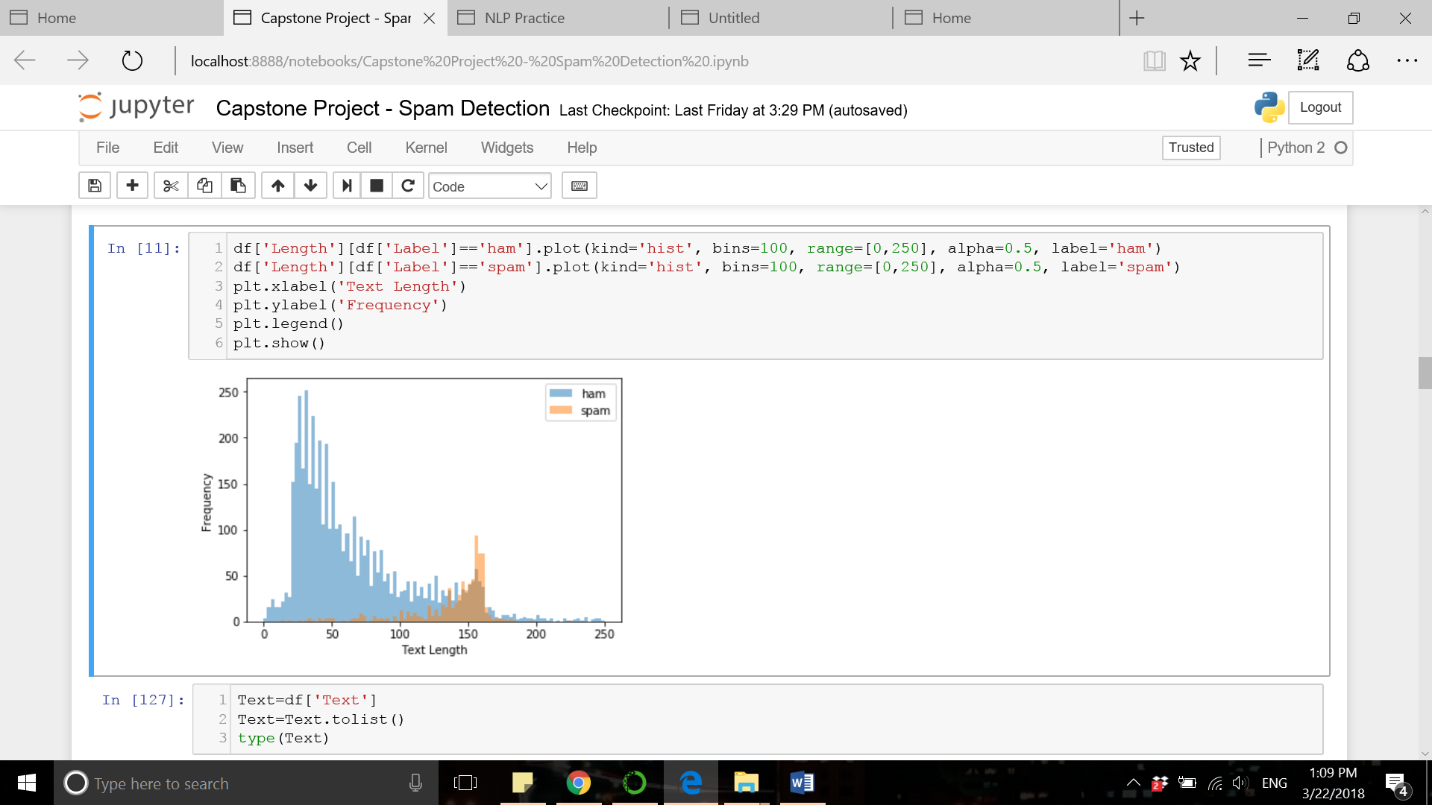
**Text length assessment**

The length of each text message was calculated and a column (‘Length’) was added to the data frame to include this variable:



The table below summarized the statistic for length in ham and spam messages:

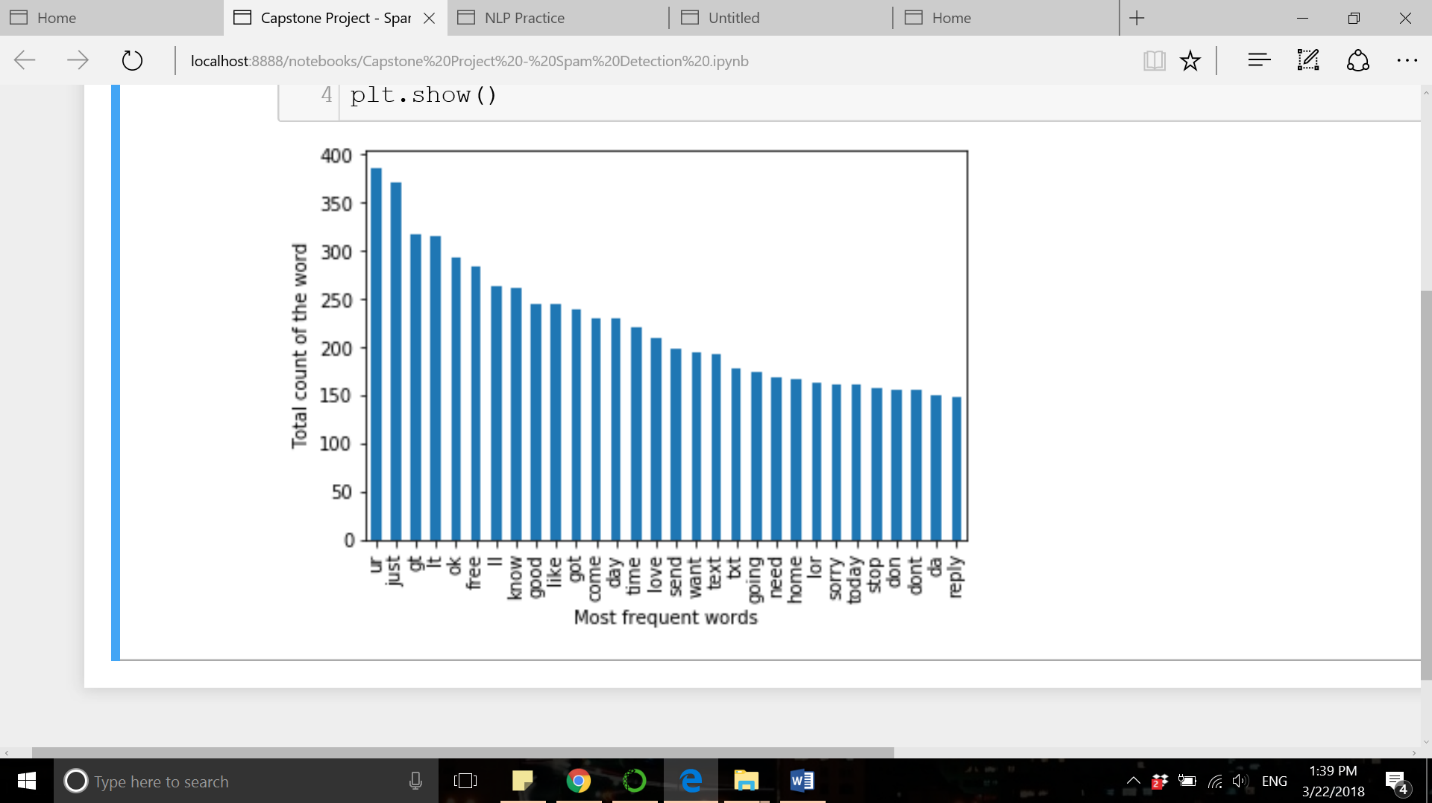


As observed above, the spam messages tend to be longer than the ham messages (mean values: 138 vs. 71 characters). We visualize the distribution of message length by creating a histogram:

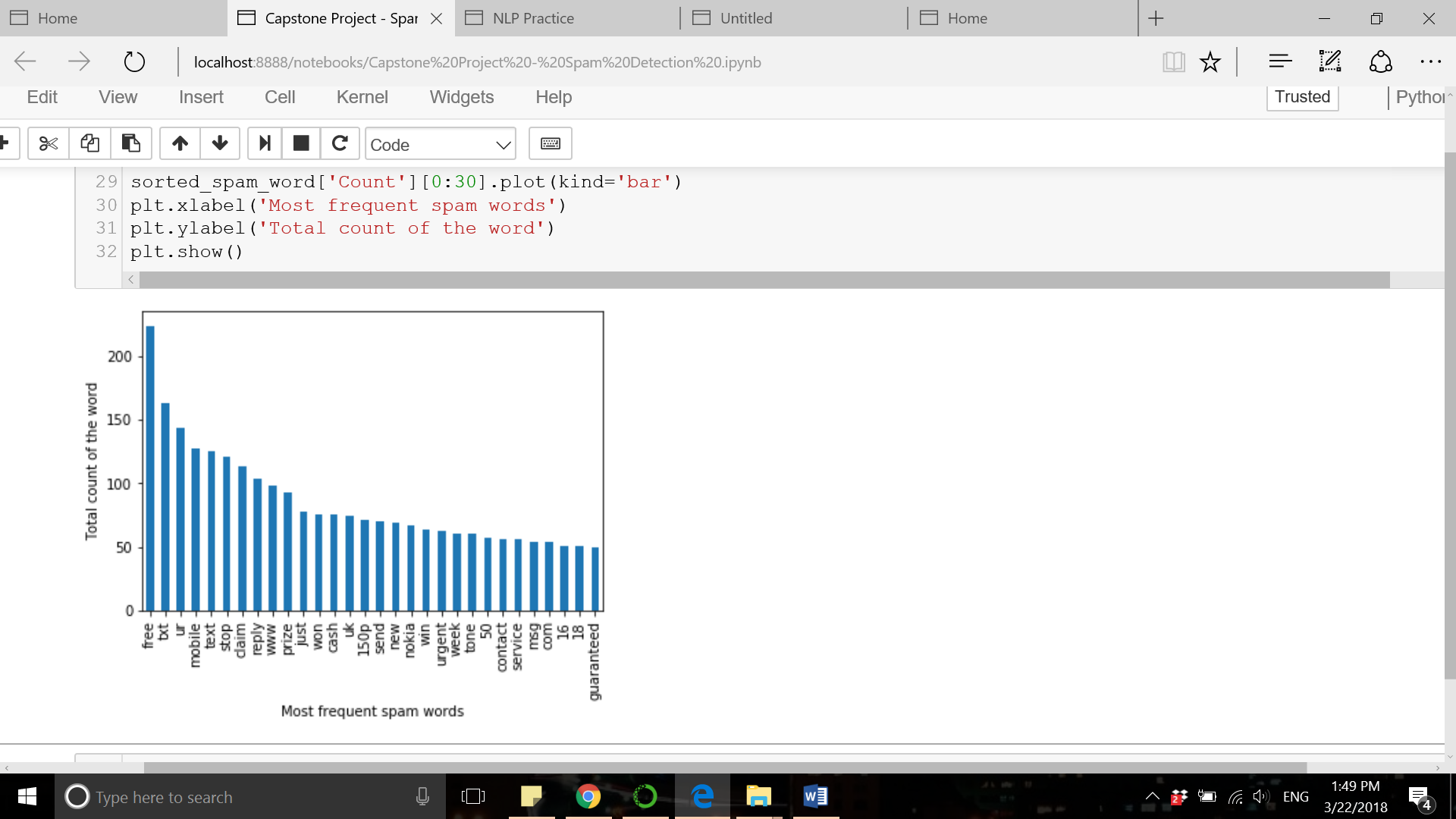
This is evident from histogram data that the ham messages are longer than spam messages and the length of message can be a feature to detect spam messages.

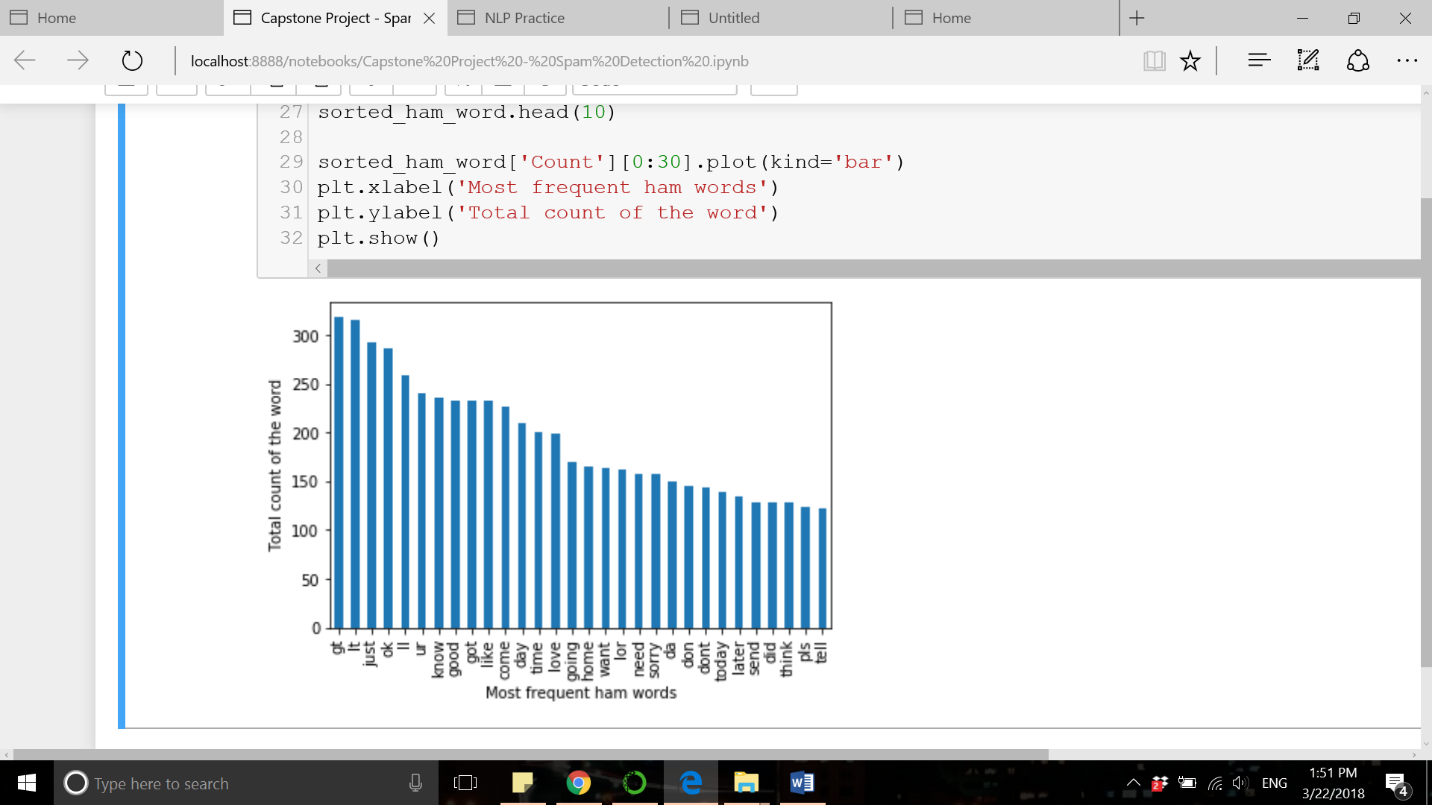
**Tokenization**

CountVectorizer function from sklearn library was used to convert the text messages to a matrix of token counts. We set the argument “stop\_words” in this function to make sure that stop words are removed during tokenization process. The bar chart below show the most frequent words in the corpus:



The bar charts below show the most frequent words in each group. Words like ‘free’, ‘claim’, ‘prize’, ‘reply’, ‘won’, ‘cash’, etc are frequent in spam messages and represent a common theme among spam messages.





Based on the word frequency charts, we realize that the spam messages have a common theme of texting about winning prizes to the receiver. So, creating feature vectors based on word frequency data can be the next step to vectorize the text messages and classify them based mathematical algorithms.