EEGR 565: Machine Learning Applications

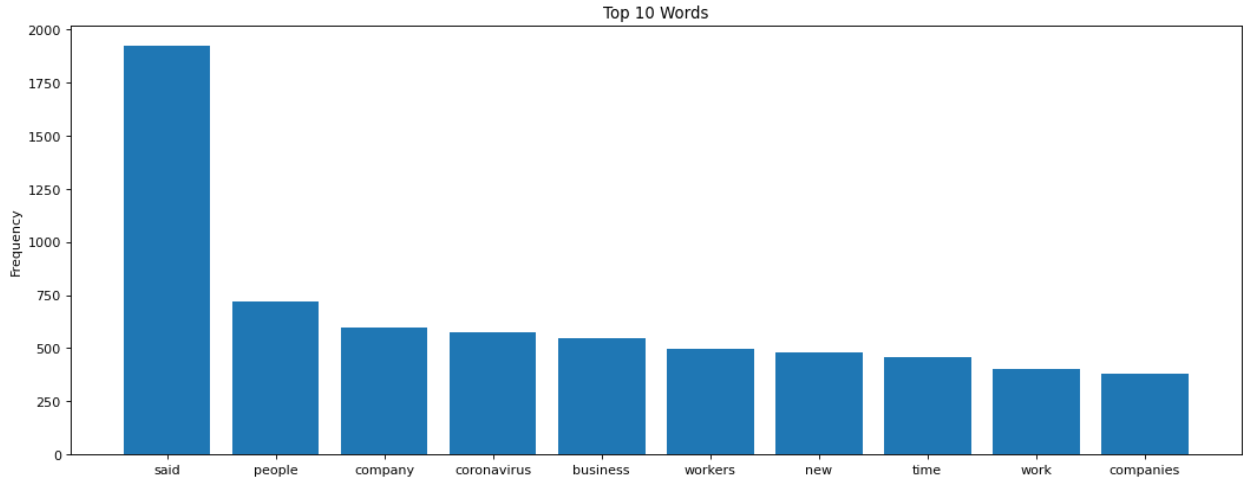
Assignment 1

As a data scientist, you are curious about how much "chatter" there was online within the "Business" section of CNN around the start of the pandemic (late March 2020 to early April 2020). You build a web scraper that runs once a day to archive articles on the "Business" pages for a few different themes, which include investing, banking, success, video games, tech, market etc. This data is provided in the file **cnn_data_4_5.csv**.

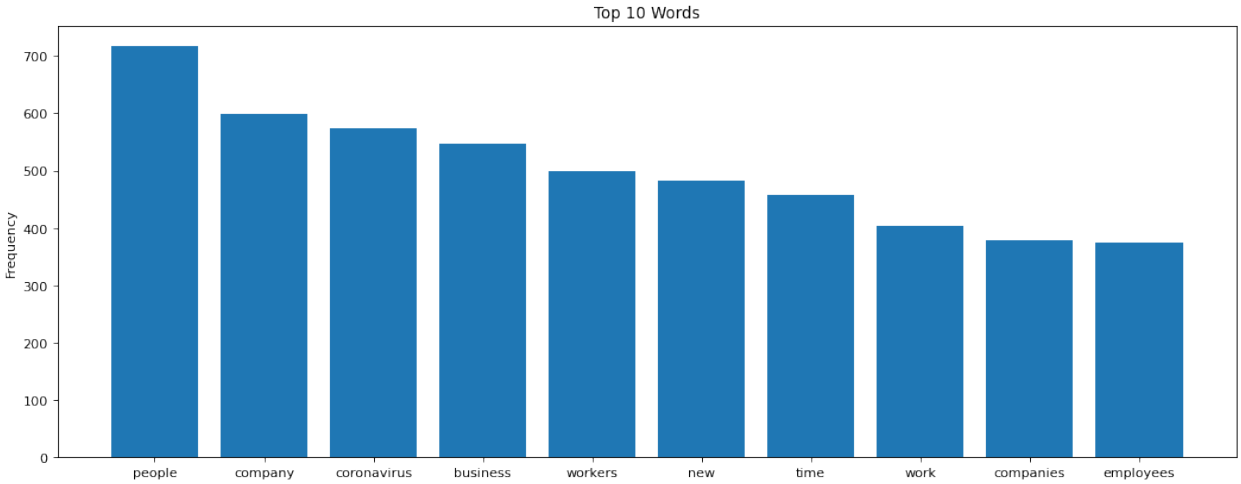
Create a python script to read this data (look up **read_csv** from the **pandas** library) and display the first few rows.

	url	title	body	date
0	https://www.cnn.com/2020/03/23/media/japan-abe...	Japan asked the international media to change ...	In the new system "Canton becomes Guangzhou an...	3/24/2020
1	https://www.cnn.com/2020/03/16/perspectives/us...	The United States is still too reliant on oil	Saudi Arabia's decision to open its taps comes...	3/24/2020
2	https://www.cnn.com/2020/03/23/investing/globa...	Global stocks and US futures rise as policymak...	The promise of unlimited support for markets f...	3/24/2020
3	https://www.cnn.com/2020/03/24/economy/china-e...	China is trying to revive its economy without ...	The country where the pandemic began was almos...	3/24/2020
4	https://www.cnn.com/2020/03/24/business/bailou...	Companies that binged on buybacks now seek bai...	Now, some of the same companies that binged on...	3/24/2020

Vectorize the body of each article using a max of 500 features. Get the frequency of each token (word) and create a histogram.



Notice how the token with the highest frequency count has a count that is much larger than all the rest. Treat this token as an outlier and filter out any token with a value larger than 1000.



Next notice how the word “coronavirus” did pop up in the number 3 spot. You are curious how many other terms related to the pandemic occur within the top 500 words (and their relative frequency), so you go to a website that lists terms associated with the pandemic and massage the data into the list of terms in **pandemic.txt**. Read this list (the first line is just a label describing the contents so can be ignored) and break up each word in each term (this can be done with the CountVectorizer). Filter out any tokens that do not fall within this pandemic word list to display the relative frequency of terms related to the pandemic used within the “Business” section of the CNN articles.