General information about data-

I have considered NBC news front page for my analysis for experiment 9, where I extracted all headlines from front page for the analysis of word content in headline.

Cleaning process-

In process of cleaning of data all numeric values are removed as it does not make any sense with words, all capital words are consider as it defines all nouns in the headlines, “stop” list is created and all words from “stopwords\_en.txt” are added to stop list. Also 3 more words were added to the list as they were conjunction words like the, and, of, in and they were the only words occurring frequently. They were removed as they would have made no sense on word cloud for analysis.

Bigram-

Bigrams were extracted and most frequently used bigrams were calculated but there were no bigrams occurring more than 2 times. Hence, adding then to stop list was net required but code to do that process is added to comment in the code.

Output/Result- Here is the result of the word cloud after all cleaning is done. Few things can be interpreted by this word cloud i.e. some hot topics going around world like problems with safety as both words are big and repeated more times. Some more problem with Housing, LGBTQ etc.

There is also occurrence of big company like Twitter, Uber, Netflix etc. this can’t explain actual problems but can understand that this companies are big and there something or other is going on with the this company may be some problems or growing big.

A close up of a piece of paper

Description automatically generated