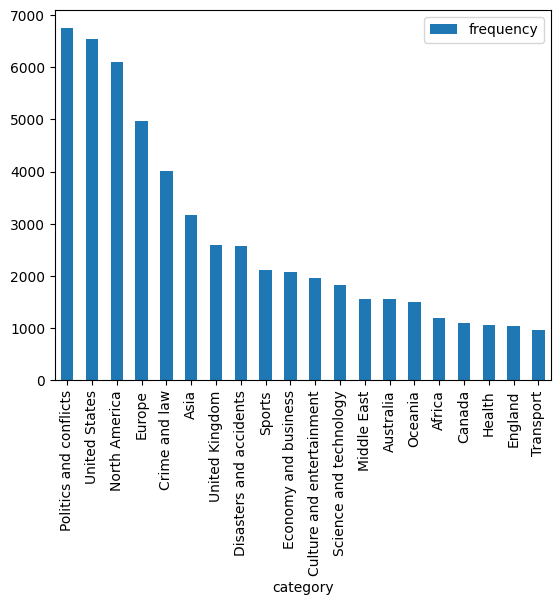
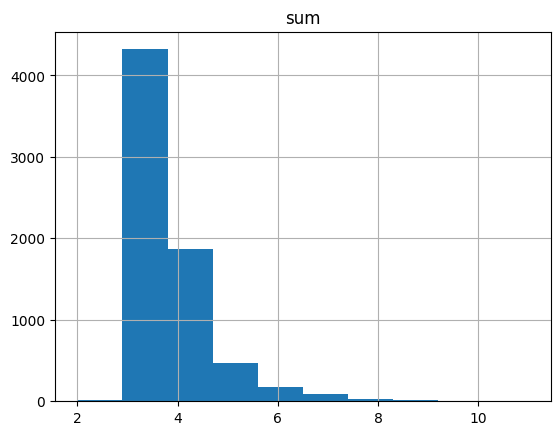
News categorization

After reading the csv, I analyzed the categories field. There were **3844** unique categories with **145860 instances.**

This is a histogram with top 20 categories, and number of appearences.



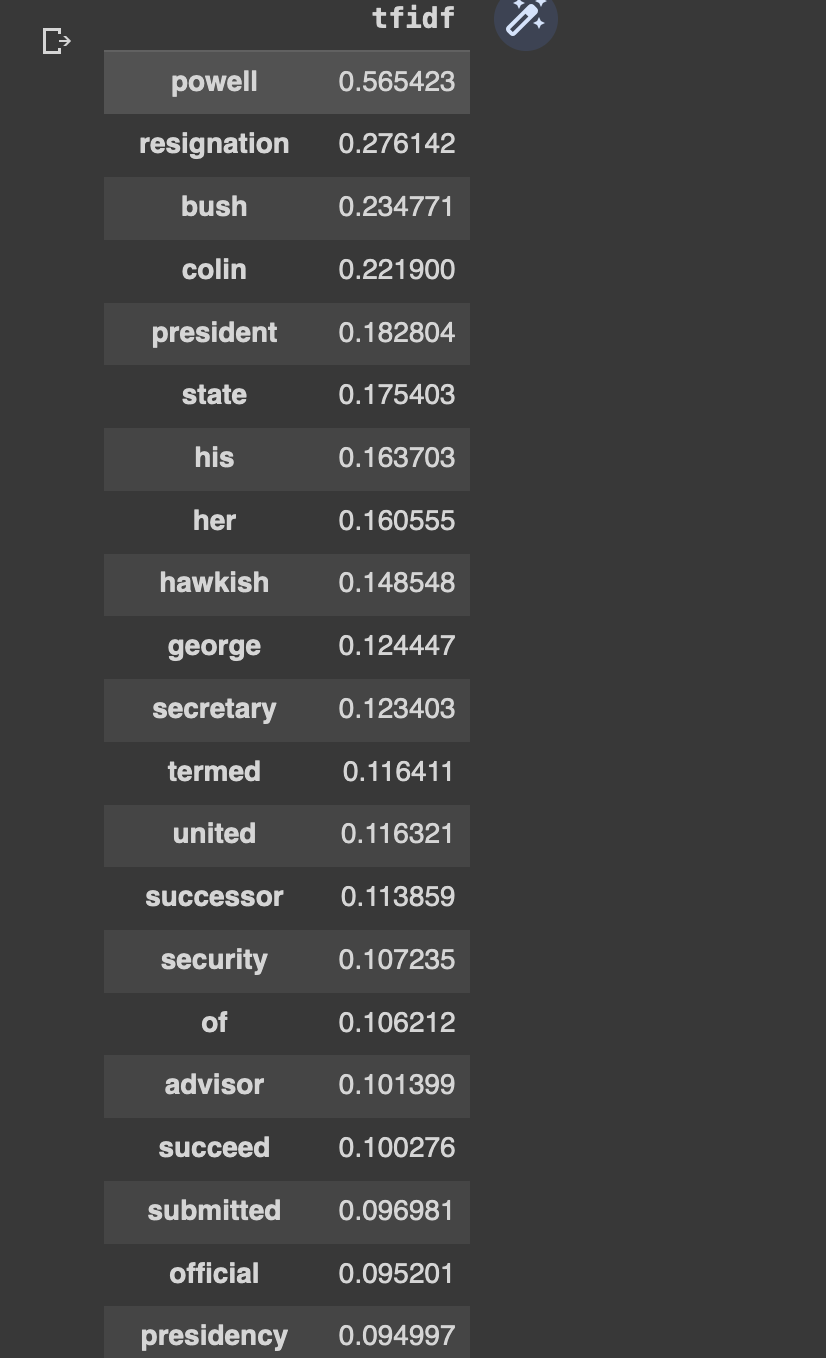
This is another histogram presenting the number of texts per number of categories. We can see that over 4500 examples, have between 3-4 categories assigned.



Also here is wordcloud with the words from texts.



For training a model, I extracted features from the text with tfidf. In this image we can see some words, and inverse frequency.



For training the model, I searched on the internet for multi labeling categorization.

The idea was to keep 100 categories at first, the most used and 16000 examples (after preprocessing).

I only used 20 categories and 10k examples. After searching the internet I also found a working repo with a model to train [here](https://github.com/skojaku/multilabel_knn).