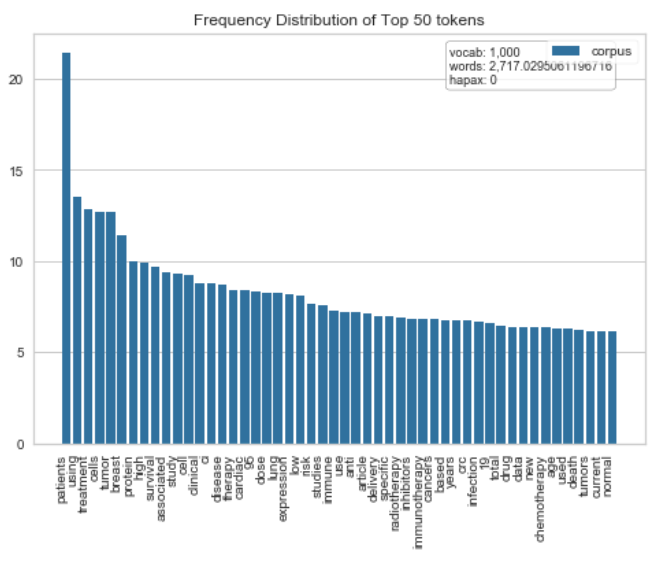
Analyzing the PubMed dataset

Ethan Go

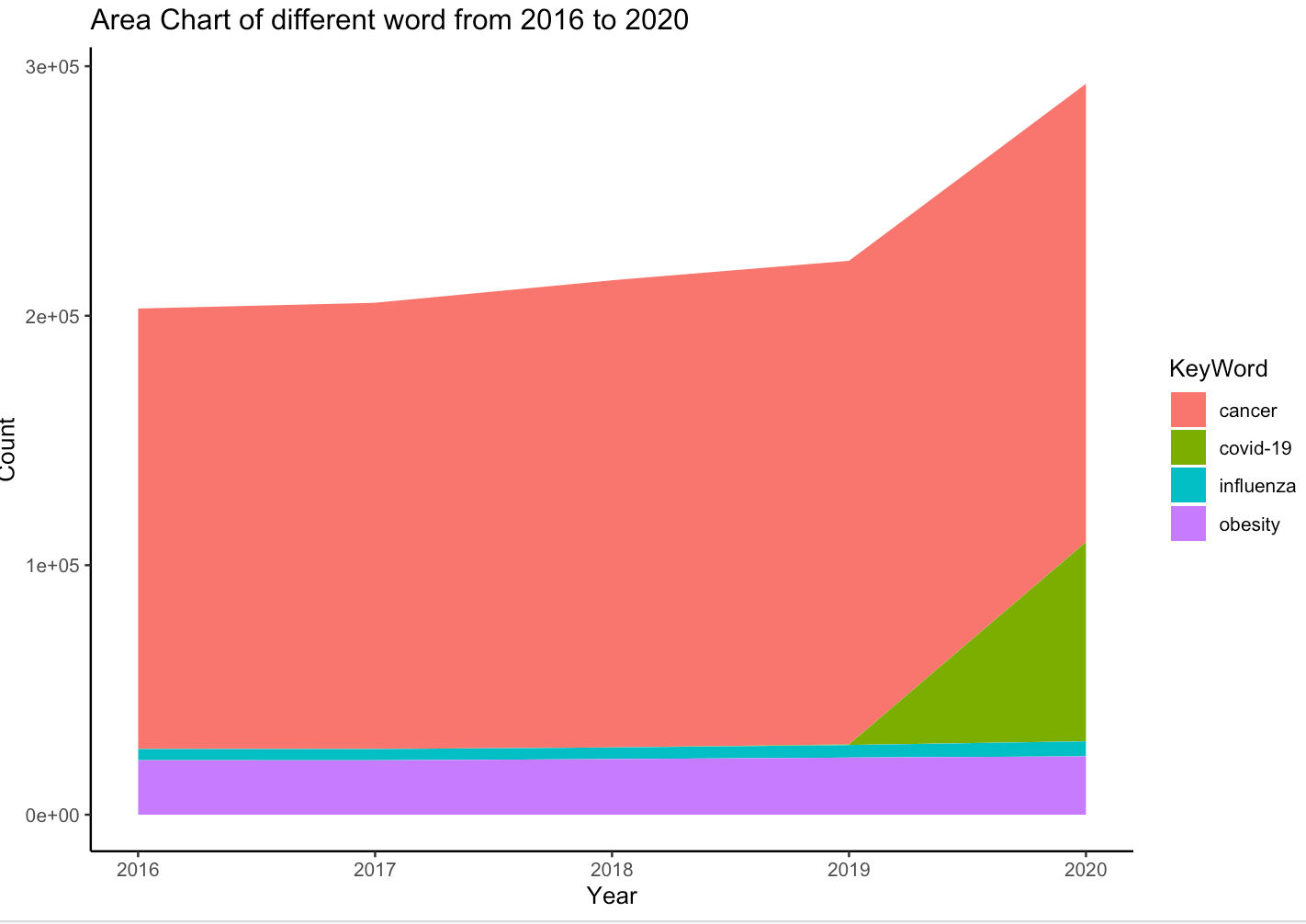
The PubMed datasets contain multiple research articles on different medical issues published by various researchers. Throughout the class, we have been exploring different ways to analyze and gain insights from these articles. Keyword analysis involves looking at each of these terms' popularity in the present age and examining their importance. One of the critical focuses of this research paper is natural language processing to understand the sentiment analysis of the different research papers on various topics. The topics focused on in this research papers are " "Obesity", "Cancer" "Covid-19"," "wearables", "Mental health", "Influenza".

In recent times, the COVID-19 pandemic has been a new and rising trend. This indicates the focus on the issue. In the same ways, analyzing the direction of different health related issues will. The trajectory of the research works also provides new insights and trends that may not exist before. One analysis that comes to mind is the theme and natural language processing to identify new keywords that may not exist back. Using a document term matrix allows the computer to provide insight on which words appear the most.



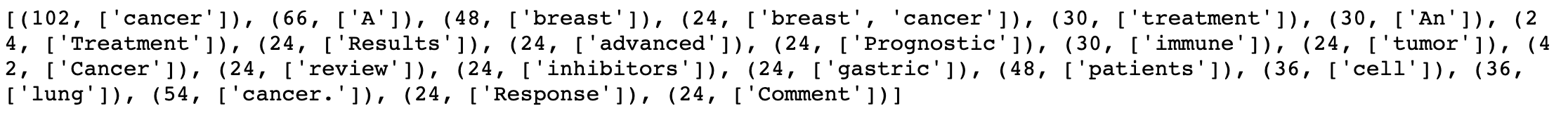
In the graph above, one interesting observation is the appearance of the word: 'patients', 'using' and 'treatment' as the top three words. This makes sense as all these terms generalize the topic. Other essential terms, such as 'protein' and 'cell,' give insight into what are the non-intuitive terms seems to be. Data analysis identifies the similarities between the keywords that were inputted. Through this, other application such as clustering may be used to

Another interesting research area is to see how popular some of the keywords have become in recent times. An area plot is one way to show this, as seen from the picture below:



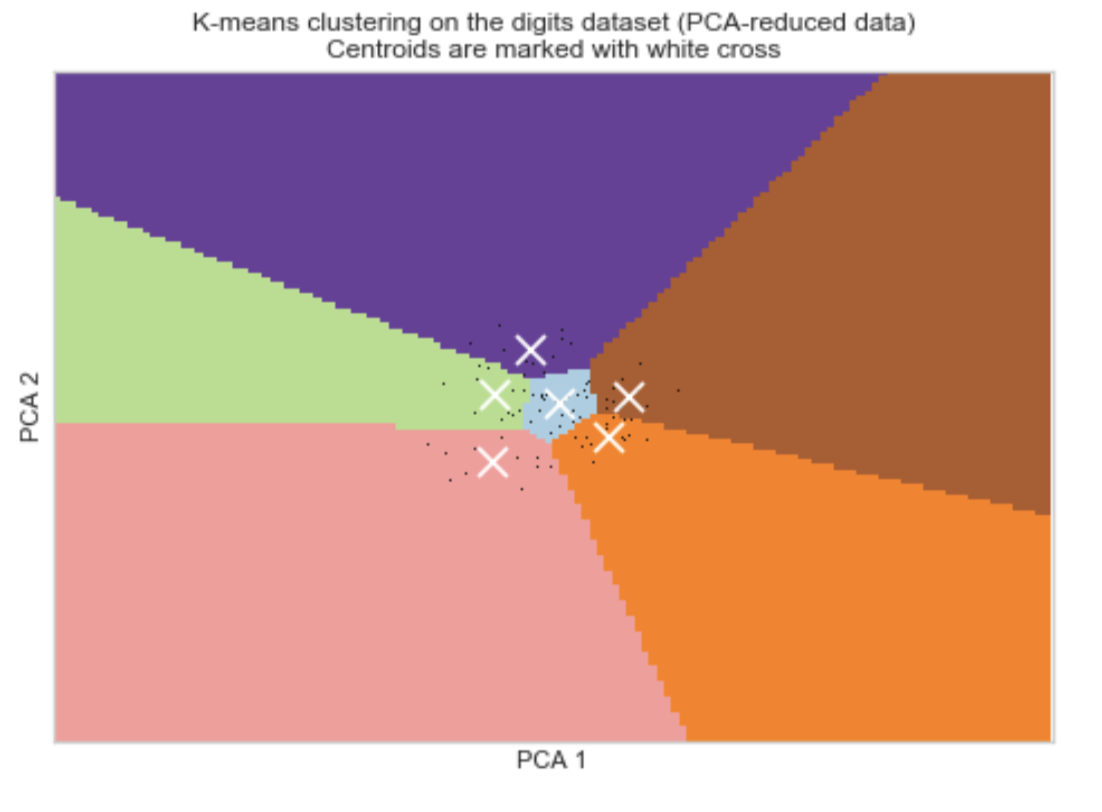
In this example, we can see the change of word in the articles from 2016 to 2020. The appearance of the covid-19 terms from 2019 to 2020 indicates that it is a rising term. The area plot is one of many data analysis tools that enable these interpretations in the PubMed articles. Using these tools allows researchers to see new trends and focus on the medical field. It also provides a comparison between which terms are more focused in the medical area. In this visualization, cancer has been such an important topic that there is a lot of research being put in that field than in any other topics.

Using one of the association rule mining, the prefix span, enables further insight as we were able to identify common frequent subsequences within the research article's title. One interesting observation from this result is breast and cancer occurrence, indicating that breast cancer appeared multiple times throughout the title. This is likely a subset of the cancer topic that was inserted and makes sense as the most occurring two sequence keywords in the title of the research article. The other results indicate the common occurrence of words in other titles. Most of these are not very useful in data analysis, as they do not provide further insights.



The PubMed dataset contains a medical research article and abstract that can is used in natural language processing. One of the possible applications in natural language processing is sentiment analysis, which is to identify which keywords have a majority positive or negative connotation in current research articles. This could prove useful in identifying the situation regarding these health issues. For example, suppose most of the research articles have a majority negative sentiment. In that case, it could be concluded that there is an outbreak or worsening of the situation somewhere around the world. On the other hand, if most research articles have a positive sentiment, there could be a new innovative medicine around the corner. In many cases, natural language processing is complex and requires a lot more further analysis to make sense.

Finally, implementing PCA and clustering allows an intuitive understanding of some of the words' connection in our datasets. Dimensionality reduction is a significant topic as it captures the ability to simplify the data while maintaining the variance within that data. Combining PCA and k-means clustering enables visualization of different words within this new space.



The x-axis represents the first principal component in the above visualization, and the y axis represents the second principal component. In this diagram, the cluster in the middle is more likely to represent likely words such as 'patients' and 'treatment' as seen earlier. The other clusters may represent other topics of interest. Clustering enables the partitioning of different segments of the data. The partition gives an idea of which words relate closely to another.

The PubMed dataset is only one type of dataset that exists out there. The truth is that there are many more different datasets that can be used to combine with the PubMed dataset to obtain further insights. Natural language processing is a very complicated process that applies mathematical concepts into practice. Machine learning and its algorithm should not be a complete substitute for human judgment, as machine excels at particular tasks. On the other hand, humans excel at very general tasks.