**Analysis of PubMed articles**

Prepared By: Neha Bais

MET – Boston University

Fall-2020

1. **Introduction**

The idea behind this report is to provide an in-sight for the article’s titles fetched from PubMed dataset based on their frequency, sentiments etc. For this task, we have extracted 9 years data from PubMed based on keywords – ‘Obesity’, ‘Influenza’, ‘Mental Health’, ‘Covid-19’, ‘Wearable’, ‘Cancer’.

The objective is to answer some research questions based on the analysis of this dataset.

1. **Scope of report**

The idea behind this is to analyze the data fetched from PubMed dataset and answer below research questions:

1. What is some knowledge extracted from trajectory of these research works in a duration of around 10 years?
2. Finding most frequent focus of academia based on these research keyword and analyze how this focus changed over years.
3. What are the sentiments about these keywords, which one are positive and which one are negative? Also, find what is the sentiment trajectory during 10 years period.
4. Find the similar research works.

In order to answer these questions, we have used some statistical tests to find out whether the distribution is same for all keywords and if not which two groups differ. We have used AFINN algorithm to see the sentiments for each keyword. Lastly, to see if there is any similarity between different research works, we have used LDA clustering method and Apriori association rule mining algorithm.

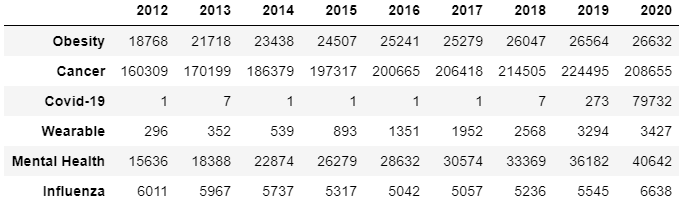
1. **Findings**

The findings will be presented in 6 sections based on different analysis of data: Analysis for each keyword over the years, Frequency distribution over years for each keyword, Statistical analysis, Sentiment analysis, finding similar research works.

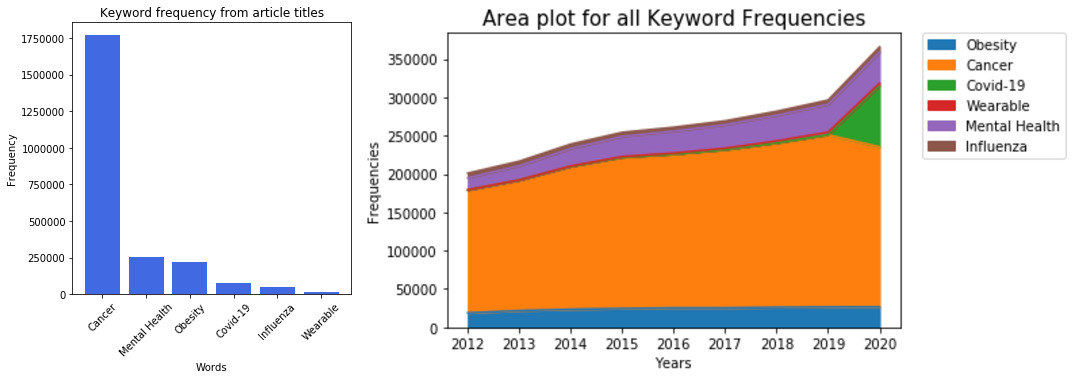
**2.1 Analysis for each keyword over years**

By looking at the count of articles based on these keywords over 9 years, we found that the keywords obesity, wearable, cancer and mental health have gained popularity over these years, however for cancer the focus is slightly dropped in 2020. It might be because in 2020, Covid-19 has become a famous topic of interest among people. If we see the results for Covid-19, it surges in 2020 and now is the hot topic in medical field. It is interesting to see some results for Covid-19 in previous years also, but those were not of the same Coronavirus which has caused a pandemic in 2020. Influenza results are approximately the same over all these years which can be justified as flu is a common topic irrespective of any year unless it is causing a pandemic.

Apart from these, there is an interesting observation about mental health academia focus which also got a spike in results in 2020, which can be obvious, as to focus on the study of mental health of people during this pandemic.



**2.2 Frequency distribution over years for each keyword**

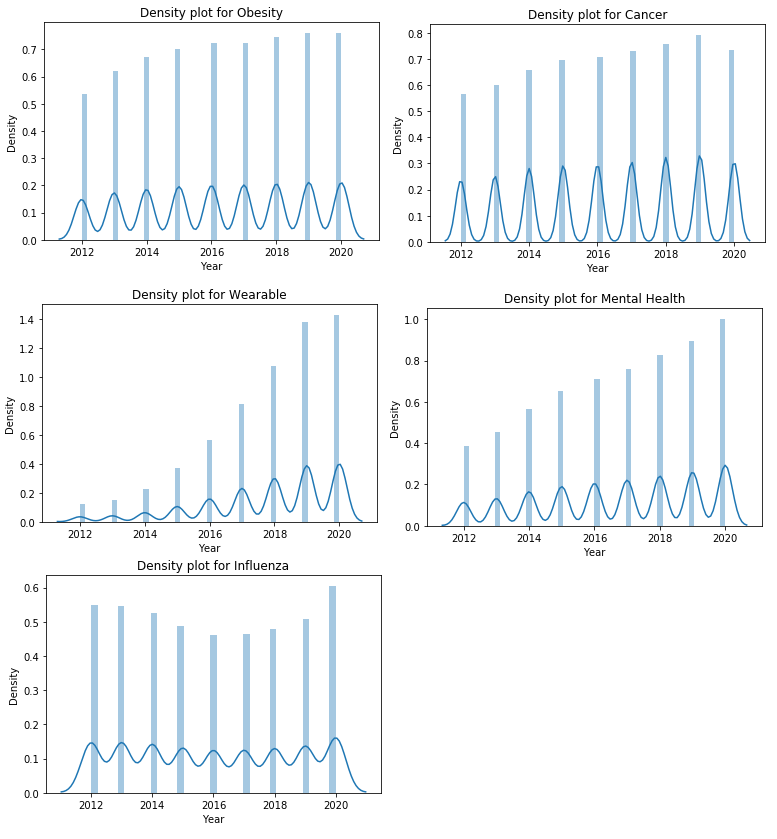


From the above distribution, cancer is the most frequent focus of academia over these 9 years. However, from the area plot we can see the focus dropped for cancer in 2020, it may be because 2020 is still left or there is a shift in research field in 2020 which is somewhat justified by Covid-19 surge in 2020.

Apart from cancer, mental health field got the second most popularity in research and then comes the obesity. Mental health is becoming a hot topic these days apart from Covid-19, which focuses on person’s emotional, psychological and social well-being. This is because now-a-days we can hear many cases of suicide due to depression, stress etc. So, this might be the reason the academia is focusing on the study of mental health of people.

**2.3 Statistical Analysis**

In order to find if there is a similarity in the distributions of these keywords over years, we performed statistical test. Below density plots gave a picture that the distribution is not normal for these keywords over years.



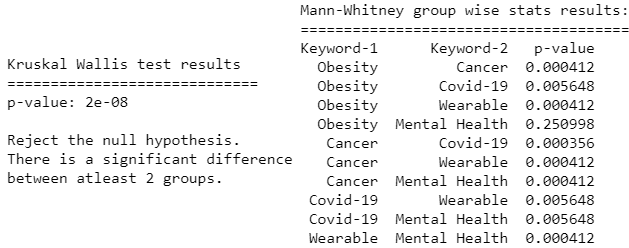
So, we used non-parametric tests, first *Kruskal Wallis* test to identify if all groups have a same distribution. We defined the null hypothesis as:

*H0: All keywords distributions are equal*

*H1: At least 2 keywords have different distribution*

From the results of this test, p-value came out to be very low and is less than 0.05. So, we rejected the null hypothesis and concluded that there is a significant difference in distribution of at least 2 groups. Then in order to identify which 2 groups differ, we performed the Mann-Whitney U test, which gives below results.

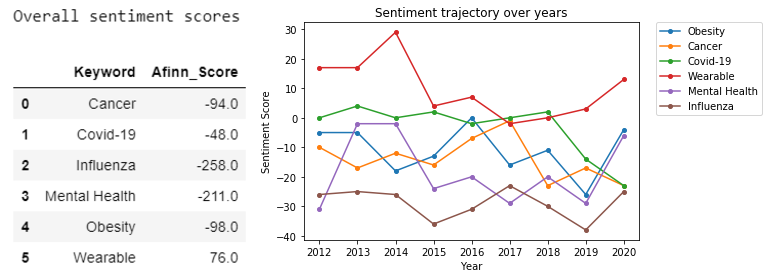
P-value was significant for all the groups, except Obesity and Mental Health, so we can say that there is not much difference in the academia research for these 2 keywords.



**2.4 Sentiment Analysis**

For finding the sentiments for each keyword, we have selected titles of the articles extracted for each keyword and then did some feature engineering like tokenizing the titles and removing stop-words, punctuations, stemming, converted them to document term matrix and then applied Afinn algorithm on them.

From the results, that we got, we can say that only articles for keyword *“Wearable”* got a positive sentiment and rest others got a negative score on an overall basis.



From the plot of sentiment trajectory over different years, we can see that for keywords ‘*Covid-19’ and ‘Mental Health’,* the negative sentiments got a surge in year 2020, which can be justifiable due to current pandemic situation. Also, we can see from the trajectory plot that only ‘*Wearable’* keyword got the positive sentiments in most of the years.

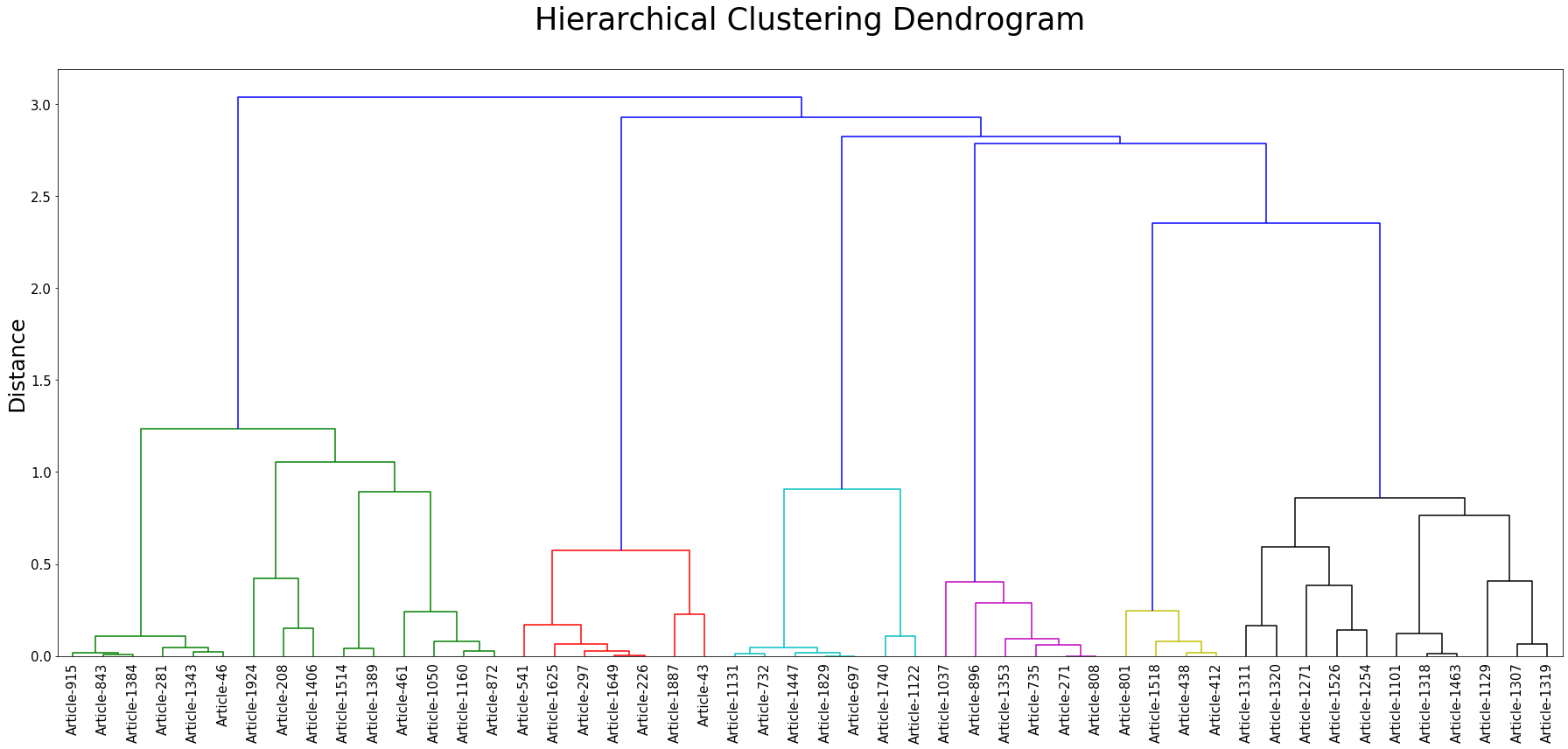
**2.5 Similar research works**

To identify which are the similar research works, we have used Latent Dirichlet Algorithm to cluster the data in 6 different topics. Again, as a part of feature engineering we did some pre-processing by tokenizing my data, converted them to lower case, removed punctuations and any special characters and stem them to their root form. Then created a document term matrix and passed it to my algorithm to cluster the data in 6 different topics. The algorithm gave a *document-topic* probability matrix which contains some keywords based on their weighted importance.

So, in order to extract theme from these results, we have sampled our data and selected the top 10 most important words in each topic, and then assigned them some theme based on these words and this way was able to cluster my entire data based on these 6 topics.

The below dendrogram shows the clustering results of randomly selected 50 articles from the data where each color represents the different clusters and the vertical lines i.e., branches indicates the distance between 2 clusters. Based on this dendrogram, we can see that the different articles combined in a cluster. The smaller the distance between two, the more similar they are based on the keywords within the articles.

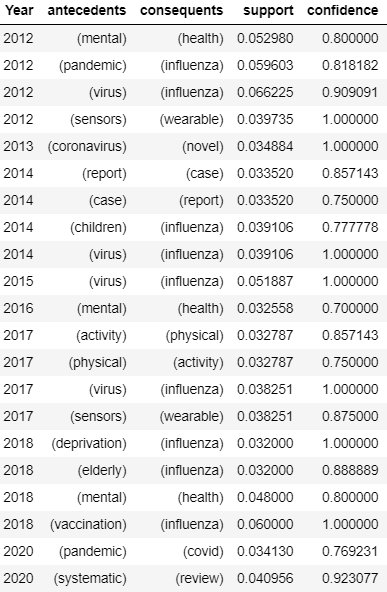
Also, if we compare two clusters, we can see that Mental Health (black cluster) and Cancer(yellow) are the ones having the least distance between them which we can interpret as the cancer patients suffer a lot and that impacts their mental health as well. Next comes Obesity (purple) and Pandemic (blue) and then Influenza (red) and at last the Wearable (green) serves as the root node which can be illustrated as wearable devices are used to monitor and study person’s health and serves as the top node in this hierarchy.



**2.6 Frequent patterns of words for each year**

In order to find what are the association between frequent occurring terms each year based on the titles of articles, Apriori algorithm was implemented with a minimum support of 0.03 and confidence level of 70%. The results that I got is illustrated below. It is interesting to see even though most of the research work is based on cancer, cancer keyword did not appear in frequent occurring terms. *Mental – Health*, *Virus – Influenza, Physical – Activity* dominated in most of the years.

So, if we inference these results, let’s say for 2020 one, we can say that around 76% of time when *COVID* term occurred in research, *pandemic* term also occurred.



**3. Conclusion**

With this analysis we found that the most frequent focus of academia over these 9 years was based on cancer and the research count was mostly increasing from 2011 to 2019, but in 2020 its focus dropped because of Covid-19 as the hot-topic due to current pandemic. Looking at the sentiments of the research works for these keywords, only wearable got a positive sentiment, rest others were showing a negative sentiment which is quite obvious as the diseases are mostly related as a negative impact on humans. While finding the similar research works, we found that Mental Health and Cancer got the top score and articles with theme Wearable are identified as the root node in the hierarchy of clusters.