Introduction:

This paper is based on finding insightful knowledge from the medical records gathered from the free archive for medical data which is the ‘pubmed’ dataset. The goal of this study is to identify the distribution of data in the downloaded articles and try to identify meaning information related to specific keywords. The methods used here are statistical analysis, feature engineering, supervised machine learning models, theme extraction and association rule mining.

Tools used: R, Python

Procedure for data download:

The data is downloaded from the pubmed website by using ‘EasyPubmed’ which is a R language library specifically developed for scrawling data from the pubmed website. Using this website scientific articles published for last 10 years, from 2011-2020 are downloaded. For each year ,the search keywords are used to target them. e.g. ‘cancer and date’ this will download data for the specific year mentioned and related to word ‘cancer’ and the other keywords were covid-19,mental health,wearable,influenza,obesity.

Feature Engineering:

Feature Engineering is the process of preparing data for applying machine learning algorithms on it. The first step after downloading data is to clean and preprocess it to prepare it for the further processing. The main goal of this study is analysis the variation of the data over the years, hence the date dimension should be extracted from the downloaded articles, also the further processing will be based on text data which is the title and abstract of the articles. Hence by using R language, year, title and abstract is extracted from the articles and stored into a separate csv file which can be used for further processing in python.

Statistical tests

To choose machine learning algorithms, we first have to identify the distribution of data. To identify if the data has normal distribution , Shapiro test is used. If the p-value is greater than 0.05 which is chosen as alpha generally. The test conducted in this study shows that the distribution is not normal and hence for further analysis normal distribution should be used. Then to identify if the distributions are independent pearson’s coefficient is calculated ,it returns value ranging from -1 to 1 indicating strong negative or positive correlation between variables. This test was used to test if there is correlation between obesity and covid-19 data and the value returned was less than 0.5 indicating very little correlation between both.

Next steps before performing theme analysis:

The downloaded data has year and titles of articles published during 2011-2020.To further analyze the data, it should be cleaned meaning all preprocessing should be performed which includes, removing special characters, punctuation marks, numbers, extra white spaces and then removing stop words, stemming and lemmatization and then storing it with respective document ids in the data frame. This process was performed in python using nltk library which has various functionalities to perform word tokenizing and other natural language processing tasks.

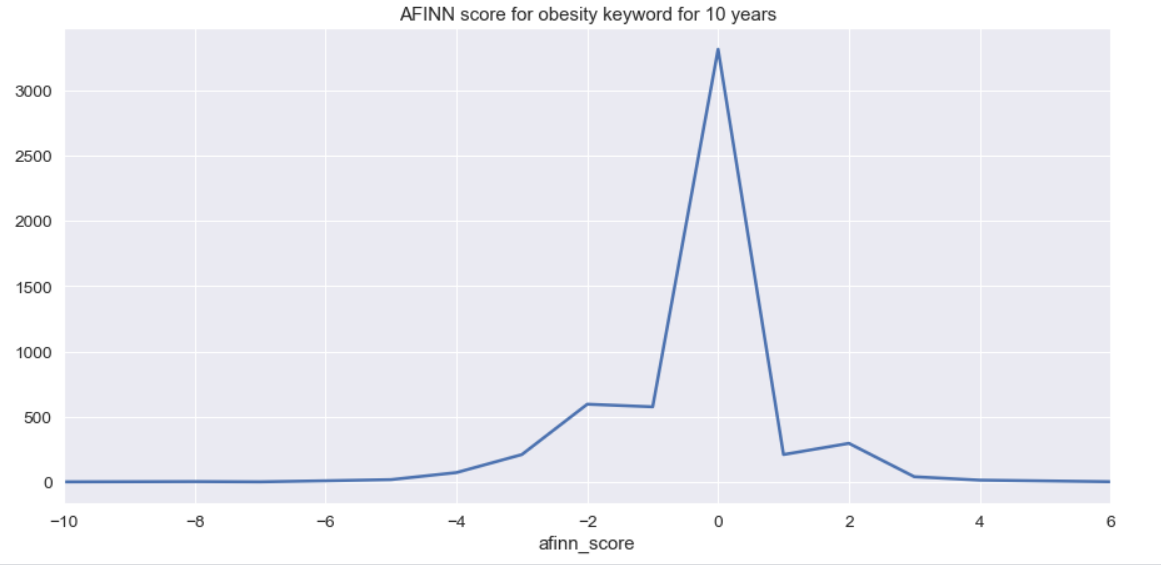
Analyzing focus of documents over the year

The number of articles for each keywords for each year are calculated using python pandas function and plotted against year in the plot.

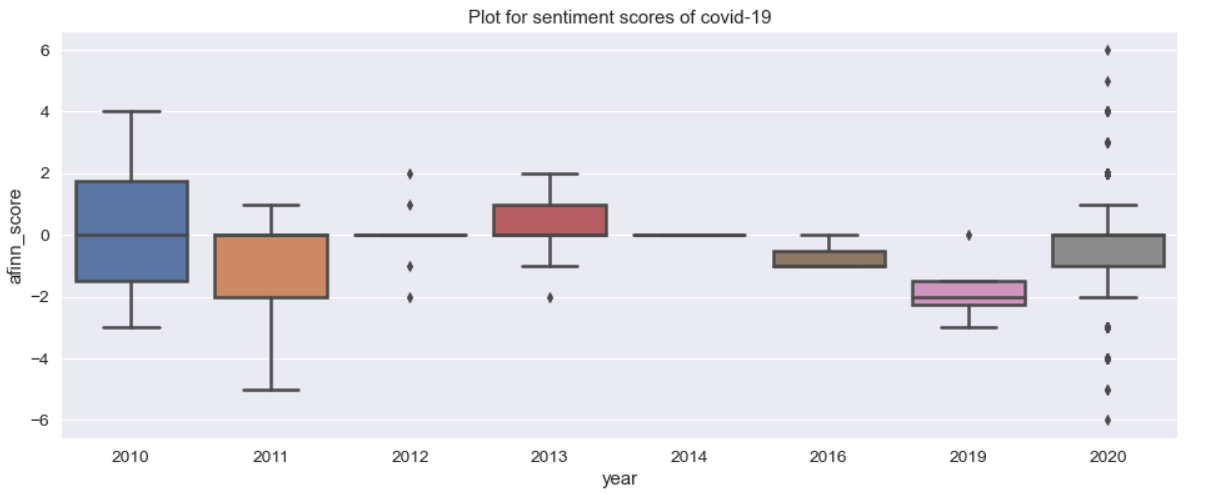
The focus on coronavirus or covid-29 has increased from 2019-2020, whereas for other keywords like cancer and obesity the majority articles were published in 2015-2016.But again total number of articles for all the keywords is highest in 2020 which is an interesting fact observed though the visualizations.

2.Sentiement Analysis

Analyzing the polarity or sentiment of the sentence helps in gaining important information which can be used for decision taking process in real life applications. To identify sentiment for all the articles the data is tokenized into separate words and then afinn library in python is used to calculate the sentiments related to the given keywords over 10 years of data.The plots show that Majority of articles related to obesity have neutral sentiment and varying sentiment from slightly positive to slightly negative.

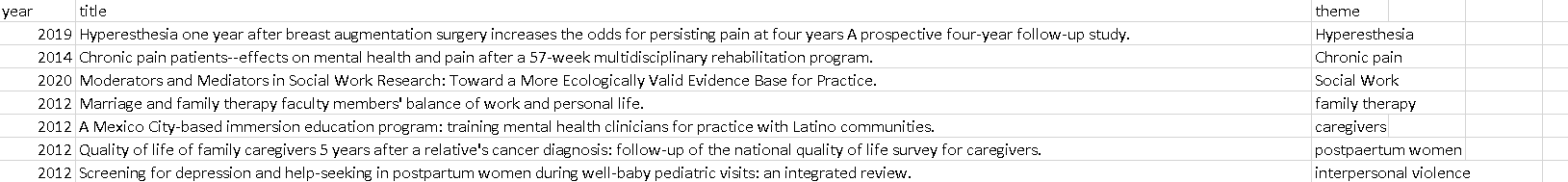


Similarly for coronavirus keyword, we can see that 2020m has highest variation in sentiments from strongly negative (-6) to positive(6)



Theme extraction:

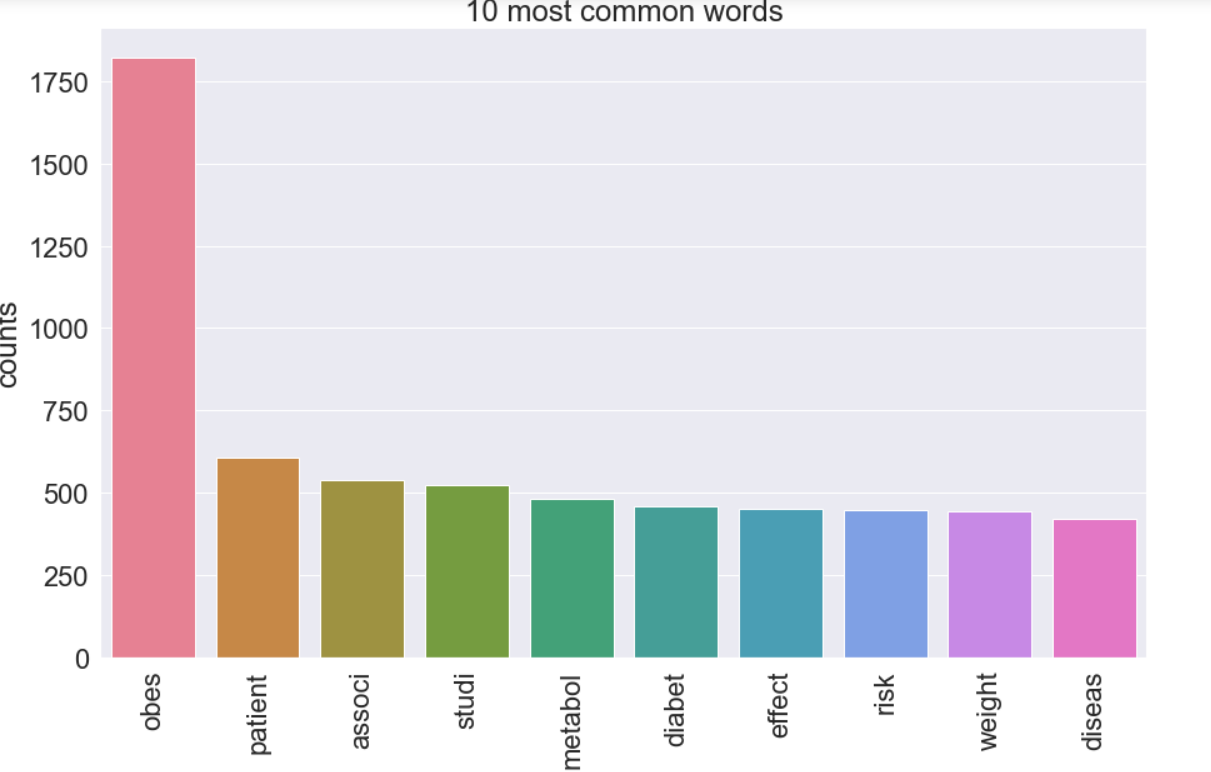
Theme extraction is the process of identifying the main idea of the content which can sufficiently describe the topic of the content. In this study, theme extraction has been performed on the medical articles manually



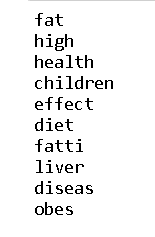
Clustering and association rule mining

The main goal of clustering is identifying hidden structures and patterns in the data and association rule mining is used to identify items which are similar to each and form rules based on the relationship between the item sets.

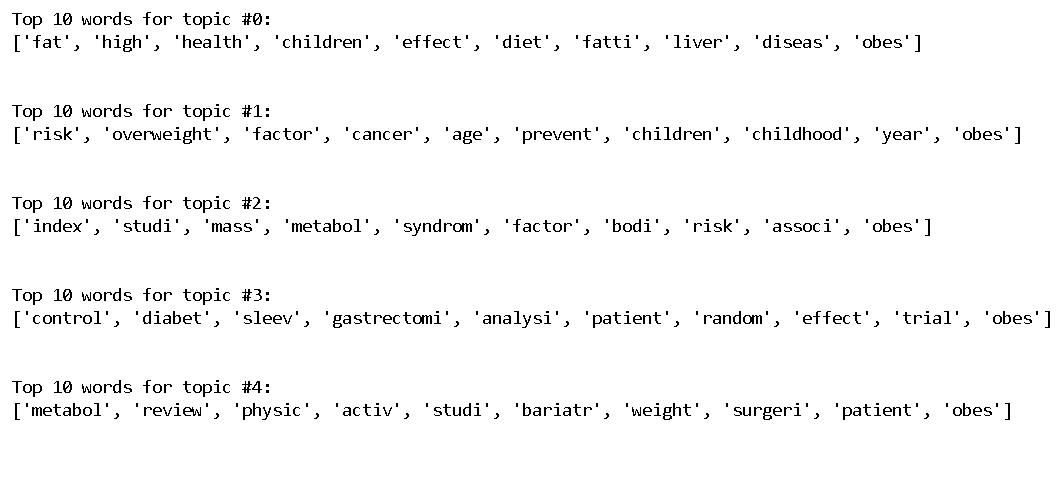
In this study ,the algorithm used is Latent Dirichlet Allocation (LDA) which is a probabilistic model that assumes that each topic is a mixture of set of words with underlying probabilities.After running LDA on the data ,the top 10 topics identifies are as follow:



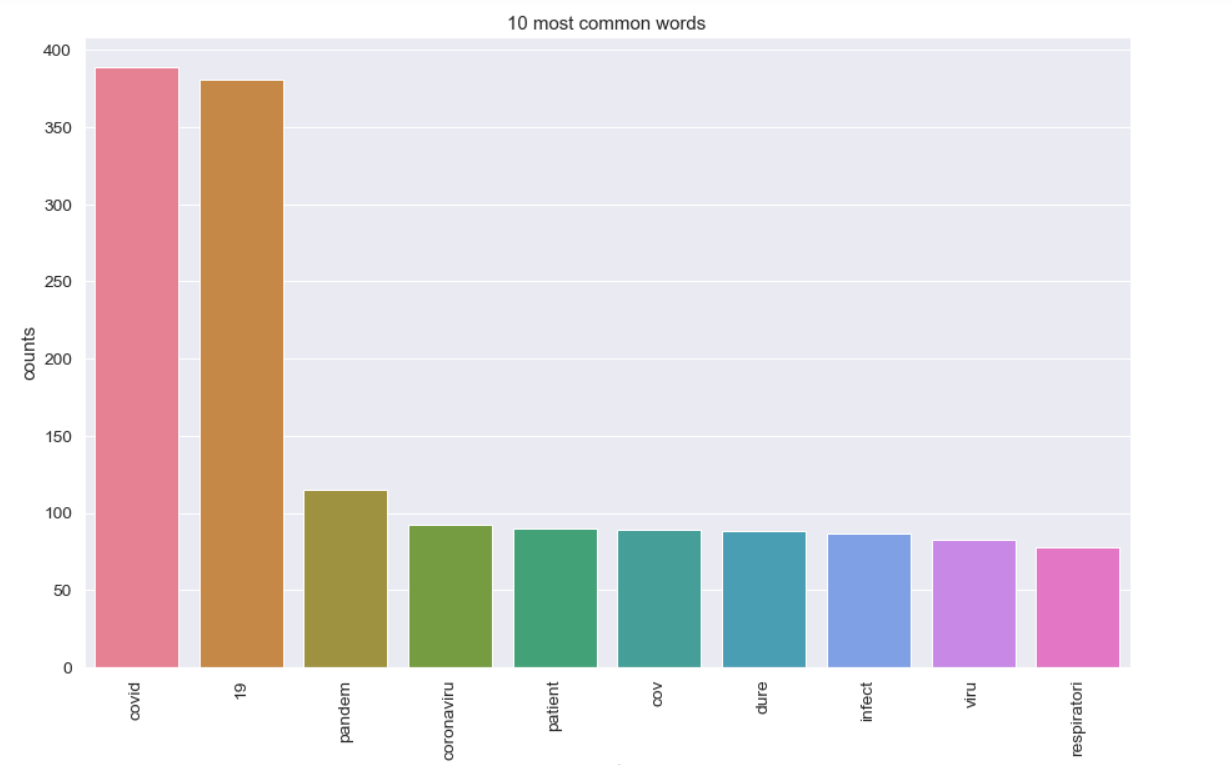
The word obese having highest number of topics, some of the words which are closest to obese are



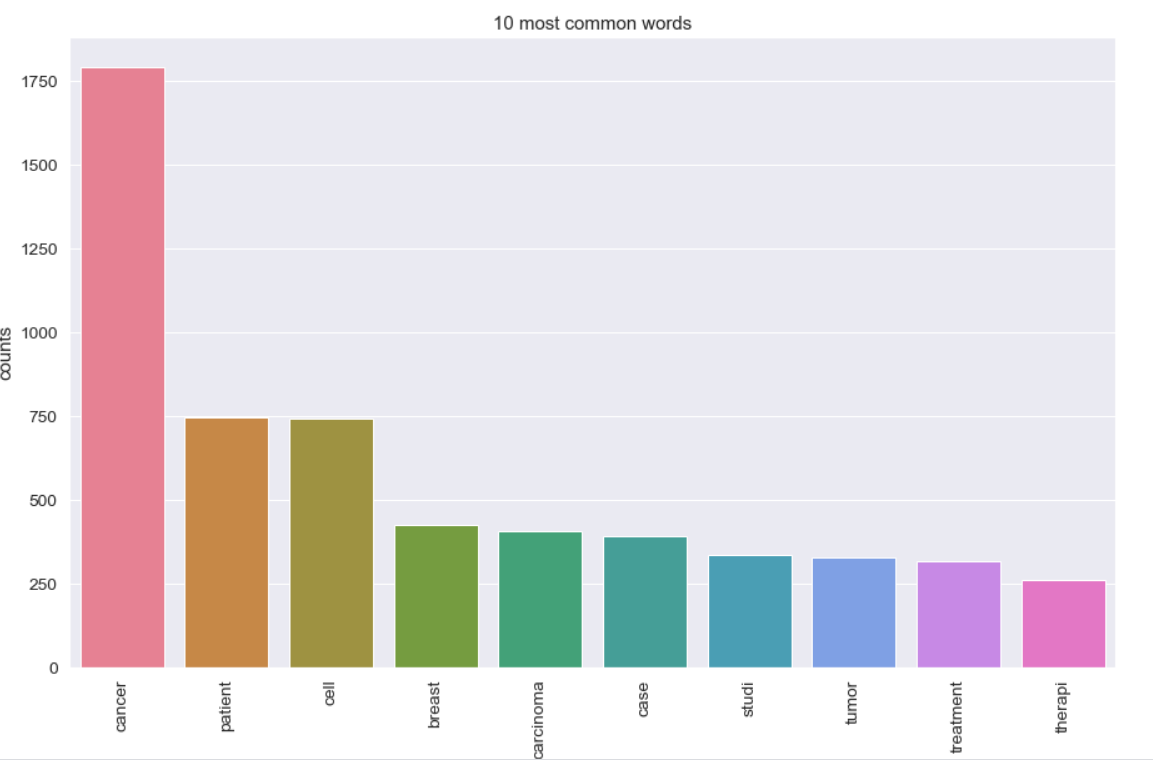
The 5 other similar topics to obesity have collection of following words



Some of the most common words similar to covid-19 are represented in the following visualization:



The most common topics related to cancer are shown in the plots below:



Conclusion:

The dataset has highest number of records for 2020 slightly bending all the calculations towards 2020.The sentiment scores are mostly neutral with slight variation to slight positive and negative articles.

References:

<https://scikit-learn.org/stable/modules/feature_selection.html>

<https://medium.com/@himanshu_23732/sentiment-analysis-with-afinn-lexicon-930533dfe75b>

<https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0>

<https://cran.r-project.org/web/packages/easyPubMed/vignettes/getting_started_with_easyPubMed.html>