Data analysis

Cs688 Term Project

Xiaoyang Wang
BUid: U57640099

1. Knowledge extracted from trajectory

1.1 Feature engineering

First, i use R library 'RISmed' to download 3000 articles from 2010 to 2020, which means I got 11 separate csv file.

So the first step of feature engineering is to combine 11 files and generate a dataframe. Then I drop the None values in dataframe. In case of the following process, I changed all letters into lower letters and removed all stop words in English.

In order to extract the trajectory in 10 years, I count the frequencies of five keywords and try to find some patterns. What I did is : I keep records of every keywords when they appears.

1.2 Keyword extraction

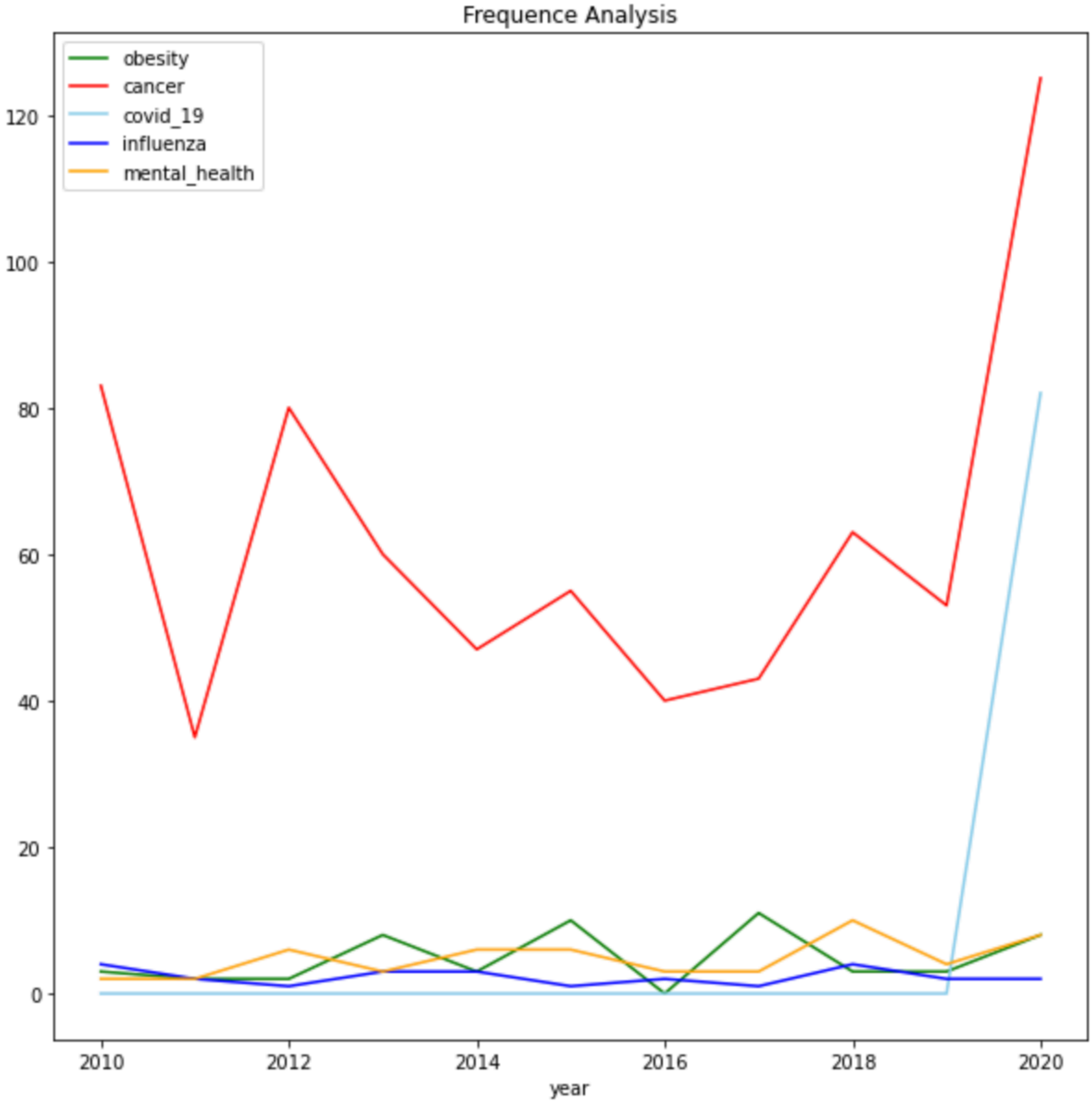
Here is the frequency table of keywords:

year	obesity	cancer	covid-19	influenza	Mental health
2010	3	83	0	4	2
2011	2	35	0	2	2
2012	2	80	0	1	6
2013	8	60	0	3	3
2014	3	47	0	3	6
2015	10	55	0	1	6
2016	0	40	0	2	3
2017	11	43	0	1	3
2018	3	63	0	4	10

2019	3	53	0	2	4
2020	8	125	82	2	8

Table1: keyword frequency table

Here is a visualization of the trajectory of the frequencies of keywords:



1.3 Statistical test

For statistical test , I choose two methods to evaluate data, which are T-test and Kruka-test. Here is the T-test result table of statistic value and p-value:

Statistic-value/p-value	**obesity**	**cancer**	**Covid-19**	**influenza**	**Mental health**
obesity		-7.29/4.72	-0.35/0.73	2.18/0.04	0.0/1.0

cancer	-7.29/4.72		5.08/5.76	7.69/2.14	7.33/4.37
Covid-19	-0.35/0.73	5.08/5.76		0.69/0.5	0.35/0.73
influenza	2.18/0.04	7.69/2.14	0.69/0.5		-3/0.007
Mental health	0.0/1.0	7.33/4.37	0.35/0.73	-3/0.007	

Here is the Kruka-test result table of statistic value and p-value:

Statistic-value/p-value	**obesity**	**cancer**	**Covid-19**	**influenza**	**Mental health**
obesity		15.89/6.71	9.82/0.0017	2.92/0.086	0.072/0.788
cancer	15.89/6.71		12.61/0.00038	15.92/6.5e-05	15.86/6.8e-05
Covid-19	9.82/0.0017	12.61/0.00038		11.76/0.0006	11.71/0.0006
influenza	2.92/0.086	15.92/6.5e-05	11.76/0.0006		6.64/0.01
Mental health	0.072/0.788	15.86/6.8e-05	11.71/0.0006	6.64/0.01	

In T-test, the p-values of obesity- influenza and mental_health -obesity are less than 0.05, which mean we can reject the hypothesis and there are not correlations.

In Kruka-test, p-values shows that the hypothesis can be only accept in obesity-cancer and obesity-mental-health.

1.4 Analysis of keyword trajectory

1. The most obvious feature is that cancer always has the highest frequency above the rest four keywords. The first reason is that cancer always the hottest area in medical research, besides, there are may sub-area in cancer researches, so it has the highest frequencies.

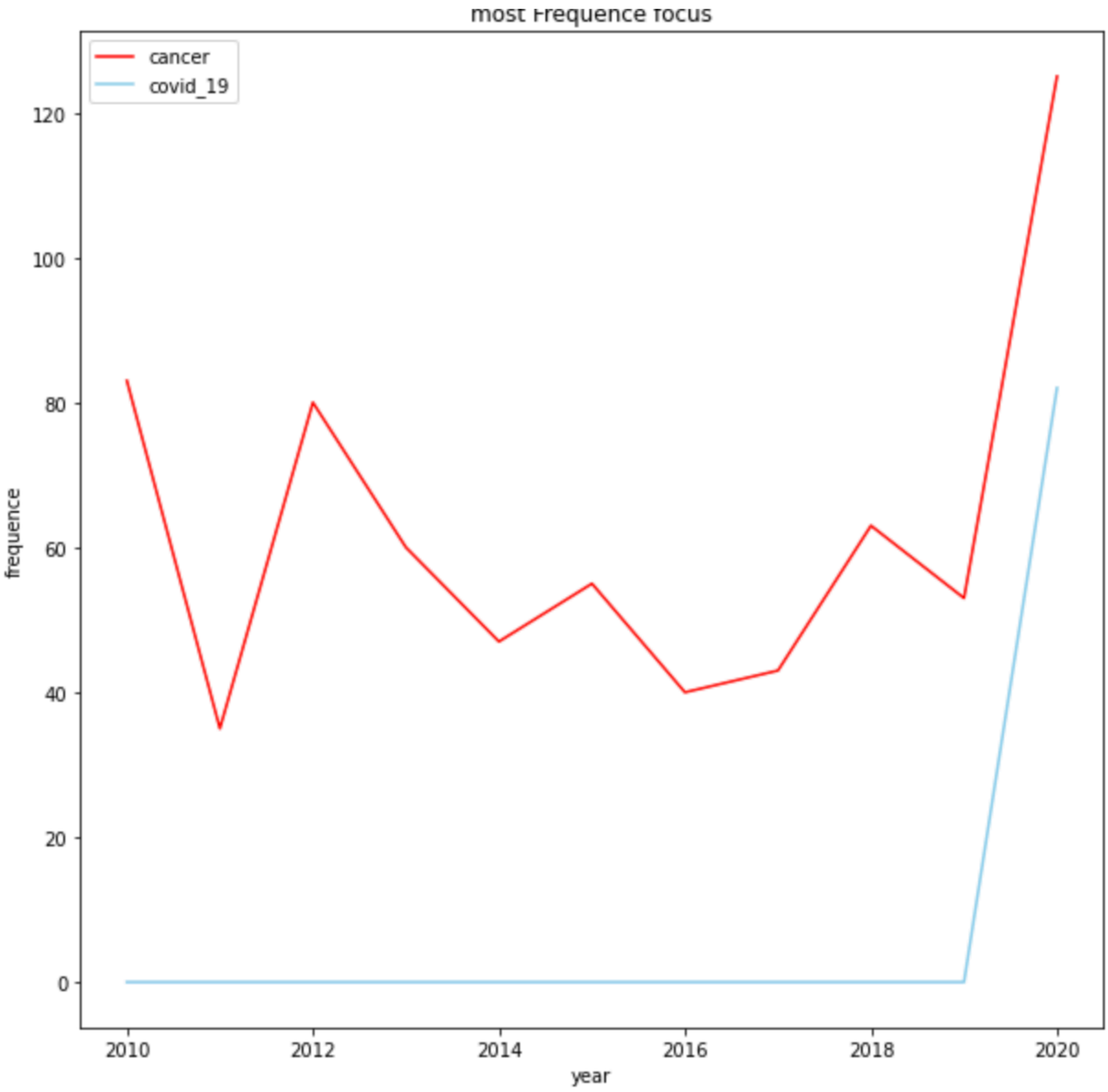
2. ‘covid-19’ didn’t appear until 2020.In 2020, the occurrences raised and there is a huge gap between that of covid-19 and other three.

3. Obesity, influenza, mental health is stable in 10 years, however, there are upward trend recently, which may have a correlation with coivd-19 and self-quarantine.

2. Focus of academia on keyword and changes

Based on the analysis, the most frequent focus of academia is cancer. Besides, in 2020, there is another hot keyword, which is covid-19, that has many focuses of academia.

Here is the visualization specially for cancer and covid-19:



3. Sentiment of keyword and trajectory

3.1 Sentiment of keywords

In this part , I choose Afinn() library to get the score of sentiment of word/sentence. Frist I check the score of particular keywords: obesity, cancer, covid-19, influenza and mental_health, here is the result :

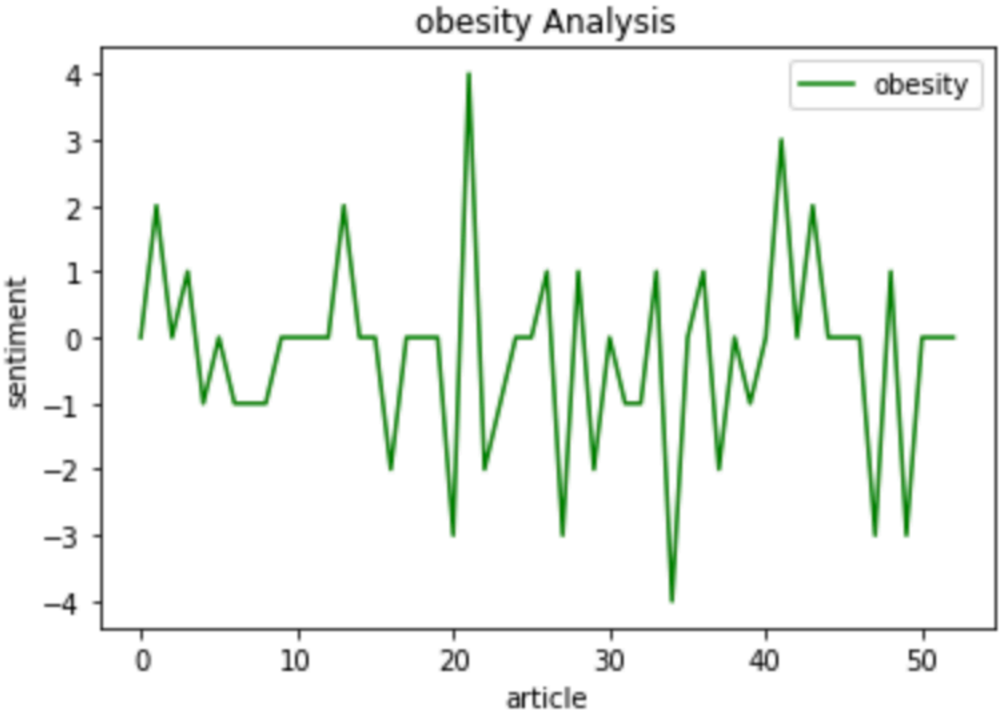
keyword	**Sentiment score**
obesity	0.0
cancer	-1.0
covid-19	0.0
influenza	0.0
mental_health	0.0

Then I implement two different methods :NRC, Bing:

	afinn	nrc	bing
obesity	0	-1	0
cancer	-1	-1	-1
Covid-19	0	0	0
influenza	0	-1	0
Mental_health	0	0	0

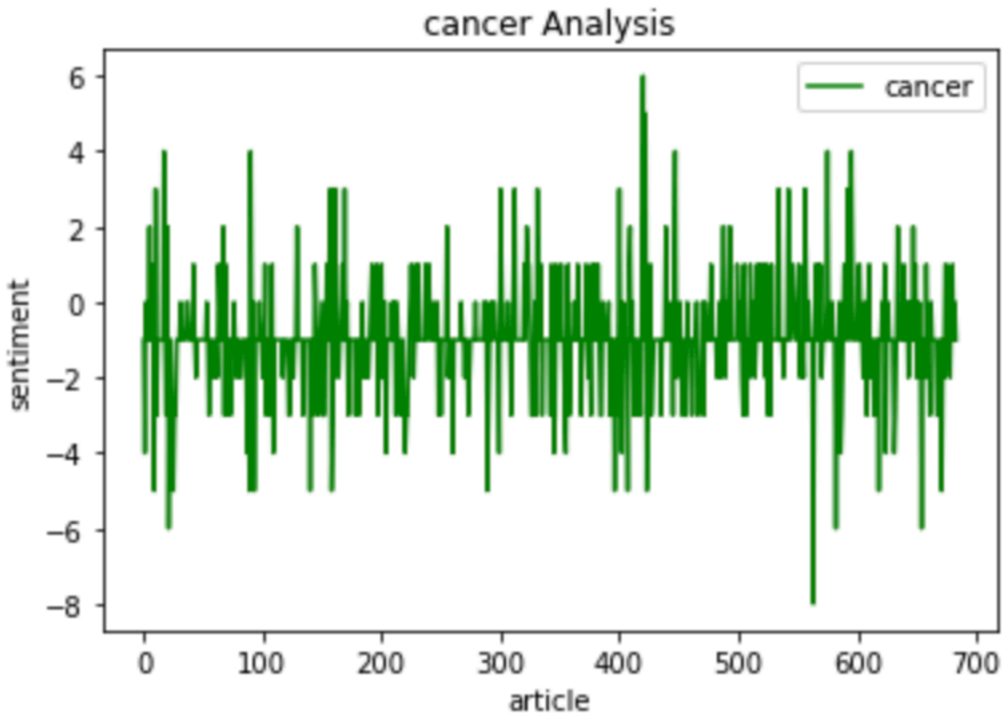
So basically, I build up a rule: if sentiment score is greater than 0, it is a positive sentiment, if sentiment score is less than 0, it is a negative sentiment, if it is 0, it is neutral sentiment. Based on this rules, only cancer has negative sentiment, the rest four keywords is neutral.

For obesity, this graph illustrate the trajectory of sentiment :



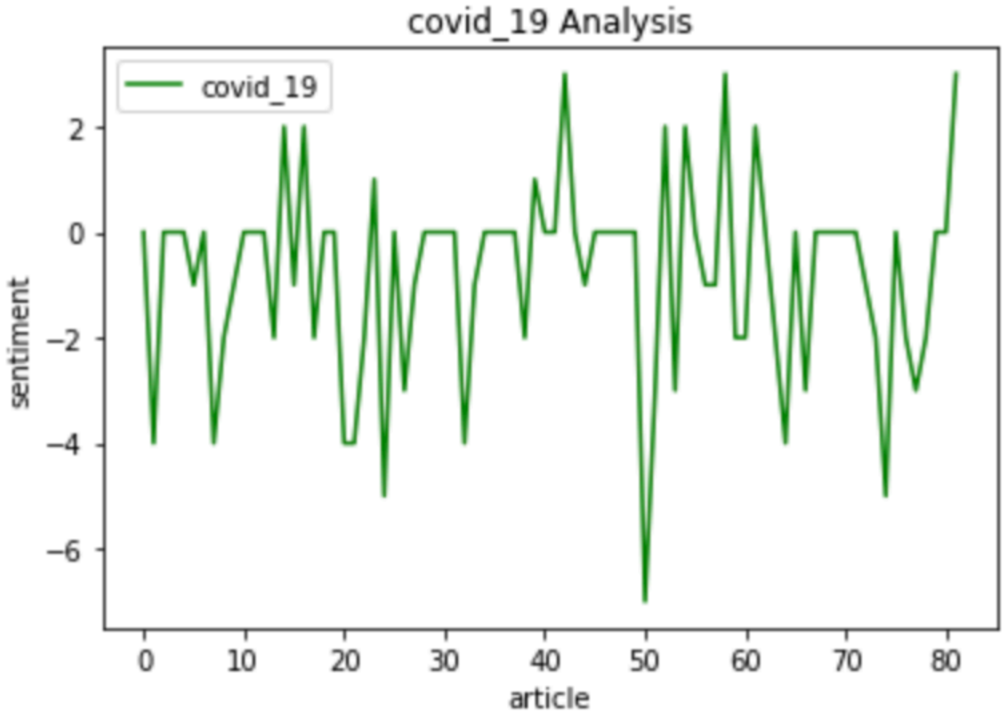
So the sentiment of articles regarding obesity had balanced distribution over years.

For cancer, this graph illustrate the trajectory of sentiment :



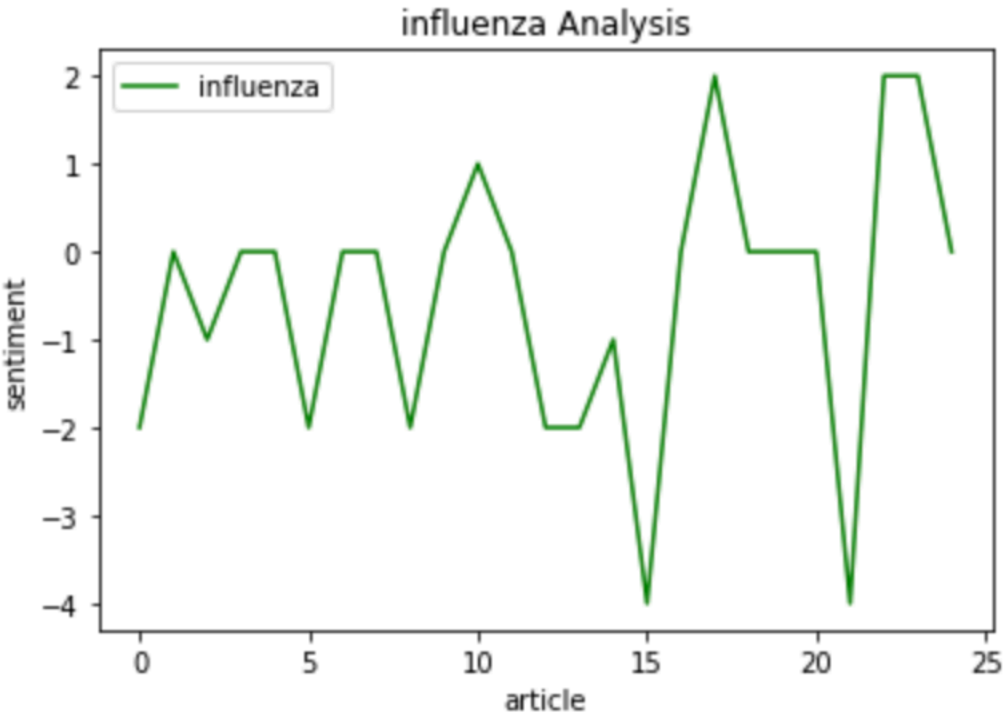
So the sentiment of articles regarding cancer had balanced distribution over years.

For covid-19, this graph illustrate the trajectory of sentiment :



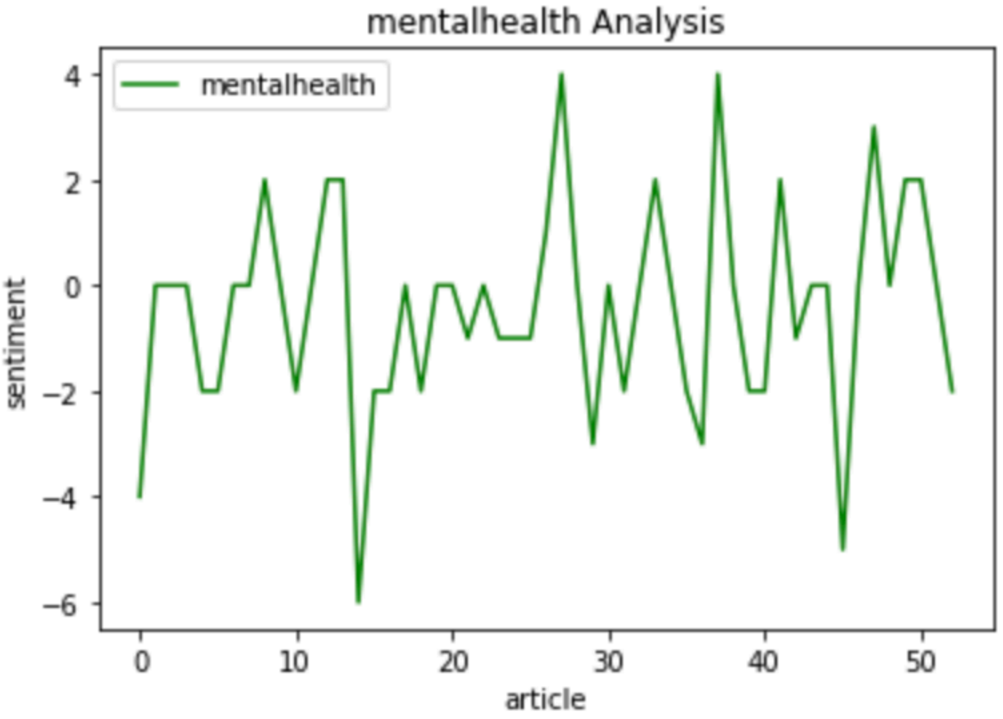
So the sentiment of articles regarding covid-19 are mostly negative.

For influenza, this graph illustrate the trajectory of sentiment :



So the sentiment of articles regarding influenza are mostly negative over first few years, and it had more positive articles recent years.

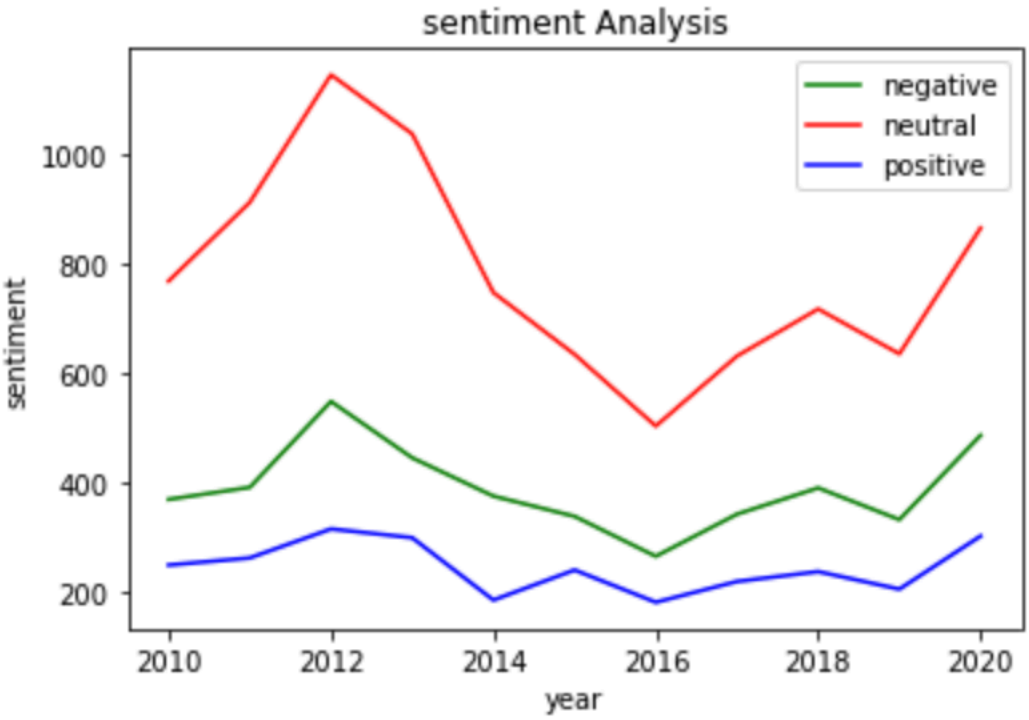
For mental_health, this graph illustrate the trajectory of sentiment :



So the sentiment of articles regarding mental health had balanced distribution over years.

3.2 Sentiment of all articles

I also focus on finding out the trajectory of sentiment of all articles in 10 years and get the result :

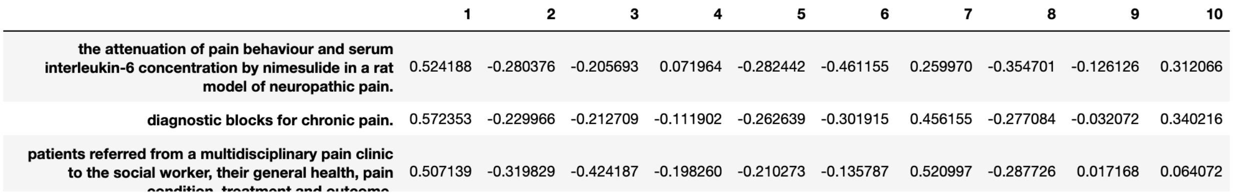


The numbers shows the same trends of three different sentiments: they raised from 2010 and reach their peeks in 2012. After falling down from 2012 to 2018, they raised again from 2019 to now.

4. Clustering and association rule

4.1 Lsa clustering for articles

Lsa model needs document term matrix. So firstly, I transfer all articles into a document term matrix, using CountVectorizer(), and then use lsa model. I set the clustering topics as 10, then got the lsa matrix:

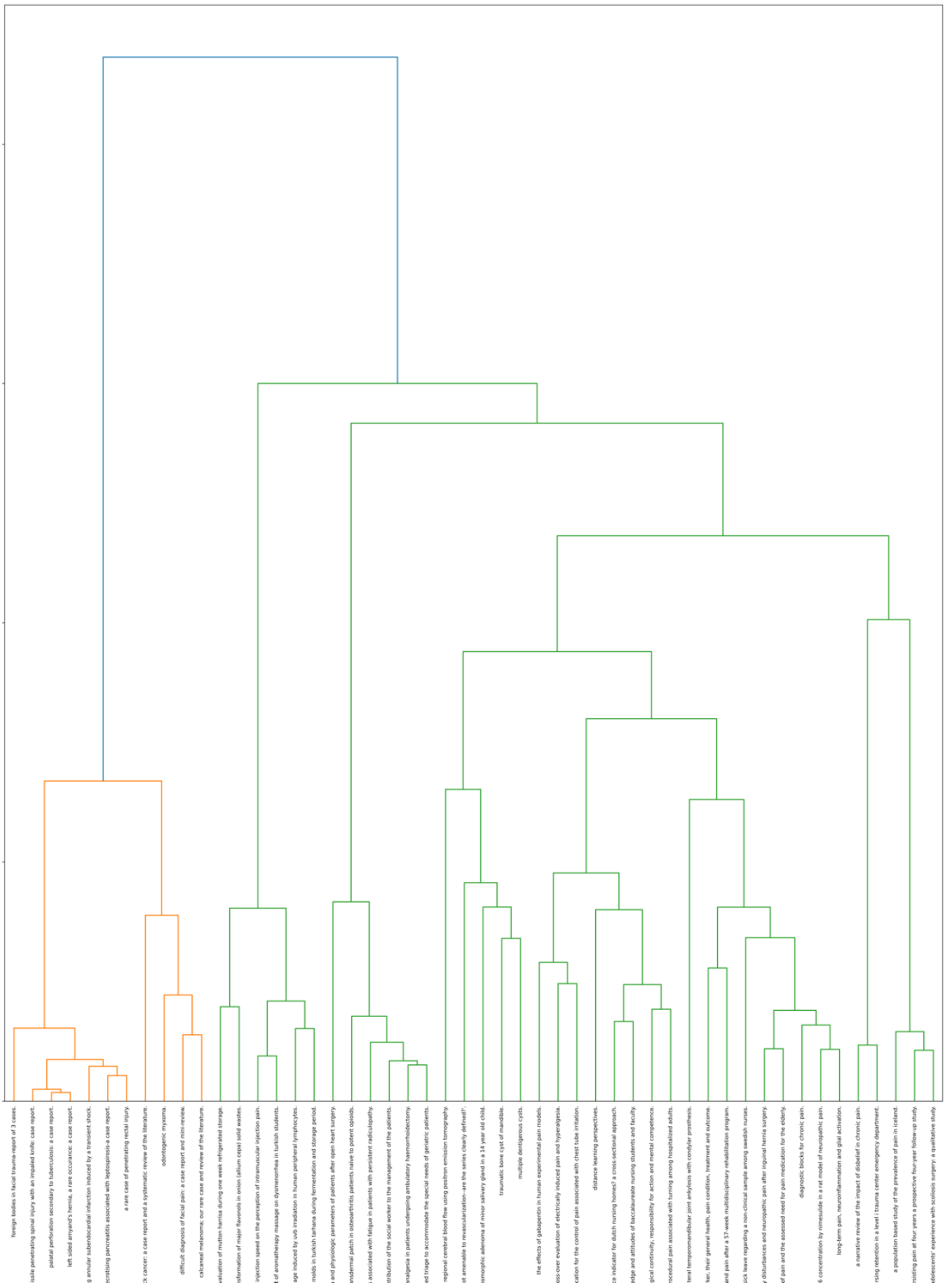


	1	2	3	4	5	6	7	8	9	10
the attenuation of pain behaviour and serum interleukin-6 concentration by nimesulide in a rat model of neuropathic pain.	0.524188	-0.280376	-0.205693	0.071964	-0.282442	-0.461155	0.259970	-0.354701	-0.126126	0.312066
diagnostic blocks for chronic pain.	0.572353	-0.229966	-0.212709	-0.111902	-0.262639	-0.301915	0.456155	-0.277084	-0.032072	0.340216
patients referred from a multidisciplinary pain clinic to the social worker, their general health, pain condition, treatment and outcome	0.507139	-0.319829	-0.424187	-0.198260	-0.210273	-0.135787	0.520997	-0.287726	0.017168	0.064072

Then I print out the topics, which are also list of words:

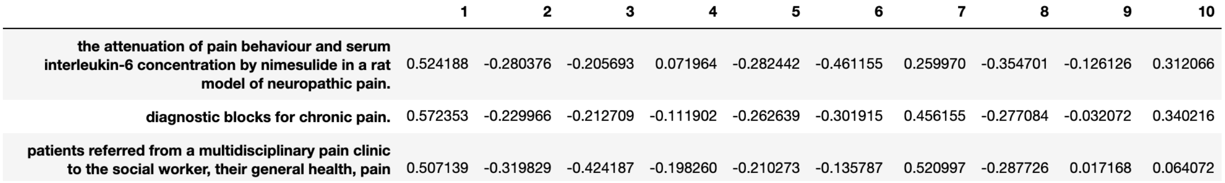
```
Topic #0:
study patients case report review cancer treatment literature cell analysis clinical disease
 using based management effect pain care evaluation carcinoma
Topic #1:
case report review literature rare syndrome presenting systematic series patient presentat
ion cases unusual cyst carcinoma tumor giant congenital old diagnosis
Topic #2:
study case prospective comparative cross pilot sectional retrospective population report
based year control observational children vitro cohort mandibular physical health
Topic #3:
cancer using analysis effect review treatment cell based cells breast systematic human cl
inical health induced evaluation activity meta quality quot
Topic #4:
cancer cell case breast carcinoma report cells study lung rare prostate therapy small tum
or non expression metastatic colorectal stem squamous
Topic #5:
review cancer study literature systematic analysis breast meta prostate cases retrospectiv
e factors colorectal lung prospective metastatic risk cells care tertiary
Topic #6:
treatment clinical surgical patient pain chronic outcomes risk care factors disease manag
ement trial health surgery fractures using use case following
Topic #7:
using analysis cell evaluation based case quot patients data method response high huma
n meta optimization detection risk surface approach gene
Topic #8:
cell treatment carcinoma disease induced stem squamous human renal acute clinical revi
ew cells small expression non associated study tumor following
Topic #9:
quot health care disease management risk factors based associated effects patient cell us
e children clinical pain prevalence tertiary induced role
```

To provide a clear show of the result of clustering, I take 50 instance as example to visualization:



4.2 Lda clustering for articles

Lda model also needs document term matrix. then use lda model. I set the clustering topics as 10, then got the lda matrix:

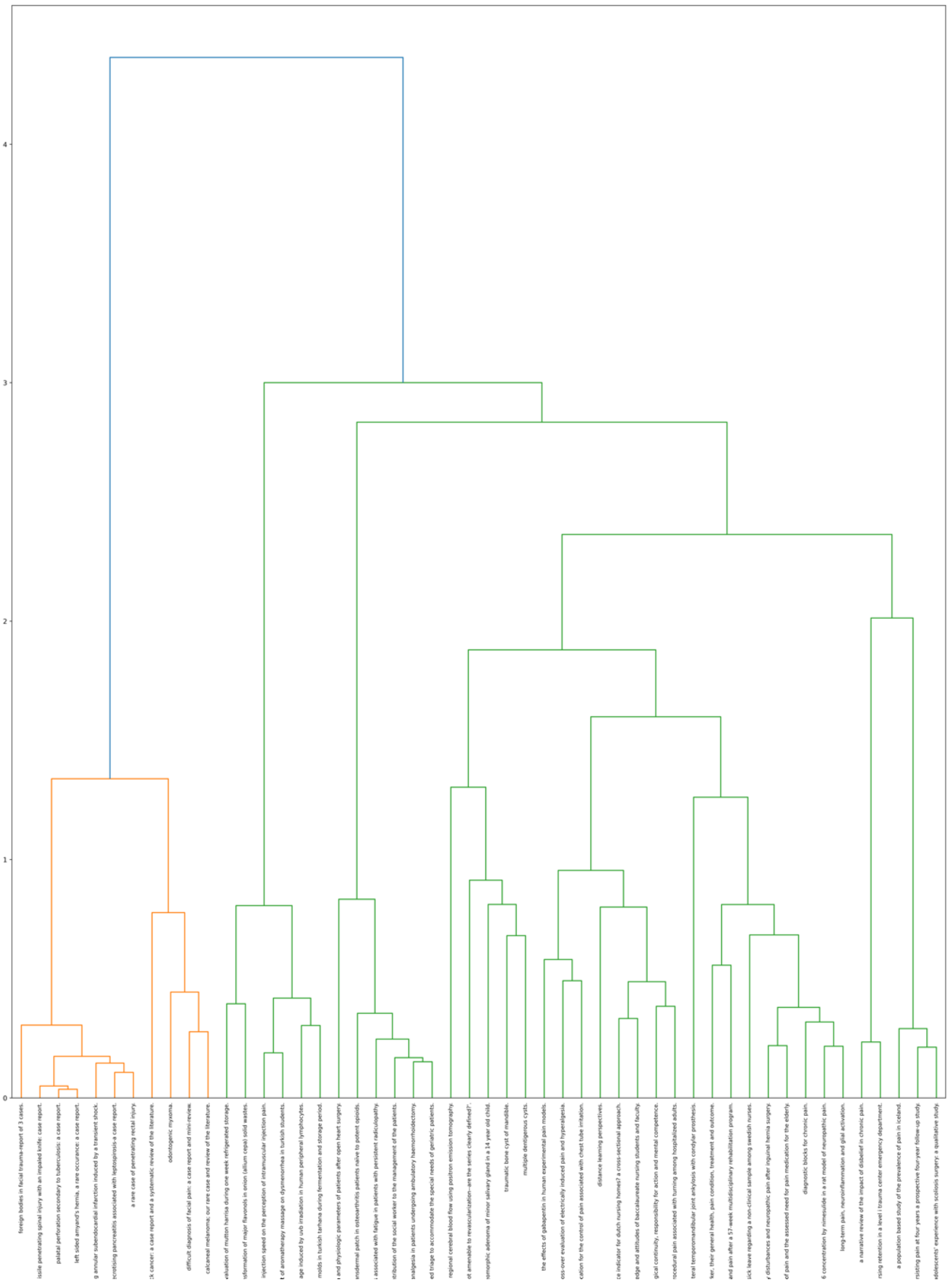


	1	2	3	4	5	6	7	8	9	10
the attenuation of pain behaviour and serum interleukin-6 concentration by nimesulide in a rat model of neuropathic pain.	0.524188	-0.280376	-0.205693	0.071964	-0.282442	-0.461155	0.259970	-0.354701	-0.126126	0.312066
diagnostic blocks for chronic pain.	0.572353	-0.229966	-0.212709	-0.111902	-0.262639	-0.301915	0.456155	-0.277084	-0.032072	0.340216
patients referred from a multidisciplinary pain clinic to the social worker, their general health, pain	0.507139	-0.319829	-0.424187	-0.198260	-0.210273	-0.135787	0.520997	-0.287726	0.017168	0.064072

Then I print out the topics, which are also list of words:

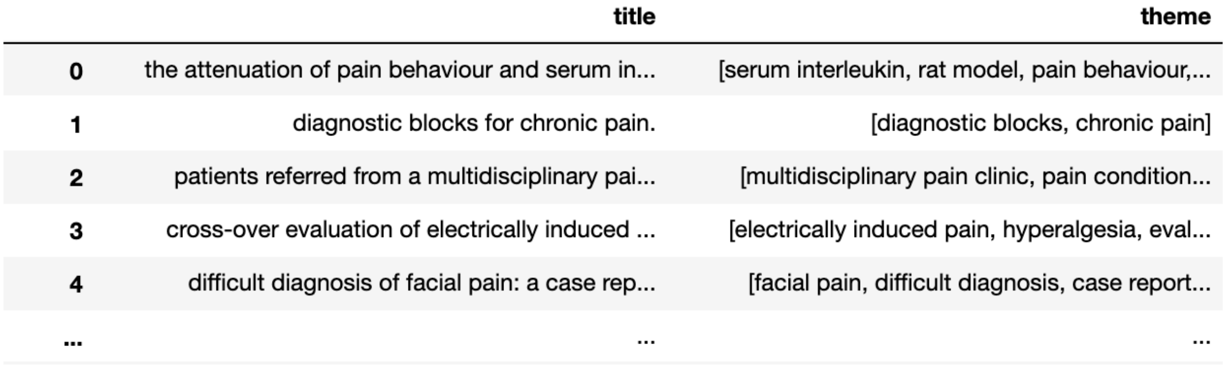
```
Topic #0:
quot high development using different drug functional model molecular vitro characteri
zation indian detection status resistance medicine new effects synthesis india
Topic #1:
patients health study treatment role based pain effects prevalence injury population gene
 changes long term liver chronic trial influence women
Topic #2:
case review report cancer cell rare management literature type surgical treatment system
atic carcinoma patients acute method breast artery new tumor
Topic #3:
impact properties levels free cardiac nasal delivery fractures conditions mediated recons
truction low stage flap screening mouse ventricular environmental biomarkers fistula
Topic #4:
syndrome effect human quality induced protein bone acid physical evaluation stress rats
 activity diabetes using dna research pressure storage treatment
Topic #5:
patient therapy cancer assessment study approach following patients infection life comp
arison early related factors blood potential cells cell lung time
Topic #6:
clinical non does water mice learning experimental effect cord people cells effects anti e
xercise cyst tool media models improve extract
Topic #7:
analysis risk children primary expression activity brain novel meta food disorders factor
 use induced endoscopic surface health evaluation age growth
Topic #8:
study patients outcomes control experience comparative associated multiple year renal p
rospective characteristics single surgery 19 retrospective adults covid identification eval
uation
Topic #9:
disease care based response association antioxidant results severe studies kidney stabilit
y cervical systems resistant ovarian brazil fat composition public modified
```

To provide a clear show of the result of clustering, I take 50 instance as example to visualization:



4.3 Theme extraction

I choose rake_nlp to automatically extract the theme of every articles.



	title	theme
0	the attenuation of pain behaviour and serum in...	[serum interleukin, rat model, pain behaviour,...
1	diagnostic blocks for chronic pain.	[diagnostic blocks, chronic pain]
2	patients referred from a multidisciplinary pai...	[multidisciplinary pain clinic, pain condition...
3	cross-over evaluation of electrically induced ...	[electrically induced pain, hyperalgesia, eval...
4	difficult diagnosis of facial pain: a case rep...	[facial pain, difficult diagnosis, case report...
...	...	...

4.4 Association rule

In this part , I use the titles that without stopwords. Firstly, I divide all articles into different groups based on years and implement the apriori method.So for each year, the rule result is (for example):

```
year:2010
[{neck} -> {head}, {head} -> {neck}, {review} -> {literature.}, {literature.} -> {revie
w}, {report.} -> {case}, {case} -> {report.}, {fine-needle} -> {aspiration}, {aspiration}
 -> {fine-needle}]

year:2011
[{report} -> {case}, {case} -> {report}, {report.} -> {case}, {case} -> {report.}, {qualit
y} -> {effect}, {effect} -> {quality}, {review} -> {literature.}, {literature.} -> {revie
w}, {literature.} -> {case}, {case} -> {literature.}, {review} -> {case}, {case} -> {revie
w}, {literature., review} -> {case}, {case, review} -> {literature.}, {case, literature.} ->
{review}, {review} -> {case, literature.}, {literature.} -> {case, review}, {case} -> {lite
rature., review}]

year:2012
[{report.} -> {case}, {case} -> {report.}, {report} -> {case}, {case} -> {report}, {rare}
-> {case}, {case} -> {rare}, {review} -> {report}, {report} -> {review}, {literature.} ->
 {case}, {case} -> {literature.}, {review} -> {case}, {case} -> {review}, {report} -> {lit
erature.}, {literature.} -> {report}, {review} -> {literature.}, {literature.} -> {review},
{literature., report} -> {case}, {case, report} -> {literature.}, {case, literature.} -> {repo
rt}, {report} -> {case, literature.}, {literature.} -> {case, report}, {case} -> {literature., r
eport}, {literature., review} -> {case}, {case, review} -> {literature.}, {case, literature.}
 -> {review}, {review} -> {case, literature.}, {literature.} -> {case, review}, {case} ->
{literature., review}, {report, review} -> {case}, {case, review} -> {report}, {case, repo
rt} -> {review}, {review} -> {case, report}, {report} -> {case, review}, {case} -> {repo
rt, review}, {report, review} -> {literature.}, {literature., review} -> {report}, {literatur
e., report} -> {review}, {review} -> {literature., report}, {report} -> {literature., revie
w}, {literature.} -> {report, review}, {literature., report, review} -> {case}, {case, repor
t, review} -> {literature.}, {case, literature., review} -> {report}, {case, literature., repor
t} -> {review}, {report, review} -> {case, literature.}, {literature., review} -> {case, rep
ort}, {literature., report} -> {case, review}, {case, review} -> {literature., report}, {cas
e, report} -> {literature., review}, {case, literature.} -> {report, review}, {review} -> {c
ase, literature., report}, {report} -> {case, literature., review}, {literature.} -> {case, rep
ort, review}, {case} -> {literature., report, review}]
```

Itemset result :

```
{1: {('clinical',): 69, ('changes',): 28, ('treatment',): 67, ('study',): 77, ('chronic',): 27,
{1: {('review',): 30, ('india.',): 11, ('care',): 14, ('experience',): 14,
{1: {('clinical',): 46, ('patient',): 24, ('tumor',): 16, ('following',): 18,
```

I found in every year, there are rules like: {review} -> {literature.}, {report} -> {case},{case, review}, and ('clinical') ('review',) are the most frequent words in itemsets.

5. Own findings

While analysing the changes of frequencies of keywords: mental health and covid-19, I found that there should be some correlations, like covid-19 forced people self-quarantine in home, which may cause some mental problem if people do not go outside for long time.

Cancer study is always the hottest field in medical research. For instance , in my lsa clustering result, there are four topics has the key word cancer.

Clinical research is very popular in pubmed , as it always appears in my association rule mining and always along with patient.