Homework 2 Report

Kholishotul Amaliah (10971108)

TEXT MINING (EE100098)
VIRTUAL EXCHANGE PROGRAM
ASIA UNIVERSITY
FALL SEMESTER

METHOD

Specification of hardware used:

- CPU : Intel Core i7-7700HQ (4 Core, 8 Thread, 2.8 GHz)
- RAM : 8GB

Specification of software used:

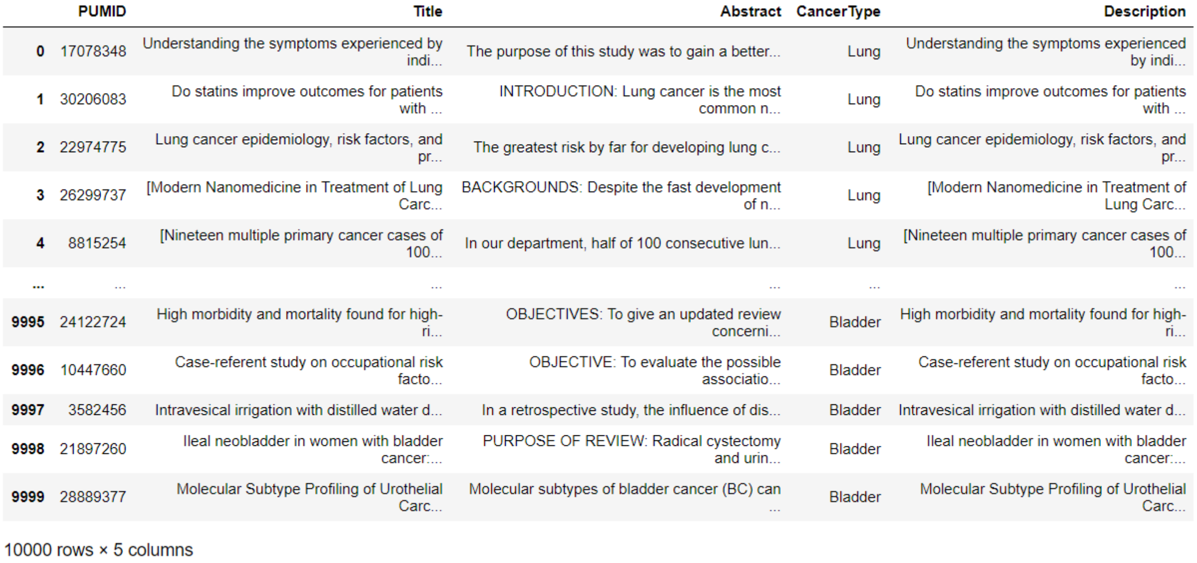
- Python version : 3.7.4

Steps :

1. Data Needed

The initial data is 100.000 information of top 10 cancer types from PubMed. The data has information about PUMID, Title, Abstract, and CancerType. To have faster computation, I choose randomly 1000 data from each cancer type. So now I have 10000 rows of data.

The content (the text to be processed) is concatenated of Title and Abstract become Description.



	PUMID	Title	Abstract	CancerType	Description
0	17078348	Understanding the symptoms experienced by indi...	The purpose of this study was to gain a better...	Lung	Understanding the symptoms experienced by indi...
1	30206083	Do statins improve outcomes for patients with ...	INTRODUCTION: Lung cancer is the most common n...	Lung	Do statins improve outcomes for patients with ...
2	22974775	Lung cancer epidemiology, risk factors, and pr...	The greatest risk by far for developing lung c...	Lung	Lung cancer epidemiology, risk factors, and pr...
3	26299737	[Modern Nanomedicine in Treatment of Lung Carc...	BACKGROUNDS: Despite the fast development of n...	Lung	[Modern Nanomedicine in Treatment of Lung Carc...
4	8815254	[Nineteen multiple primary cancer cases of 100...	In our department, half of 100 consecutive lun...	Lung	[Nineteen multiple primary cancer cases of 100...
...	...	...	...	...	...
9995	24122724	High morbidity and mortality found for high-ri...	OBJECTIVES: To give an updated review concerni...	Bladder	High morbidity and mortality found for high-ri...
9996	10447660	Case-referent study on occupational risk facto...	OBJECTIVE: To evaluate the possible associatio...	Bladder	Case-referent study on occupational risk facto...
9997	3582456	Intravesical irrigation with distilled water d...	In a retrospective study, the influence of dis...	Bladder	Intravesical irrigation with distilled water d...
9998	21897260	Ileal neobladder in women with bladder cancer:...	PURPOSE OF REVIEW: Radical cystectomy and urin...	Bladder	Ileal neobladder in women with bladder cancer:...
9999	28889377	Molecular Subtype Profiling of Urothelial Carc...	Molecular subtypes of bladder cancer (BC) can ...	Bladder	Molecular Subtype Profiling of Urothelial Carc...

10000 rows × 5 columns

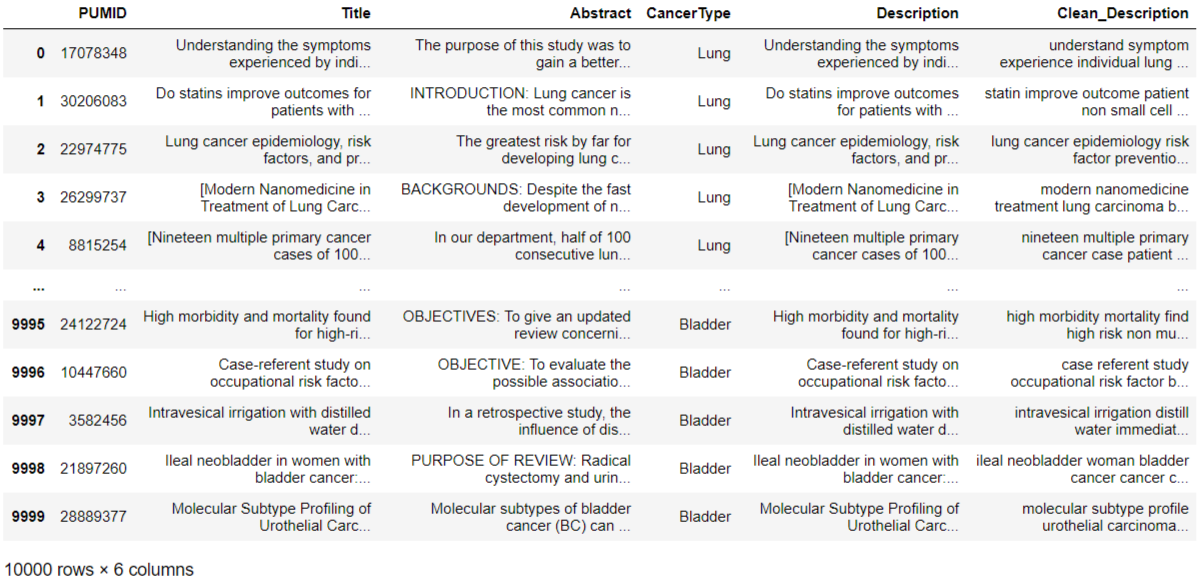
2. Text preprocessing

Text wrangling (also called preprocessing or normalization) is a process that consists of a series of steps to wrangle, clean, and standardize textual data into a form that could be consumed by other NLP and intelligent systems powered by machine learning and deep learning. We use accented char removal, text lower case, text lemmatization, special character removal, digits removal, and stopwords removal methods.

For implementation, first I need to lower case and remove special characters by using **regex** from the text. Then, for the tokenization I use the **nltk** library. After that, I need to remove the stopwords from the text with the help of nltk stopwords. But I keep the negation (no and not) for the bi-grams. And last, I need to have the basic form of word, so I lemmatize the text with the help of **spacy** library.

3. Dataset

So now, I have clean description for text mining with 10000 totals of data. It is saved under the name (Cleaned)pubmed-CancerType_Top1-10-set_10000-data.xlsx in the Src folder.



	PUMID	Title	Abstract	CancerType	Description	Clean_Description
0	17078348	Understanding the symptoms experienced by indi...	The purpose of this study was to gain a better...	Lung	Understanding the symptoms experienced by indi...	understand symptom experience individual lung ...
1	30206083	Do statins improve outcomes for patients with ...	INTRODUCTION: Lung cancer is the most common n...	Lung	Do statins improve outcomes for patients with ...	statin improve outcome patient non small cell ...
2	22974775	Lung cancer epidemiology, risk factors, and pr...	The greatest risk by far for developing lung c...	Lung	Lung cancer epidemiology, risk factors, and pr...	lung cancer epidemiology risk factor preventio...
3	26299737	[Modern Nanomedicine in Treatment of Lung Carc...	BACKGROUNDS: Despite the fast development of n...	Lung	[Modern Nanomedicine in Treatment of Lung Carc...	modern nanomedicine treatment lung carcinoma b...
4	8815254	[Nineteen multiple primary cancer cases of 100...	In our department, half of 100 consecutive lun...	Lung	[Nineteen multiple primary cancer cases of 100...	nineteen multiple primary cancer case patient ...
...	...	...	...	...	...	...
9995	24122724	High morbidity and mortality found for high-ri...	OBJECTIVES: To give an updated review concerni...	Bladder	High morbidity and mortality found for high-ri...	high morbidity mortality find high risk non mu...
9996	10447660	Case-referent study on occupational risk facto...	OBJECTIVE: To evaluate the possible associatio...	Bladder	Case-referent study on occupational risk facto...	case referent study occupational risk factor b...
9997	3582456	Intravesical irrigation with distilled water d...	In a retrospective study, the influence of dis...	Bladder	Intravesical irrigation with distilled water d...	intravesical irrigation distill water immediat...
9998	21897260	Ileal neobladder in women with bladder cancer:...	PURPOSE OF REVIEW: Radical cystectomy and urin...	Bladder	Ileal neobladder in women with bladder cancer:...	ileal neobladder woman bladder cancer cancer c...
9999	28889377	Molecular Subtype Profiling of Urothelial Carc...	Molecular subtypes of bladder cancer (BC) can ...	Bladder	Molecular Subtype Profiling of Urothelial Carc...	molecular subtype profile urothelial carcinoma...

10000 rows × 6 columns

4. Text Mining

The text mining focus is text clustering. The text clustering methods used are K-Means, Affinity Propagation, and Ward Algorithm.

1) K-Means

The KMeans algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares (see below). This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has been used across a large range of application areas in many different fields.

2) Affinity Propagation

AffinityPropagation creates clusters by sending messages between pairs of samples until convergence. A dataset is then described using a small number of exemplars, which are identified as those most representative of other samples. The messages sent between pairs represent the suitability for one sample to be the exemplar of the other, which is updated in response to the values from other pairs. This updating happens iteratively until convergence, at which point the final exemplars are chosen, and hence the final clustering is given.

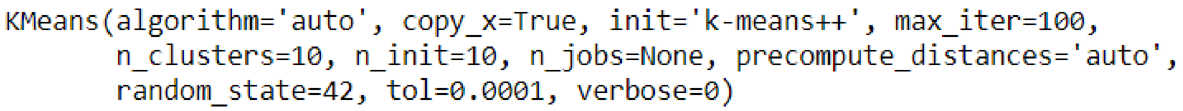
3) Ward Algorithm

Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample. Ward minimizes the sum of squared differences within all clusters. It is a variance-minimizing approach and in this sense is similar to the k-means objective function but tackled with an agglomerative hierarchical approach.

RESULT AND DISCUSSION

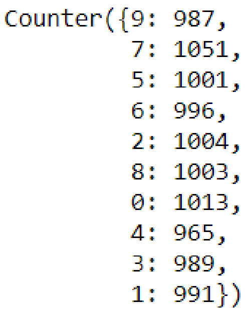
1. Text Clustering Using K-Means Cluster

First, I use TF-IDF for feature extraction. I limit the feature by ngram_range is unigram and bigram, minimum document frequency is 10, and maximum document frequency is 0.8. Then for the clustering, using K-Means with the number of clusters is 10, the maximum iteration is 100, and the n_init is 10.



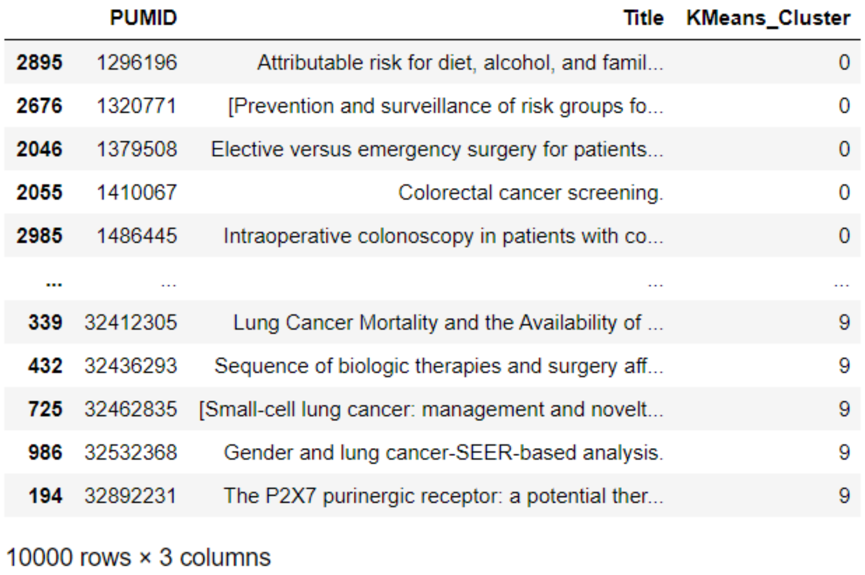
```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=100,
       n_clusters=10, n_init=10, n_jobs=None, precompute_distances='auto',
       random_state=42, tol=0.0001, verbose=0)
```

From the clustering, here is what I got.



```
Counter({9: 987,
         7: 1051,
         5: 1001,
         6: 996,
         2: 1004,
         8: 1003,
         0: 1013,
         4: 965,
         3: 989,
         1: 991})
```

There are 10 clusters, but each of the cluster has different total number of data. The most minimum data is cluster 4 with 965 data, and the most maximum data is cluster 7 with 1051 data. Here is the detail information about the clustering result.

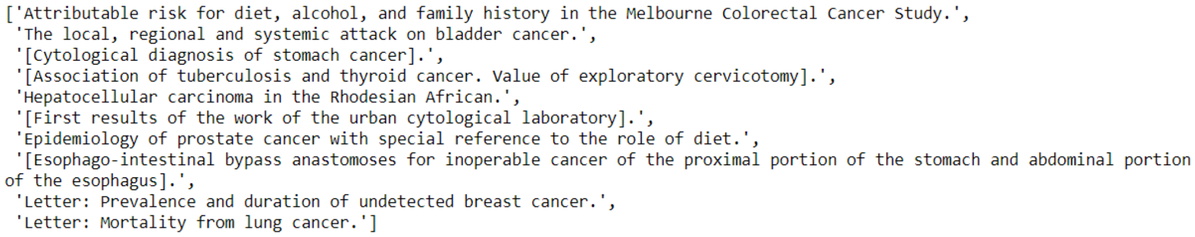


	PUMID	Title	KMeans_Cluster
2895	1296196	Attributable risk for diet, alcohol, and famil...	0
2676	1320771	[Prevention and surveillance of risk groups fo...	0
2046	1379508	Elective versus emergency surgery for patients...	0
2055	1410067	Colorectal cancer screening.	0
2985	1486445	Intraoperative colonoscopy in patients with co...	0
...	...	...	...
339	32412305	Lung Cancer Mortality and the Availability of ...	9
432	32436293	Sequence of biologic therapies and surgery aff...	9
725	32462835	[Small-cell lung cancer: management and novelt...	9
986	32532368	Gender and lung cancer-SEER-based analysis.	9
194	32892231	The P2X7 purinergic receptor: a potential ther...	9

10000 rows × 3 columns

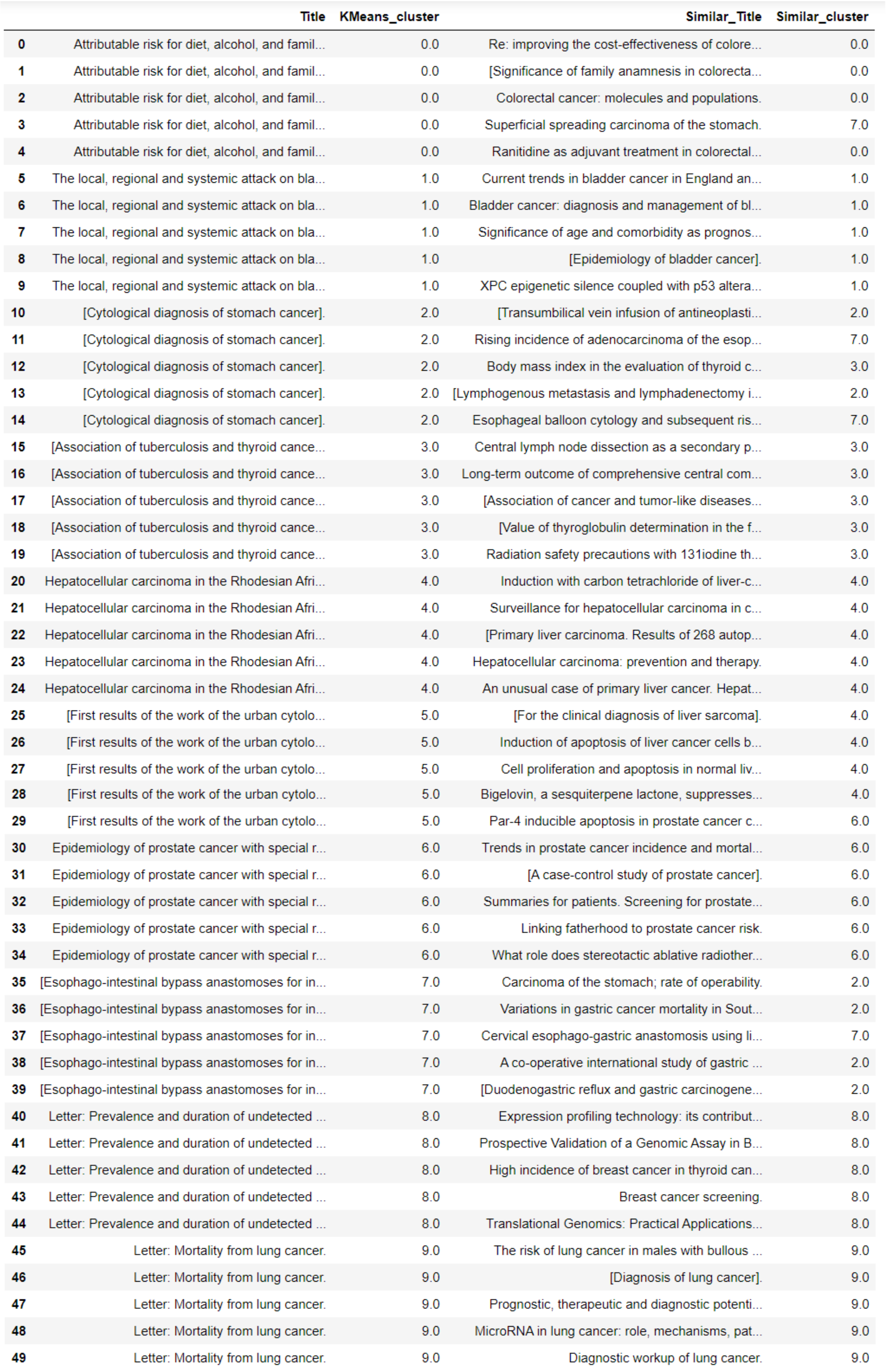
Select one article (using "PUMID"), from each cluster and extract top 5 similarest articles (instances)

To do that, I need to compute the cosine similarity of the feature extraction matrix with the help of **sklearn** library. Then, I take 1 pubmed title from each cluster.



```
['Attributable risk for diet, alcohol, and family history in the Melbourne Colorectal Cancer Study.',
 'The local, regional and systemic attack on bladder cancer.',
 '[Cytological diagnosis of stomach cancer].',
 '[Association of tuberculosis and thyroid cancer. Value of exploratory cervicotomy].',
 'Hepatocellular carcinoma in the Rhodesian African.',
 '[First results of the work of the urban cytological laboratory].',
 'Epidemiology of prostate cancer with special reference to the role of diet.',
 '[Esophago-intestinal bypass anastomoses for inoperable cancer of the proximal portion of the stomach and abdominal portion
of the esophagus].',
 'Letter: Prevalence and duration of undetected breast cancer.',
 'Letter: Mortality from lung cancer.']
```

From each title above, I use the similarity matrix and take 5 greatest value to have the most 5 similar articles. Then I use **numpy** library to get the title of the similar articles. And here is the snippet of the result.

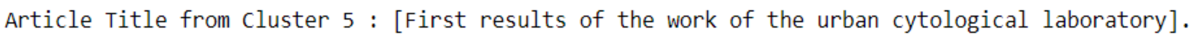


	Title	KMeans_cluster	Similar_Title	Similar_cluster
0	Attributable risk for diet, alcohol, and famil...	0.0	Re: improving the cost-effectiveness of colore...	0.0
1	Attributable risk for diet, alcohol, and famil...	0.0	[Significance of family anamnesis in colorecta...	0.0
2	Attributable risk for diet, alcohol, and famil...	0.0	Colorectal cancer: molecules and populations.	0.0
3	Attributable risk for diet, alcohol, and famil...	0.0	Superficial spreading carcinoma of the stomach.	7.0
4	Attributable risk for diet, alcohol, and famil...	0.0	Ranitidine as adjuvant treatment in colorectal...	0.0
5	The local, regional and systemic attack on bla...	1.0	Current trends in bladder cancer in England an...	1.0
6	The local, regional and systemic attack on bla...	1.0	Bladder cancer: diagnosis and management of bl...	1.0
7	The local, regional and systemic attack on bla...	1.0	Significance of age and comorbidity as prognos...	1.0
8	The local, regional and systemic attack on bla...	1.0	[Epidemiology of bladder cancer].	1.0
9	The local, regional and systemic attack on bla...	1.0	XPC epigenetic silence coupled with p53 altera...	1.0
10	[Cytological diagnosis of stomach cancer].	2.0	[Transumbilical vein infusion of antineoplasti...	2.0
11	[Cytological diagnosis of stomach cancer].	2.0	Rising incidence of adenocarcinoma of the esop...	7.0
12	[Cytological diagnosis of stomach cancer].	2.0	Body mass index in the evaluation of thyroid c...	3.0
13	[Cytological diagnosis of stomach cancer].	2.0	[Lymphogenous metastasis and lymphadenectomy i...	2.0
14	[Cytological diagnosis of stomach cancer].	2.0	Esophageal balloon cytology and subsequent ris...	7.0
15	[Association of tuberculosis and thyroid cance...	3.0	Central lymph node dissection as a secondary p...	3.0
16	[Association of tuberculosis and thyroid cance...	3.0	Long-term outcome of comprehensive central com...	3.0
17	[Association of tuberculosis and thyroid cance...	3.0	[Association of cancer and tumor-like diseases...	3.0
18	[Association of tuberculosis and thyroid cance...	3.0	[Value of thyroglobulin determination in the f...	3.0
19	[Association of tuberculosis and thyroid cance...	3.0	Radiation safety precautions with 131iodine th...	3.0
20	Hepatocellular carcinoma in the Rhodesian Afri...	4.0	Induction with carbon tetrachloride of liver-c...	4.0
21	Hepatocellular carcinoma in the Rhodesian Afri...	4.0	Surveillance for hepatocellular carcinoma in c...	4.0
22	Hepatocellular carcinoma in the Rhodesian Afri...	4.0	[Primary liver carcinoma. Results of 268 autop...	4.0
23	Hepatocellular carcinoma in the Rhodesian Afri...	4.0	Hepatocellular carcinoma: prevention and therapy.	4.0
24	Hepatocellular carcinoma in the Rhodesian Afri...	4.0	An unusual case of primary liver cancer. Hepat...	4.0
25	[First results of the work of the urban cytolo...	5.0	[For the clinical diagnosis of liver sarcoma].	4.0
26	[First results of the work of the urban cytolo...	5.0	Induction of apoptosis of liver cancer cells b...	4.0
27	[First results of the work of the urban cytolo...	5.0	Cell proliferation and apoptosis in normal liv...	4.0
28	[First results of the work of the urban cytolo...	5.0	Bigelovin, a sesquiterpene lactone, suppresses...	4.0
29	[First results of the work of the urban cytolo...	5.0	Par-4 inducible apoptosis in prostate cancer c...	6.0
30	Epidemiology of prostate cancer with special r...	6.0	Trends in prostate cancer incidence and mortal...	6.0
31	Epidemiology of prostate cancer with special r...	6.0	[A case-control study of prostate cancer].	6.0
32	Epidemiology of prostate cancer with special r...	6.0	Summaries for patients. Screening for prostate...	6.0
33	Epidemiology of prostate cancer with special r...	6.0	Linking fatherhood to prostate cancer risk.	6.0
34	Epidemiology of prostate cancer with special r...	6.0	What role does stereotactic ablative radiother...	6.0
35	[Esophago-intestinal bypass anastomoses for in...	7.0	Carcinoma of the stomach; rate of operability.	2.0
36	[Esophago-intestinal bypass anastomoses for in...	7.0	Variations in gastric cancer mortality in Sout...	2.0
37	[Esophago-intestinal bypass anastomoses for in...	7.0	Cervical esophago-gastric anastomosis using li...	7.0
38	[Esophago-intestinal bypass anastomoses for in...	7.0	A co-operative international study of gastric ...	2.0
39	[Esophago-intestinal bypass anastomoses for in...	7.0	[Duodenogastric reflux and gastric carcinogene...	2.0
40	Letter: Prevalence and duration of undetected ...	8.0	Expression profiling technology: its contribut...	8.0
41	Letter: Prevalence and duration of undetected ...	8.0	Prospective Validation of a Genomic Assay in B...	8.0
42	Letter: Prevalence and duration of undetected ...	8.0	High incidence of breast cancer in thyroid can...	8.0
43	Letter: Prevalence and duration of undetected ...	8.0	Breast cancer screening.	8.0
44	Letter: Prevalence and duration of undetected ...	8.0	Translational Genomics: Practical Applications...	8.0
45	Letter: Mortality from lung cancer.	9.0	The risk of lung cancer in males with bullous ...	9.0
46	Letter: Mortality from lung cancer.	9.0	[Diagnosis of lung cancer].	9.0
47	Letter: Mortality from lung cancer.	9.0	Prognostic, therapeutic and diagnostic potenti...	9.0
48	Letter: Mortality from lung cancer.	9.0	MicroRNA in lung cancer: role, mechanisms, pat...	9.0
49	Letter: Mortality from lung cancer.	9.0	Diagnostic workup of lung cancer.	9.0

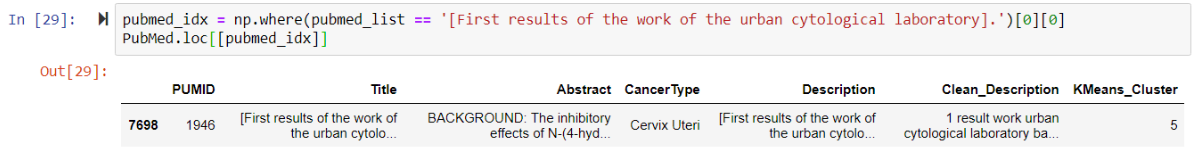
From the result above, we can see that not all representative titles have similar article from the same cluster. For the first title from cluster 0, there is 1 similar article from cluster 7. Then for the second title from cluster 2, the most 5 similar articles are from the same cluster. And etc.

There is something strange with the result for the representative title from cluster 5. All the most 5 similar articles are not from the same cluster. The most 5 similar articles are from cluster 4 and 6. This may happen because the representative title has less-key feature of the cluster. Let us take look for it.

The representative title from cluster 5 is



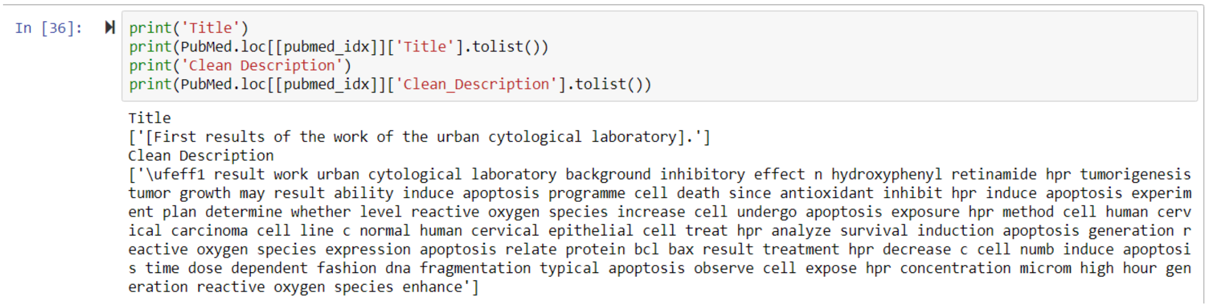
```
Article Title from Cluster 5 : [First results of the work of the urban cytological laboratory].
```



```
In [29]:  pubmed_idx = np.where(pubmed_list == '[First results of the work of the urban cytological laboratory].')[0][0]
          PubMed.loc[[pubmed_idx]]
```

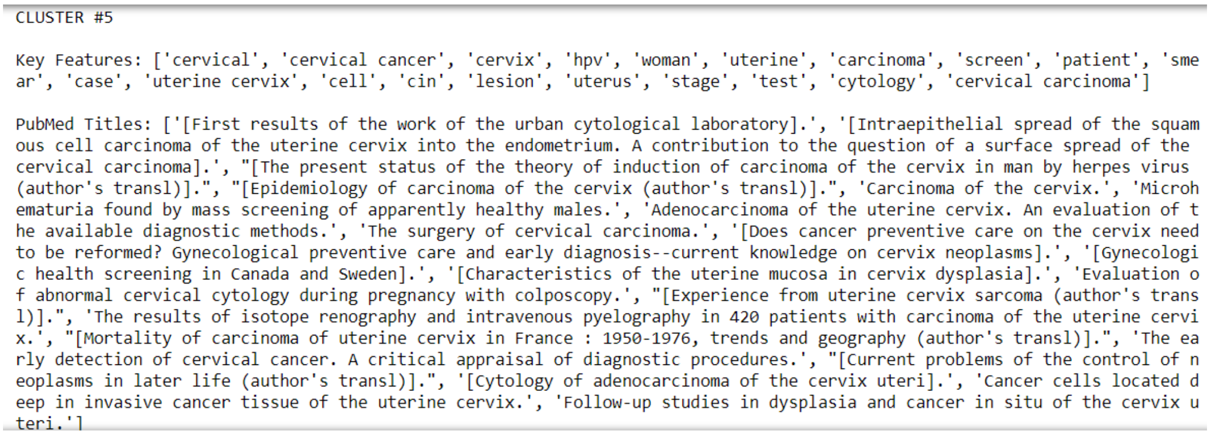
Out[29]:

	PUMID	Title	Abstract	CancerType	Description	Clean_Description	KMeans_Cluster
7698	1946	[First results of the work of the urban cytolo...	BACKGROUND: The inhibitory effects of N-(4-hyd...	Cervix Uteri	[First results of the work of the urban cytolo...	1 result work urban cytological laboratory ba...	5



```
In [36]:  print('Title')
          print(PubMed.loc[[pubmed_idx]]['Title'].tolist())
          print('Clean Description')
          print(PubMed.loc[[pubmed_idx]]['Clean_Description'].tolist())

          Title
          ['[First results of the work of the urban cytological laboratory].']
          Clean Description
          ['\ufeff1 result work urban cytological laboratory background inhibitory effect n hydroxyphenyl retinamide hpr tumorigenesis
          tumor growth may result ability induce apoptosis programme cell death since antioxidant inhibit hpr induce apoptosis experim
          ent plan determine whether level reactive oxygen species increase cell undergo apoptosis exposure hpr method cell human cerv
          ical carcinoma cell line c normal human cervical epithelial cell treat hpr analyze survival induction apoptosis generation r
          eactive oxygen species expression apoptosis relate protein bcl bax result treatment hpr decrease c cell numb induce apoptosi
          s time dose dependent fashion dna fragmentation typical apoptosis observe cell expose hpr concentration microm high hour gen
          eration reactive oxygen species enhance']
```

While the key features of cluster 5 are

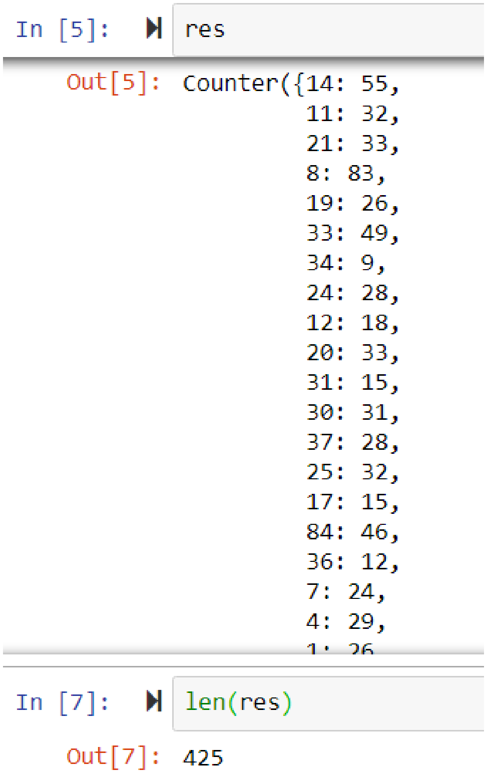


```
CLUSTER #5

Key Features: ['cervical', 'cervical cancer', 'cervix', 'hpv', 'woman', 'uterine', 'carcinoma', 'screen', 'patient', 'sme
ar', 'case', 'uterine cervix', 'cell', 'cin', 'lesion', 'uterus', 'stage', 'test', 'cytology', 'cervical carcinoma']

PubMed Titles: ['[First results of the work of the urban cytological laboratory].', '[Intraepithelial spread of the squam
ous cell carcinoma of the uterine cervix into the endometrium. A contribution to the question of a surface spread of the
cervical carcinoma].', "[The present status of the theory of induction of carcinoma of the cervix in man by herpes virus
(author's transl)].", "[Epidemiology of carcinoma of the cervix (author's transl)].", 'Carcinoma of the cervix.', 'Microh
ematuria found by mass screening of apparently healthy males.', 'Adenocarcinoma of the uterine cervix. An evaluation of t
he available diagnostic methods.', 'The surgery of cervical carcinoma.', '[Does cancer preventive care on the cervix need
to be reformed? Gynecological preventive care and early diagnosis--current knowledge on cervix neoplasms].', '[Gynecologi
c health screening in Canada and Sweden].', '[Characteristics of the uterine mucosa in cervix dysplasia].', 'Evaluation o
f abnormal cervical cytology during pregnancy with colposcopy.', "[Experience from uterine cervix sarcoma (author's trans
l)].", 'The results of isotope renography and intravenous pyelography in 420 patients with carcinoma of the uterine cervi
x.', "[Mortality of carcinoma of uterine cervix in France : 1950-1976, trends and geography (author's transl)].", 'The ea
rly detection of cervical cancer. A critical appraisal of diagnostic procedures.', "[Current problems of the control of n
eoplasms in later life (author's transl)].", '[Cytology of adenocarcinoma of the cervix uteri].', 'Cancer cells located d
eep in invasive cancer tissue of the uterine cervix.', 'Follow-up studies in dysplasia and cancer in situ of the cervix u
teri.']
```

As we can see that the article has less key feature which in the top 20 features. I think these is one of the reasons why.

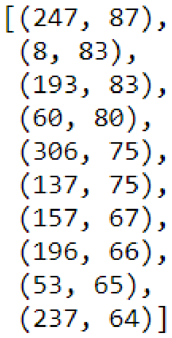
2. Text Clustering Using Affinity Propagation Cluster

First, I use TF-IDF for feature extraction. I limit the feature by ngram_range is unigram and bigram, minimum document frequency is 10, and maximum document frequency is 0.8. Then for the clustering, I firstly compute the cosine similarity of the feature extraction matrix. Then using Affinity Propagation with the number of maximum iterations is 10 to have less computing time. Here is the result of the clustering. The left item is the cluster number and the right one is the total number of articles in the cluster.



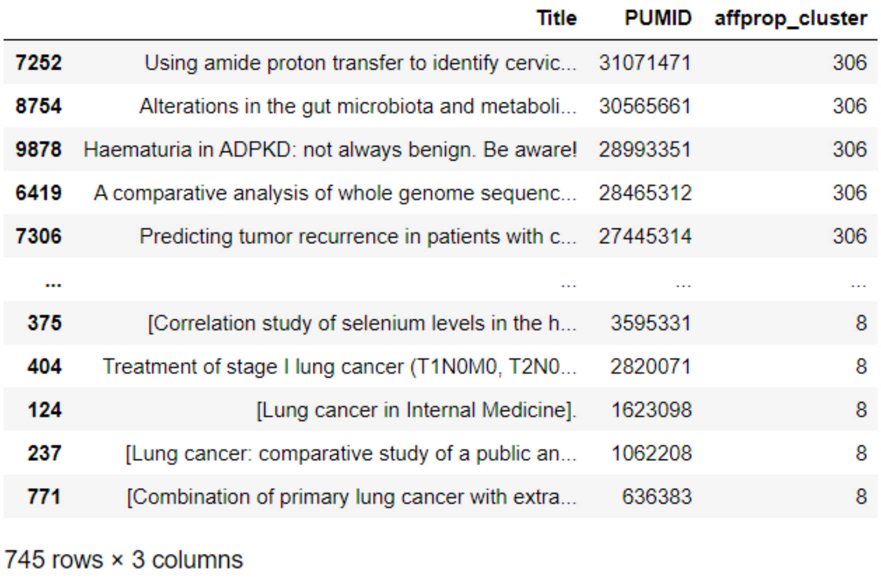
```
In [5]:  res

Out[5]: Counter({14: 55,
                 11: 32,
                 21: 33,
                 8: 83,
                 19: 26,
                 33: 49,
                 34: 9,
                 24: 28,
                 12: 18,
                 20: 33,
                 31: 15,
                 30: 31,
                 37: 28,
                 25: 32,
                 17: 15,
                 84: 46,
                 36: 12,
                 7: 24,
                 4: 29,
                 1: 26

In [7]:  len(res)

Out[7]: 425
```

The Affinity Propagation give result of 425 cluster. But, to have the same number of clusters, I take the 10 most common cluster. And here is the result.



```
[(247, 87),
 (8, 83),
 (193, 83),
 (60, 80),
 (306, 75),
 (137, 75),
 (157, 67),
 (196, 66),
 (53, 65),
 (237, 64)]
```

This would affect the number of data used to become 745 data for all the 10 clusters. Here is the detail information about the clustering result.

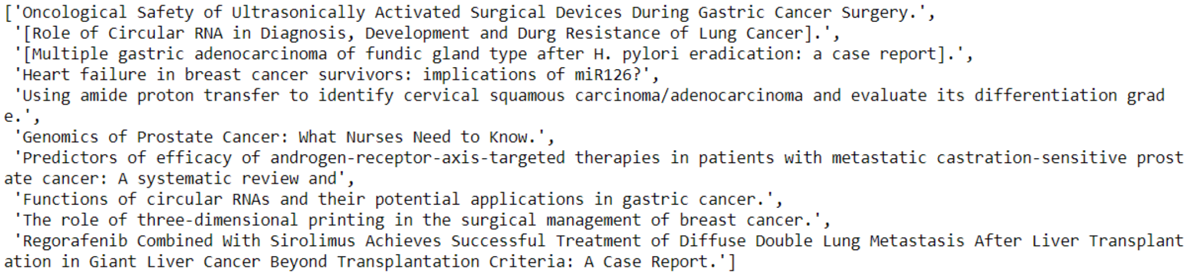


	Title	PUMID	affprop_cluster
7252	Using amide proton transfer to identify cervic...	31071471	306
8754	Alterations in the gut microbiota and metaboli...	30565661	306
9878	Haematuria in ADPKD: not always benign. Be aware!	28993351	306
6419	A comparative analysis of whole genome sequenc...	28465312	306
7306	Predicting tumor recurrence in patients with c...	27445314	306
...	...	...	...
375	[Correlation study of selenium levels in the h...	3595331	8
404	Treatment of stage I lung cancer (T1N0M0, T2N0...	2820071	8
124	[Lung cancer in Internal Medicine].	1623098	8
237	[Lung cancer: comparative study of a public an...	1062208	8
771	[Combination of primary lung cancer with extra...	636383	8

745 rows × 3 columns

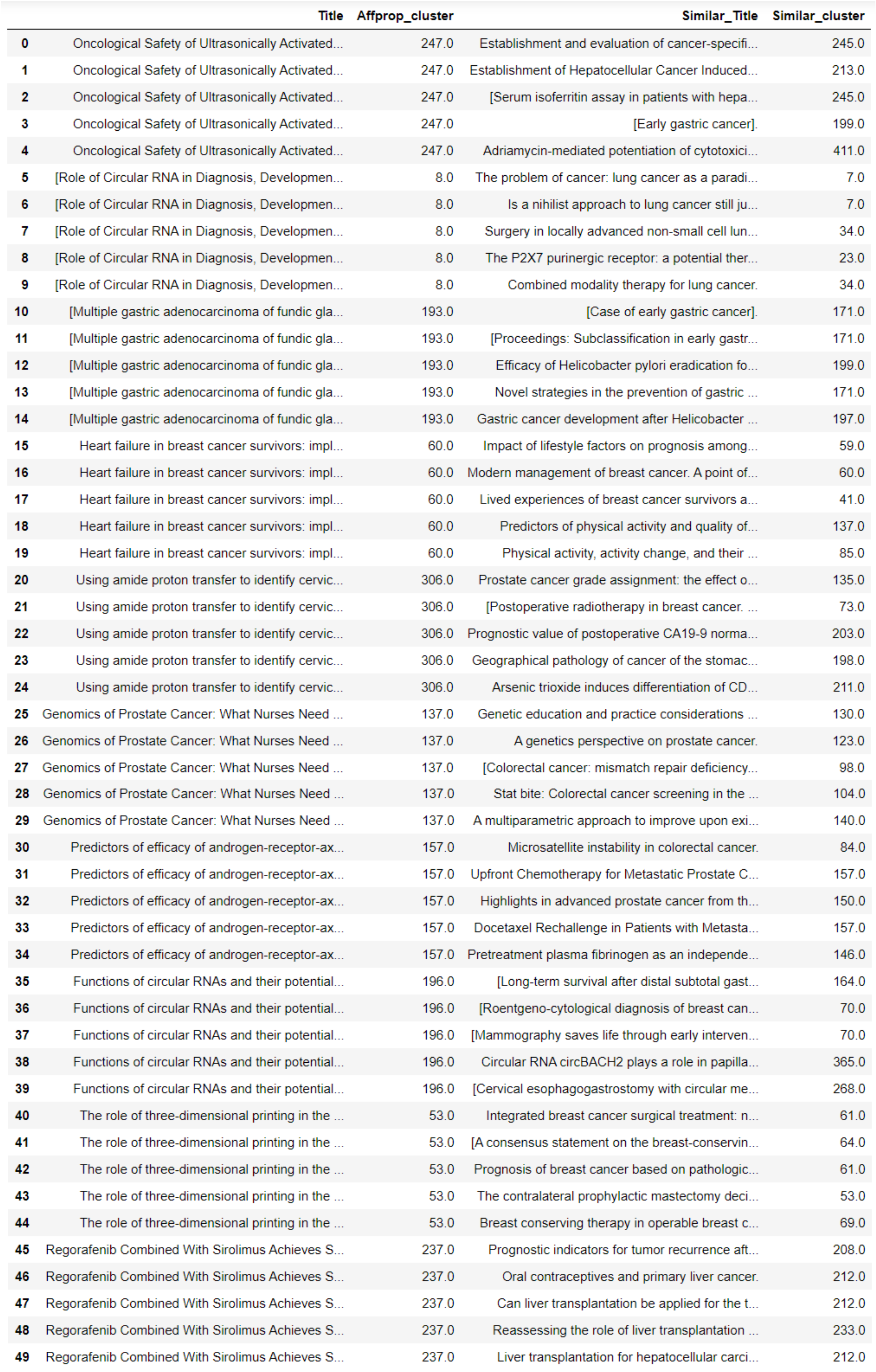
Select one article (using "PUMID"), from each cluster and extract top 5 similarest articles (instances)

To do that, I need to compute the cosine similarity of the feature extraction matrix with the help of **sklearn** library. Then, I take 1 pubmed title from each cluster.



```
['Oncological Safety of Ultrasonically Activated Surgical Devices During Gastric Cancer Surgery.',
 '[Role of Circular RNA in Diagnosis, Development and Durg Resistance of Lung Cancer].',
 '[Multiple gastric adenocarcinoma of fundic gland type after H. pylori eradication: a case report].',
 'Heart failure in breast cancer survivors: implications of miR126?',
 'Using amide proton transfer to identify cervical squamous carcinoma/adenocarcinoma and evaluate its differentiation grad
e.',
 'Genomics of Prostate Cancer: What Nurses Need to Know.',
 'Predictors of efficacy of androgen-receptor-axis-targeted therapies in patients with metastatic castration-sensitive prost
ate cancer: A systematic review and',
 'Functions of circular RNAs and their potential applications in gastric cancer.',
 'The role of three-dimensional printing in the surgical management of breast cancer.',
 'Regorafenib Combined With Sirolimus Achieves Successful Treatment of Diffuse Double Lung Metastasis After Liver Transplant
ation in Giant Liver Cancer Beyond Transplantation Criteria: A Case Report.']
```

From each title above, I use the similarity matrix and take 5 greatest value to have the most 5 similar articles. Then I use **numpy** library to get the title of the similar articles. And here is the snippet of the result.



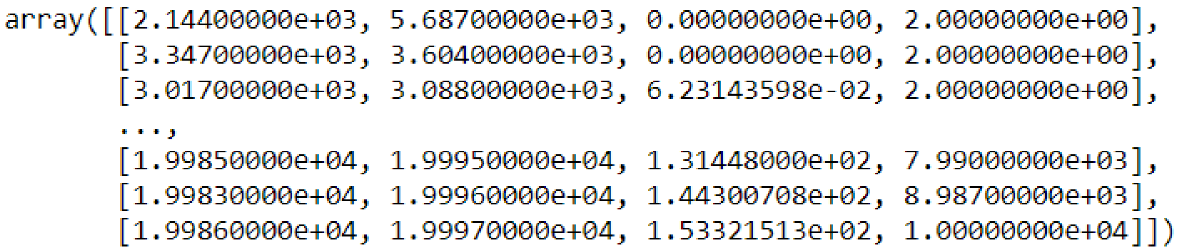
	Title	Affprop_cluster	Similar_Title	Similar_cluster
0	Oncological Safety of Ultrasonically Activated...	247.0	Establishment and evaluation of cancer-specifi...	245.0
1	Oncological Safety of Ultrasonically Activated...	247.0	Establishment of Hepatocellular Cancer Induced...	213.0
2	Oncological Safety of Ultrasonically Activated...	247.0	[Serum isoferritin assay in patients with hepa...	245.0
3	Oncological Safety of Ultrasonically Activated...	247.0	[Early gastric cancer].	199.0
4	Oncological Safety of Ultrasonically Activated...	247.0	Adriamycin-mediated potentiation of cytotoxici...	411.0
5	[Role of Circular RNA in Diagnosis, Developmen...	8.0	The problem of cancer: lung cancer as a paradi...	7.0
6	[Role of Circular RNA in Diagnosis, Developmen...	8.0	Is a nihilist approach to lung cancer still ju...	7.0
7	[Role of Circular RNA in Diagnosis, Developmen...	8.0	Surgery in locally advanced non-small cell lun...	34.0
8	[Role of Circular RNA in Diagnosis, Developmen...	8.0	The P2X7 purinergic receptor: a potential ther...	23.0
9	[Role of Circular RNA in Diagnosis, Developmen...	8.0	Combined modality therapy for lung cancer.	34.0
10	[Multiple gastric adenocarcinoma of fundic gla...	193.0	[Case of early gastric cancer].	171.0
11	[Multiple gastric adenocarcinoma of fundic gla...	193.0	[Proceedings: Subclassification in early gastr...	171.0
12	[Multiple gastric adenocarcinoma of fundic gla...	193.0	Efficacy of Helicobacter pylori eradication fo...	199.0
13	[Multiple gastric adenocarcinoma of fundic gla...	193.0	Novel strategies in the prevention of gastric ...	171.0
14	[Multiple gastric adenocarcinoma of fundic gla...	193.0	Gastric cancer development after Helicobacter ...	197.0
15	Heart failure in breast cancer survivors: impl...	60.0	Impact of lifestyle factors on prognosis among...	59.0
16	Heart failure in breast cancer survivors: impl...	60.0	Modern management of breast cancer. A point of...	60.0
17	Heart failure in breast cancer survivors: impl...	60.0	Lived experiences of breast cancer survivors a...	41.0
18	Heart failure in breast cancer survivors: impl...	60.0	Predictors of physical activity and quality of...	137.0
19	Heart failure in breast cancer survivors: impl...	60.0	Physical activity, activity change, and their ...	85.0
20	Using amide proton transfer to identify cervic...	306.0	Prostate cancer grade assignment: the effect o...	135.0
21	Using amide proton transfer to identify cervic...	306.0	[Postoperative radiotherapy in breast cancer. ...	73.0
22	Using amide proton transfer to identify cervic...	306.0	Prognostic value of postoperative CA19-9 norma...	203.0
23	Using amide proton transfer to identify cervic...	306.0	Geographical pathology of cancer of the stomac...	198.0
24	Using amide proton transfer to identify cervic...	306.0	Arsenic trioxide induces differentiation of CD...	211.0
25	Genomics of Prostate Cancer: What Nurses Need ...	137.0	Genetic education and practice considerations ...	130.0
26	Genomics of Prostate Cancer: What Nurses Need ...	137.0	A genetics perspective on prostate cancer.	123.0
27	Genomics of Prostate Cancer: What Nurses Need ...	137.0	[Colorectal cancer: mismatch repair deficiency...	98.0
28	Genomics of Prostate Cancer: What Nurses Need ...	137.0	Stat bite: Colorectal cancer screening in the ...	104.0
29	Genomics of Prostate Cancer: What Nurses Need ...	137.0	A multiparametric approach to improve upon exi...	140.0
30	Predictors of efficacy of androgen-receptor-ax...	157.0	Microsatellite instability in colorectal cancer.	84.0
31	Predictors of efficacy of androgen-receptor-ax...	157.0	Upfront Chemotherapy for Metastatic Prostate C...	157.0
32	Predictors of efficacy of androgen-receptor-ax...	157.0	Highlights in advanced prostate cancer from th...	150.0
33	Predictors of efficacy of androgen-receptor-ax...	157.0	Docetaxel Rechallenge in Patients with Metasta...	157.0
34	Predictors of efficacy of androgen-receptor-ax...	157.0	Pretreatment plasma fibrinogen as an independe...	146.0
35	Functions of circular RNAs and their potential...	196.0	[Long-term survival after distal subtotal gast...	164.0
36	Functions of circular RNAs and their potential...	196.0	[Roentgeno-cytological diagnosis of breast can...	70.0
37	Functions of circular RNAs and their potential...	196.0	[Mammography saves life through early interven...	70.0
38	Functions of circular RNAs and their potential...	196.0	Circular RNA circBACH2 plays a role in papilla...	365.0
39	Functions of circular RNAs and their potential...	196.0	[Cervical esophagogastrostomy with circular me...	268.0
40	The role of three-dimensional printing in the ...	53.0	Integrated breast cancer surgical treatment: n...	61.0
41	The role of three-dimensional printing in the ...	53.0	[A consensus statement on the breast-conservin...	64.0
42	The role of three-dimensional printing in the ...	53.0	Prognosis of breast cancer based on pathologic...	61.0
43	The role of three-dimensional printing in the ...	53.0	The contralateral prophylactic mastectomy deci...	53.0
44	The role of three-dimensional printing in the ...	53.0	Breast conserving therapy in operable breast c...	69.0
45	Regorafenib Combined With Sirolimus Achieves S...	237.0	Prognostic indicators for tumor recurrence aft...	208.0
46	Regorafenib Combined With Sirolimus Achieves S...	237.0	Oral contraceptives and primary liver cancer.	212.0
47	Regorafenib Combined With Sirolimus Achieves S...	237.0	Can liver transplantation be applied for the t...	212.0
48	Regorafenib Combined With Sirolimus Achieves S...	237.0	Reassessing the role of liver transplantation ...	233.0
49	Regorafenib Combined With Sirolimus Achieves S...	237.0	Liver transplantation for hepatocellular carci...	212.0

From the result above, we can see that almost all the similar articles have different cluster with the representative title. For example, in the first title from cluster 247. The similar articles are from cluster 245, 213, 199, 411. While for the representative title from cluster 60 have only 1 similar article from the same cluster (index number 16).

It may happen because the number of clusters is 425 that makes the clustering becomes more diverse. So that most of all the similar articles are derived from the different cluster.

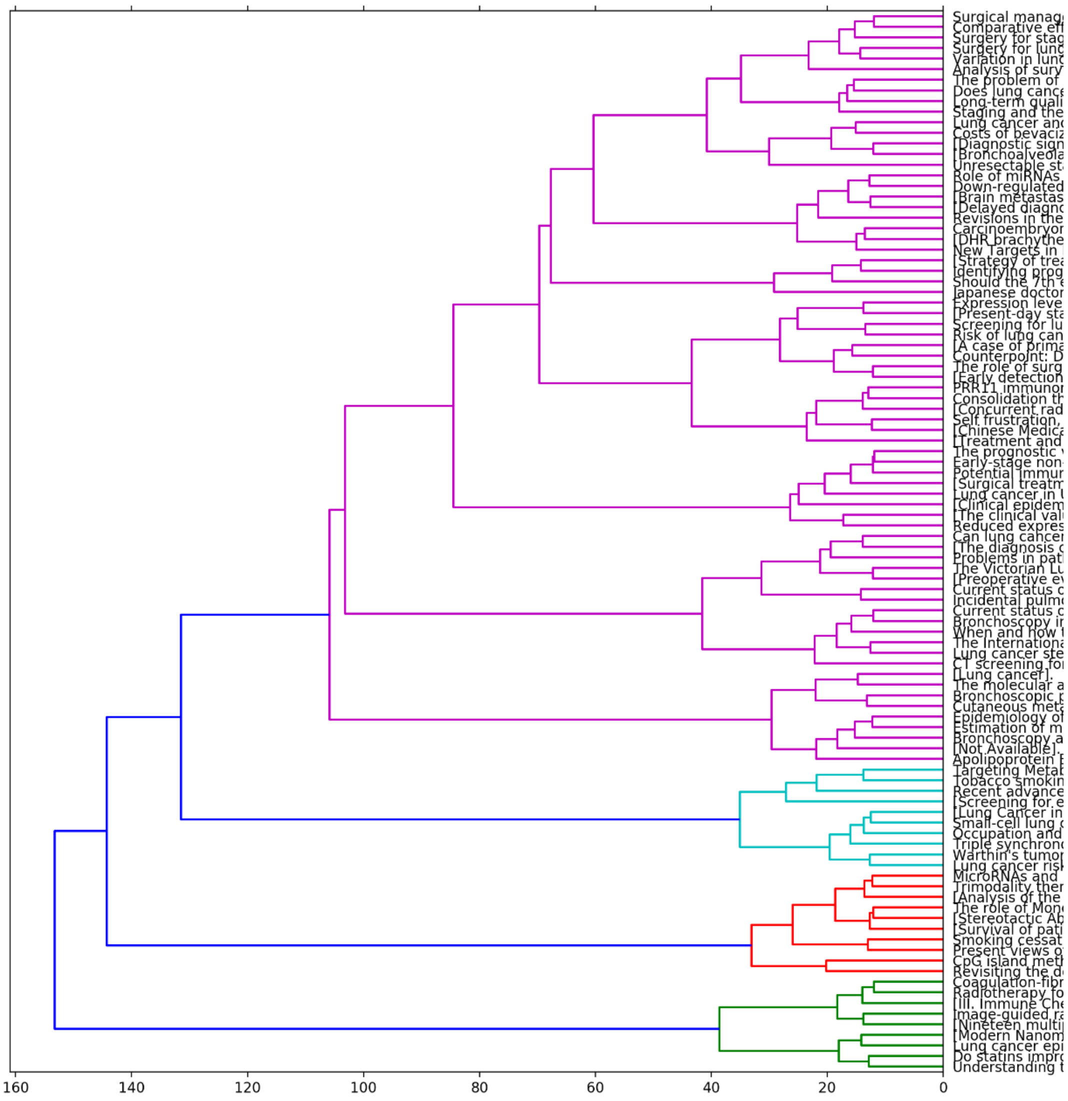
3. Text Clustering Using Ward Cluster Algorithm

First, I use TF-IDF for feature extraction. I limit the feature by ngram_range is unigram and bigram, minimum document frequency is 10, and maximum document frequency is 0.8. Then for the clustering, I firstly compute the cosine similarity of the feature extraction matrix. Then using Ward Algorithm of the cosine distance to have linkage matrix. Here is the linkage matrix of the clustering.

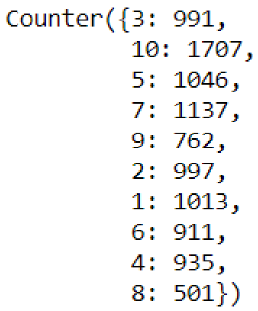


```
array([[2.14400000e+03, 5.68700000e+03, 0.00000000e+00, 2.00000000e+00],
       [3.34700000e+03, 3.60400000e+03, 0.00000000e+00, 2.00000000e+00],
       [3.01700000e+03, 3.08800000e+03, 6.23143598e-02, 2.00000000e+00],
       ...,
       [1.99850000e+04, 1.99950000e+04, 1.31448000e+02, 7.99000000e+03],
       [1.99830000e+04, 1.99960000e+04, 1.44300708e+02, 8.98700000e+03],
       [1.99860000e+04, 1.99970000e+04, 1.53321513e+02, 1.00000000e+04]])
```

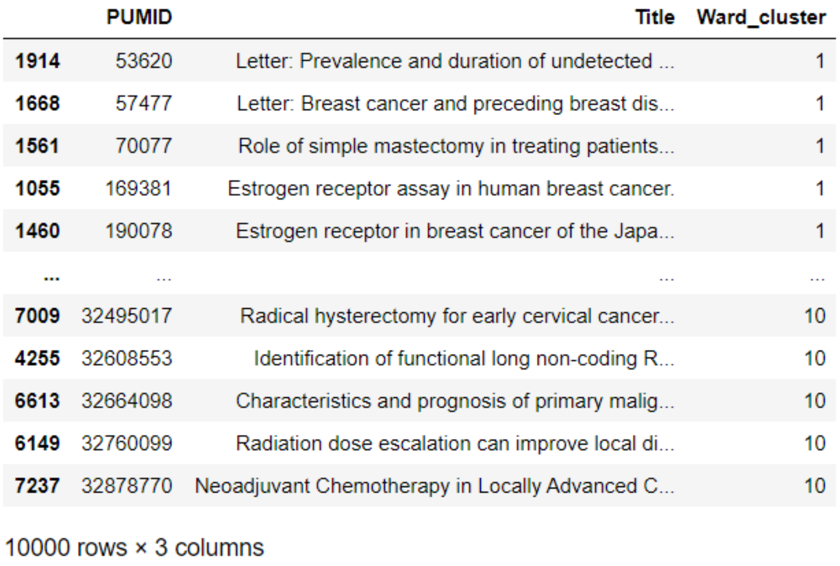
And it is plotted using dendogram as shown below.



Because ward algorithm links all the data become 1 cluster, so I need to cut the link. I use **fcluster** and cut at the 60th link to get 10 cluster. And here is the result.



```
Counter({3: 991,
         10: 1707,
         5: 1046,
         7: 1137,
         9: 762,
         2: 997,
         1: 1013,
         6: 911,
         4: 935,
         8: 501})
```

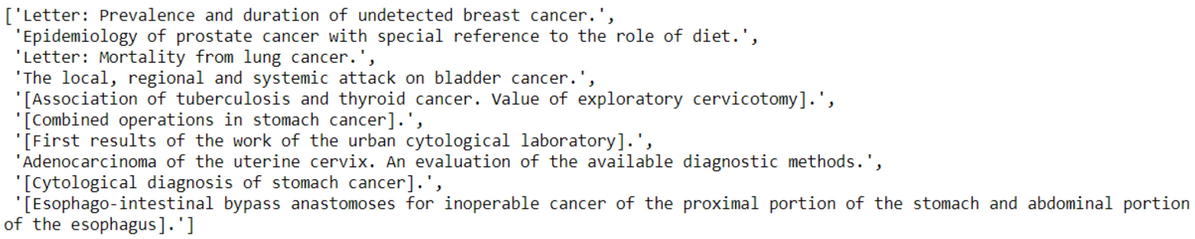


	PUMID	Title	Ward_cluster
1914	53620	Letter: Prevalence and duration of undetected ...	1
1668	57477	Letter: Breast cancer and preceding breast dis...	1
1561	70077	Role of simple mastectomy in treating patients...	1
1055	169381	Estrogen receptor assay in human breast cancer.	1
1460	190078	Estrogen receptor in breast cancer of the Japa...	1
...	...	...	...
7009	32495017	Radical hysterectomy for early cervical cancer...	10
4255	32608553	Identification of functional long non-coding R...	10
6613	32664098	Characteristics and prognosis of primary malig...	10
6149	32760099	Radiation dose escalation can improve local di...	10
7237	32878770	Neoadjuvant Chemotherapy in Locally Advanced C...	10

10000 rows × 3 columns

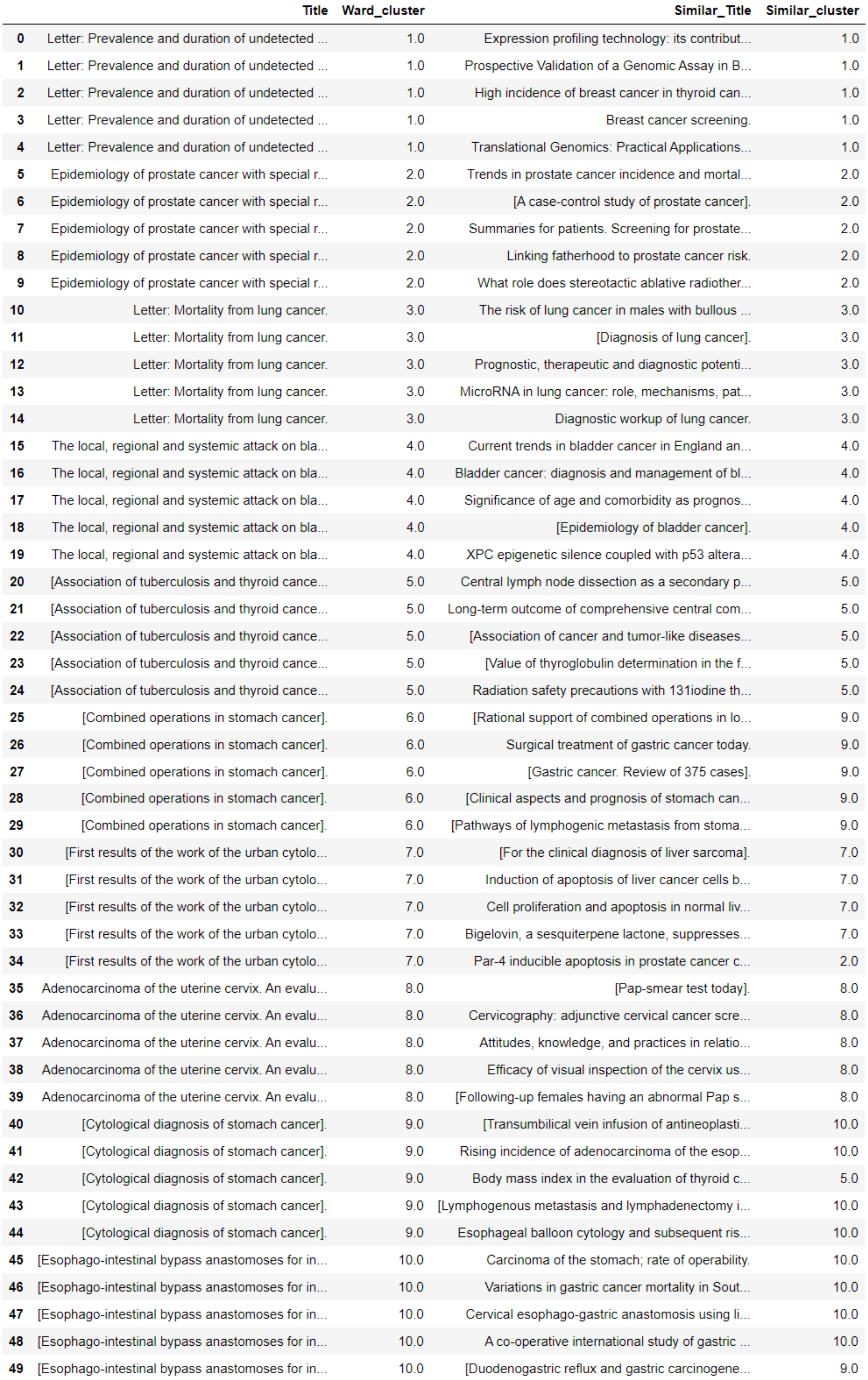
Select one article (using "PUMID"), from each cluster and extract top 5 similarest articles (instances)

To do that, I need to compute the cosine similarity of the feature extraction matrix with the help of **sklearn** library. Then, I take 1 pubmed title from each cluster.



```
['Letter: Prevalence and duration of undetected breast cancer.',
 'Epidemiology of prostate cancer with special reference to the role of diet.',
 'Letter: Mortality from lung cancer.',
 'The local, regional and systemic attack on bladder cancer.',
 '[Association of tuberculosis and thyroid cancer. Value of exploratory cervicotomy].',
 '[Combined operations in stomach cancer].',
 '[First results of the work of the urban cytological laboratory].',
 'Adenocarcinoma of the uterine cervix. An evaluation of the available diagnostic methods.',
 '[Cytological diagnosis of stomach cancer].',
 '[Esophago-intestinal bypass anastomoses for inoperable cancer of the proximal portion of the stomach and abdominal portion
of the esophagus].']
```

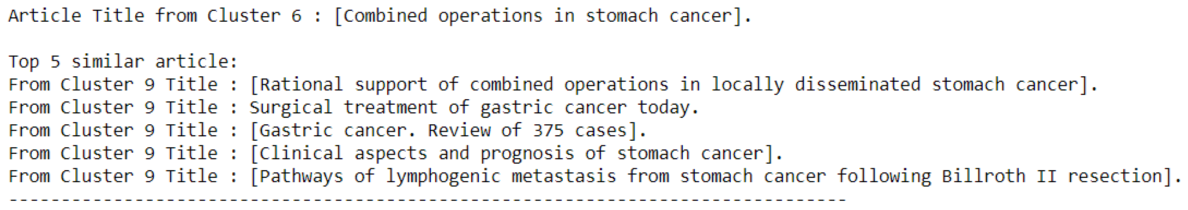
From each title above, I use the similarity matrix and take 5 greatest value to have the most 5 similar articles. Then I use **numpy** library to get the title of the similar articles. And here is the snippet of the result.



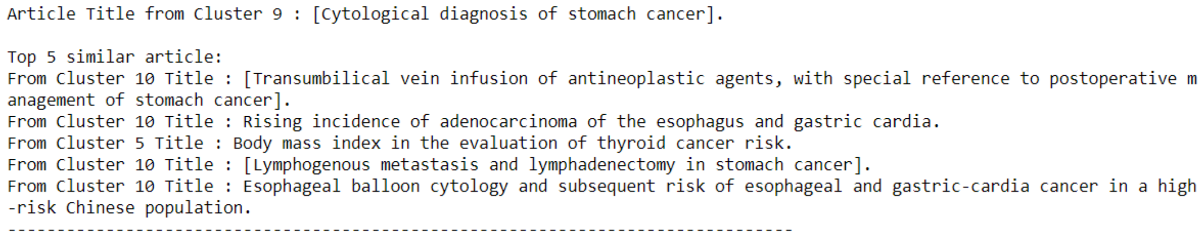
	Title	Ward_cluster	Similar_Title	Similar_cluster
0	Letter: Prevalence and duration of undetected ...	1.0	Expression profiling technology: its contribut...	1.0
1	Letter: Prevalence and duration of undetected ...	1.0	Prospective Validation of a Genomic Assay in B...	1.0
2	Letter: Prevalence and duration of undetected ...	1.0	High incidence of breast cancer in thyroid can...	1.0
3	Letter: Prevalence and duration of undetected ...	1.0	Breast cancer screening.	1.0
4	Letter: Prevalence and duration of undetected ...	1.0	Translational Genomics: Practical Applications...	1.0
5	Epidemiology of prostate cancer with special r...	2.0	Trends in prostate cancer incidence and mortal...	2.0
6	Epidemiology of prostate cancer with special r...	2.0	[A case-control study of prostate cancer].	2.0
7	Epidemiology of prostate cancer with special r...	2.0	Summaries for patients. Screening for prostate...	2.0
8	Epidemiology of prostate cancer with special r...	2.0	Linking fatherhood to prostate cancer risk.	2.0
9	Epidemiology of prostate cancer with special r...	2.0	What role does stereotactic ablative radiother...	2.0
10	Letter: Mortality from lung cancer.	3.0	The risk of lung cancer in males with bullous ...	3.0
11	Letter: Mortality from lung cancer.	3.0	[Diagnosis of lung cancer].	3.0
12	Letter: Mortality from lung cancer.	3.0	Prognostic, therapeutic and diagnostic potenti...	3.0
13	Letter: Mortality from lung cancer.	3.0	MicroRNA in lung cancer: role, mechanisms, pat...	3.0
14	Letter: Mortality from lung cancer.	3.0	Diagnostic workup of lung cancer.	3.0
15	The local, regional and systemic attack on bla...	4.0	Current trends in bladder cancer in England an...	4.0
16	The local, regional and systemic attack on bla...	4.0	Bladder cancer: diagnosis and management of bl...	4.0
17	The local, regional and systemic attack on bla...	4.0	Significance of age and comorbidity as prognos...	4.0
18	The local, regional and systemic attack on bla...	4.0	[Epidemiology of bladder cancer].	4.0
19	The local, regional and systemic attack on bla...	4.0	XPC epigenetic silence coupled with p53 altera...	4.0
20	[Association of tuberculosis and thyroid cance...	5.0	Central lymph node dissection as a secondary p...	5.0
21	[Association of tuberculosis and thyroid cance...	5.0	Long-term outcome of comprehensive central com...	5.0
22	[Association of tuberculosis and thyroid cance...	5.0	[Association of cancer and tumor-like diseases...	5.0
23	[Association of tuberculosis and thyroid cance...	5.0	[Value of thyroglobulin determination in the f...	5.0
24	[Association of tuberculosis and thyroid cance...	5.0	Radiation safety precautions with 131iodine th...	5.0
25	[Combined operations in stomach cancer].	6.0	[Rational support of combined operations in lo...	9.0
26	[Combined operations in stomach cancer].	6.0	Surgical treatment of gastric cancer today.	9.0
27	[Combined operations in stomach cancer].	6.0	[Gastric cancer. Review of 375 cases].	9.0
28	[Combined operations in stomach cancer].	6.0	[Clinical aspects and prognosis of stomach can...	9.0
29	[Combined operations in stomach cancer].	6.0	[Pathways of lymphogenic metastasis from stoma...	9.0
30	[First results of the work of the urban cytolo...	7.0	[For the clinical diagnosis of liver sarcoma].	7.0
31	[First results of the work of the urban cytolo...	7.0	Induction of apoptosis of liver cancer cells b...	7.0
32	[First results of the work of the urban cytolo...	7.0	Cell proliferation and apoptosis in normal liv...	7.0
33	[First results of the work of the urban cytolo...	7.0	Bigelovin, a sesquiterpene lactone, suppresses...	7.0
34	[First results of the work of the urban cytolo...	7.0	Par-4 inducible apoptosis in prostate cancer c...	2.0
35	Adenocarcinoma of the uterine cervix. An evalu...	8.0	[Pap-smear test today].	8.0
36	Adenocarcinoma of the uterine cervix. An evalu...	8.0	Cervicography: adjunctive cervical cancer scre...	8.0
37	Adenocarcinoma of the uterine cervix. An evalu...	8.0	Attitudes, knowledge, and practices in relatio...	8.0
38	Adenocarcinoma of the uterine cervix. An evalu...	8.0	Efficacy of visual inspection of the cervix us...	8.0
39	Adenocarcinoma of the uterine cervix. An evalu...	8.0	[Following-up females having an abnormal Pap s...	8.0
40	[Cytological diagnosis of stomach cancer].	9.0	[Transumbilical vein infusion of antineoplasti...	10.0
41	[Cytological diagnosis of stomach cancer].	9.0	Rising incidence of adenocarcinoma of the esop...	10.0
42	[Cytological diagnosis of stomach cancer].	9.0	Body mass index in the evaluation of thyroid c...	5.0
43	[Cytological diagnosis of stomach cancer].	9.0	[Lymphogenous metastasis and lymphadenectomy i...	10.0
44	[Cytological diagnosis of stomach cancer].	9.0	Esophageal balloon cytology and subsequent ris...	10.0
45	[Esophago-intestinal bypass anastomoses for in...	10.0	Carcinoma of the stomach; rate of operability.	10.0
46	[Esophago-intestinal bypass anastomoses for in...	10.0	Variations in gastric cancer mortality in Sout...	10.0
47	[Esophago-intestinal bypass anastomoses for in...	10.0	Cervical esophago-gastric anastomosis using li...	10.0
48	[Esophago-intestinal bypass anastomoses for in...	10.0	A co-operative international study of gastric ...	10.0
49	[Esophago-intestinal bypass anastomoses for in...	10.0	[Duodenogastric reflux and gastric carcinogene...	9.0

From the result above, we can see that almost all the representative titles have 5 most similar articles from the same cluster. For the representative title from cluster 7 and 10, there is 1 similar article that is from the different cluster (index 34 and 49). This may happened because the similar article has the more key features to the representative cluster.

While for the representative title from cluster 6 and 9, all the 5 most similar articles are from different cluster.



```
Article Title from Cluster 6 : [Combined operations in stomach cancer].

Top 5 similar article:
From Cluster 9 Title : [Rational support of combined operations in locally disseminated stomach cancer].
From Cluster 9 Title : Surgical treatment of gastric cancer today.
From Cluster 9 Title : [Gastric cancer. Review of 375 cases].
From Cluster 9 Title : [Clinical aspects and prognosis of stomach cancer].
From Cluster 9 Title : [Pathways of lymphogenic metastasis from stomach cancer following Billroth II resection].
--------------------------------------------------------------------------------
```



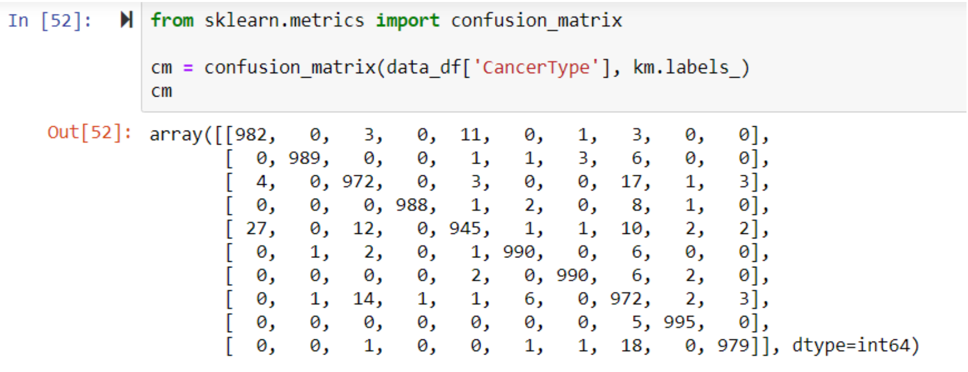
```
Article Title from Cluster 9 : [Cytological diagnosis of stomach cancer].

Top 5 similar article:
From Cluster 10 Title : [Transumbilical vein infusion of antineoplastic agents, with special reference to postoperative m
anagement of stomach cancer].
From Cluster 10 Title : Rising incidence of adenocarcinoma of the esophagus and gastric cardia.
From Cluster 5 Title : Body mass index in the evaluation of thyroid cancer risk.
From Cluster 10 Title : [Lymphogenous metastasis and lymphadenectomy in stomach cancer].
From Cluster 10 Title : Esophageal balloon cytology and subsequent risk of esophageal and gastric-cardia cancer in a high
-risk Chinese population.
--------------------------------------------------------------------------------
```

In my opinion, the result become seen above because of representative titles. As we can see that the representative title is categorical as stomach cancer, but it is from different cluster. So that the most 5 similar articles also become different.

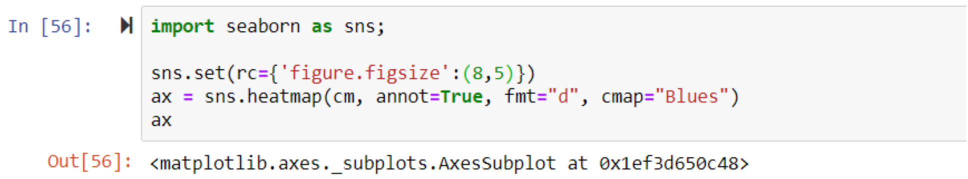
4. Comparison of Clustering Method

a. K-Means

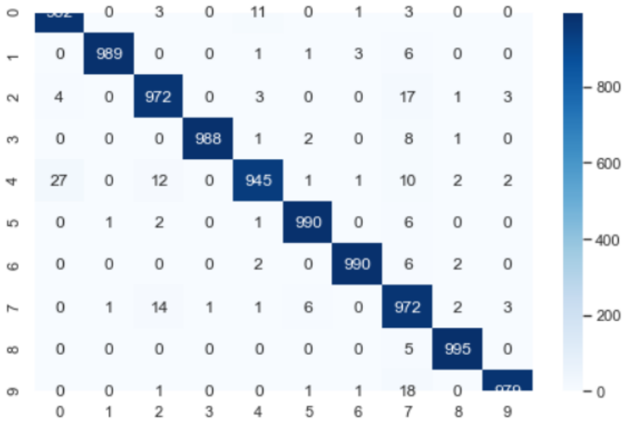
Here is the confusion matrix of predicted labels (using K-Means) and the actual labels (from the cancer type).



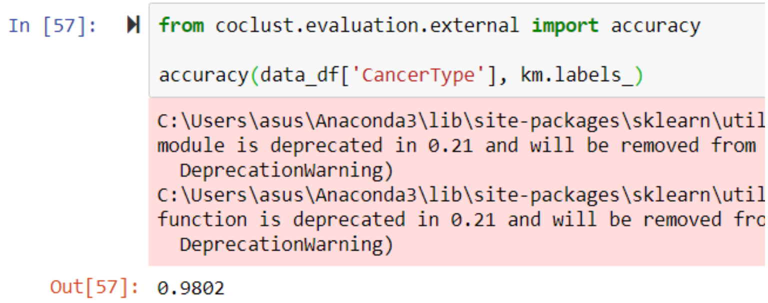
```
In [52]:  from sklearn.metrics import confusion_matrix

          cm = confusion_matrix(data_df['CancerType'], km.labels_)
          cm

Out[52]:  array([[982,   0,   3,   0,  11,   0,   1,   3,   0,   0],
                 [  0, 989,   0,   0,   1,   1,   3,   6,   0,   0],
                 [  4,   0, 972,   0,   3,   0,   0,  17,   1,   3],
                 [  0,   0,   0, 988,   1,   2,   0,   8,   1,   0],
                 [ 27,   0,  12,   0, 945,   1,   1,  10,   2,   2],
                 [  0,   1,   2,   0,   1, 990,   0,   6,   0,   0],
                 [  0,   0,   0,   0,   2,   0, 990,   6,   2,   0],
                 [  0,   1,  14,   1,   1,   6,   0, 972,   2,   3],
                 [  0,   0,   0,   0,   0,   0,   0,   5, 995,   0],
                 [  0,   0,   1,   0,   0,   1,   1,  18,   0, 979]], dtype=int64)
```



```
In [56]:  import seaborn as sns;

          sns.set(rc={'figure.figsize':(8,5)})
          ax = sns.heatmap(cm, annot=True, fmt="d", cmap="Blues")
          ax

Out[56]:  <matplotlib.axes._subplots.AxesSubplot at 0x1ef3d650c48>
```

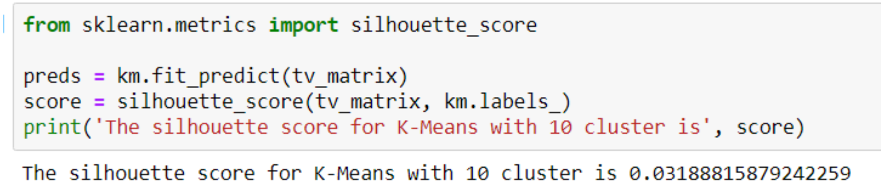


Then, for computing the accuracy of K-Means, I use accuracy from **coclust** library (https://smorbieu.gitlab.io/accuracy-from-classification-to-clustering-evaluation/). And the result of the K-Means accuracy is 98.02%.



```
In [57]:  from coclust.evaluation.external import accuracy

          accuracy(data_df['CancerType'], km.labels_)

          C:\Users\asus\Anaconda3\lib\site-packages\sklearn\util
          module is deprecated in 0.21 and will be removed from
            DeprecationWarning)
          C:\Users\asus\Anaconda3\lib\site-packages\sklearn\util
          function is deprecated in 0.21 and will be removed fro
            DeprecationWarning)

Out[57]:  0.9802
```

I also use silhouette score from the **sklearn** library for the 10 cluster, and here is the result.

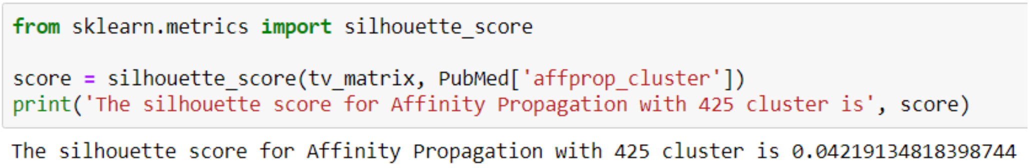


```
from sklearn.metrics import silhouette_score

preds = km.fit_predict(tv_matrix)
score = silhouette_score(tv_matrix, km.labels_)
print('The silhouette score for K-Means with 10 cluster is', score)
```

```
The silhouette score for K-Means with 10 cluster is 0.03188815879242259
```

b. Affinity

Because affinity propagation’s result has 425 cluster, so I cannot use **coclust** accuracy. While using silhouette score, the score is 0.04219134818398744.

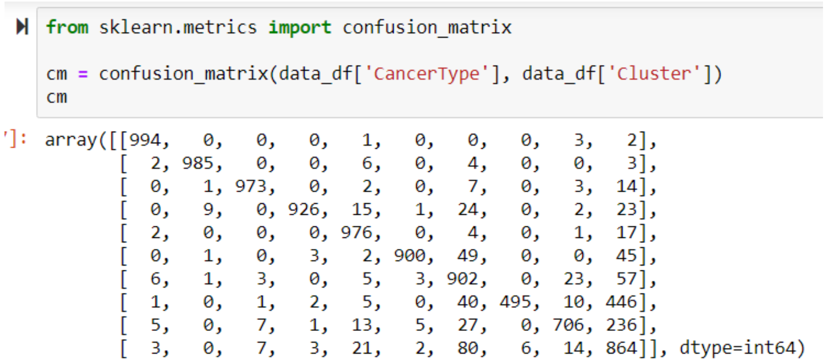


```
from sklearn.metrics import silhouette_score

score = silhouette_score(tv_matrix, PubMed['affprop_cluster'])
print('The silhouette score for Affinity Propagation with 425 cluster is', score)
```

```
The silhouette score for Affinity Propagation with 425 cluster is 0.04219134818398744
```

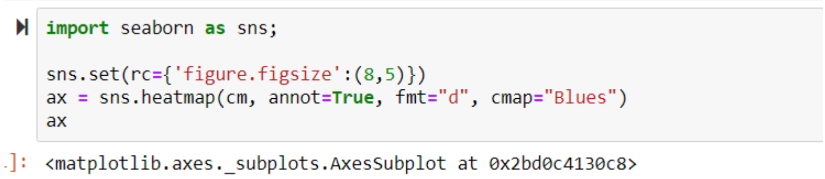
c. Ward Algorithm

Here is the confusion matrix of predicted labels (using Ward) and the actual labels (from the cancer type).



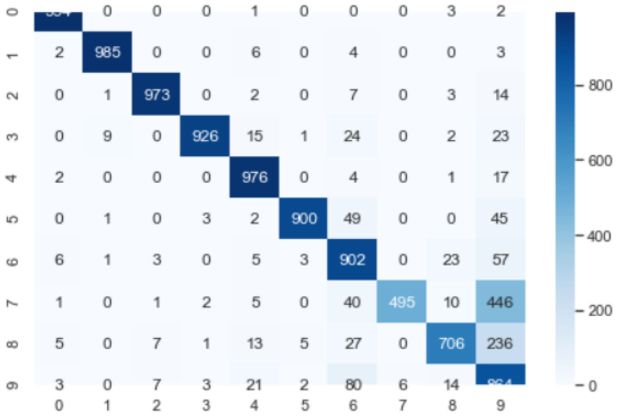
```
from sklearn.metrics import confusion_matrix

cm = confusion_matrix(data_df['CancerType'], data_df['Cluster'])
cm
```

```
']: array([[994,   0,   0,   0,   1,   0,   0,   0,   3,   2],
           [  2, 985,   0,   0,   6,   0,   4,   0,   0,   3],
           [  0,   1, 973,   0,   2,   0,   7,   0,   3,  14],
           [  0,   9,   0, 926,  15,   1,  24,   0,   2,  23],
           [  2,   0,   0,   0, 976,   0,   4,   0,   1,  17],
           [  0,   1,   0,   3,   2, 900,  49,   0,   0,  45],
           [  6,   1,   3,   0,   5,   3, 902,   0,  23,  57],
           [  1,   0,   1,   2,   5,   0,  40, 495,  10, 446],
           [  5,   0,   7,   1,  13,   5,  27,   0, 706, 236],
           [  3,   0,   7,   3,  21,   2,  80,   6,  14, 864]], dtype=int64)
```

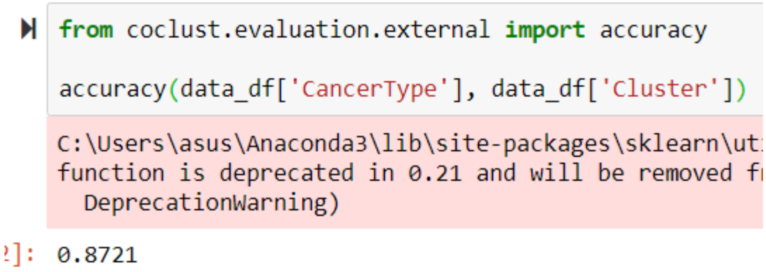


```
import seaborn as sns;

sns.set(rc={'figure.figsize':(8,5)})
ax = sns.heatmap(cm, annot=True, fmt="d", cmap="Blues")
ax
```

```
.]: <matplotlib.axes._subplots.AxesSubplot at 0x2bd0c4130c8>
```



Then, for computing the accuracy of Ward, I use accuracy from **coclust** library (https://smorbieu.gitlab.io/accuracy-from-classification-to-clustering-evaluation/). And the result of the Ward accuracy is 87.21%.

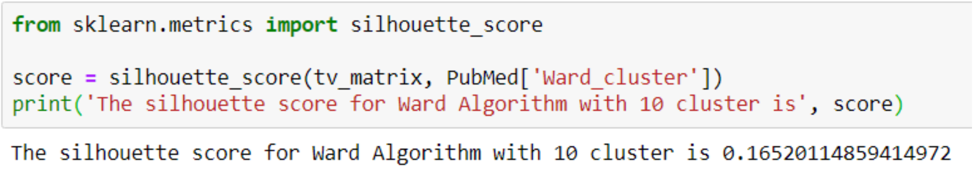


```
from coclust.evaluation.external import accuracy

accuracy(data_df['CancerType'], data_df['Cluster'])
```

```
C:\Users\asus\Anaconda3\lib\site-packages\sklearn\ut
function is deprecated in 0.21 and will be removed fr
  DeprecationWarning)
```

```
]: 0.8721
```

I also use silhouette score from the **sklearn** library for the 10 cluster, and here is the result.

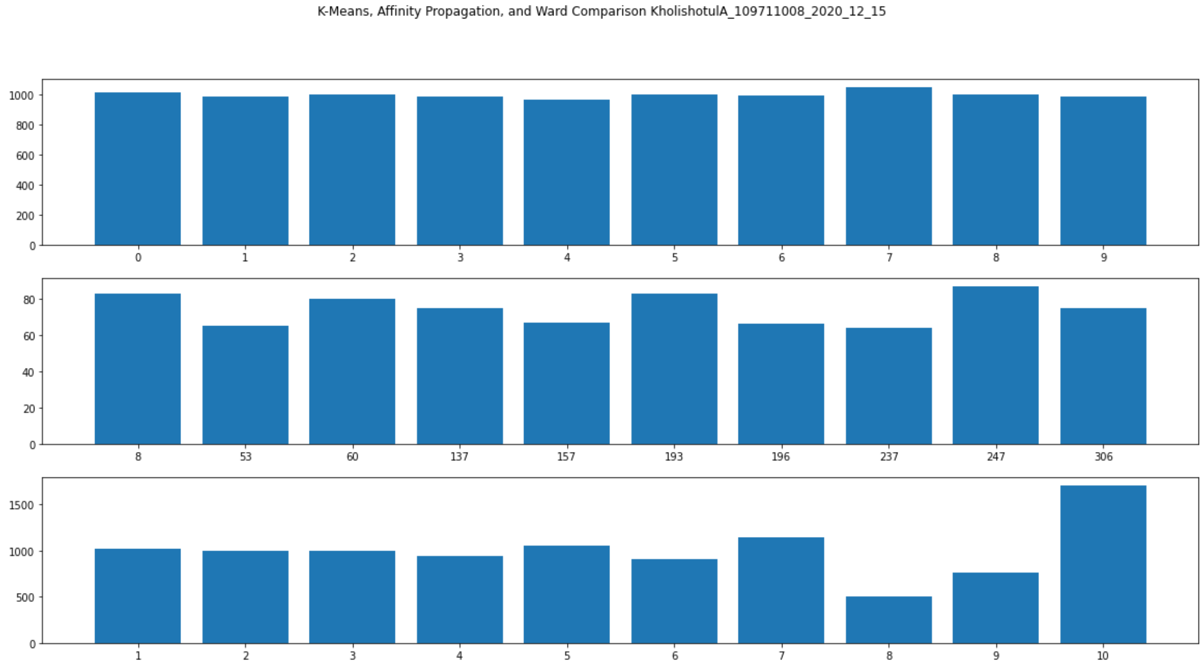


```
from sklearn.metrics import silhouette_score

score = silhouette_score(tv_matrix, PubMed['Ward_cluster'])
print('The silhouette score for Ward Algorithm with 10 cluster is', score)
```

```
The silhouette score for Ward Algorithm with 10 cluster is 0.16520114859414972
```

Observation :

The spread of the article within cluster are shown below. The graph ordered as K-Means, Affinity Propagation, Ward.



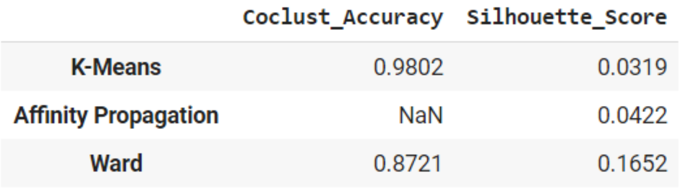
K-Means, Affinity Propagation, and Ward Comparison KholishotulA_109711008_2020_12_15

As we can see that the spread of K-Means is balance in range of 950 – 1100 articles for each cluster. I cannot say anything for the Affinity Propagation, because I used only 10 most common cluster in the figure above. While Ward Algorithm has imbalanced spread. The difference of cluster 8 and 10 are big.

The :mod:`coclust.evaluation.external` module provides functions to evaluate clustering or co-clustering results with external information such as the true labeling of the clusters.

The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b). To clarify, b is the distance between a sample and the nearest cluster that the sample is not a part of.

Based on the accuracy, K-Means has higher accuracy than Ward Algorithm. But I have no idea compared to Affinity Propagation. While based on the silhouette score, Ward Algorithm has highest score.

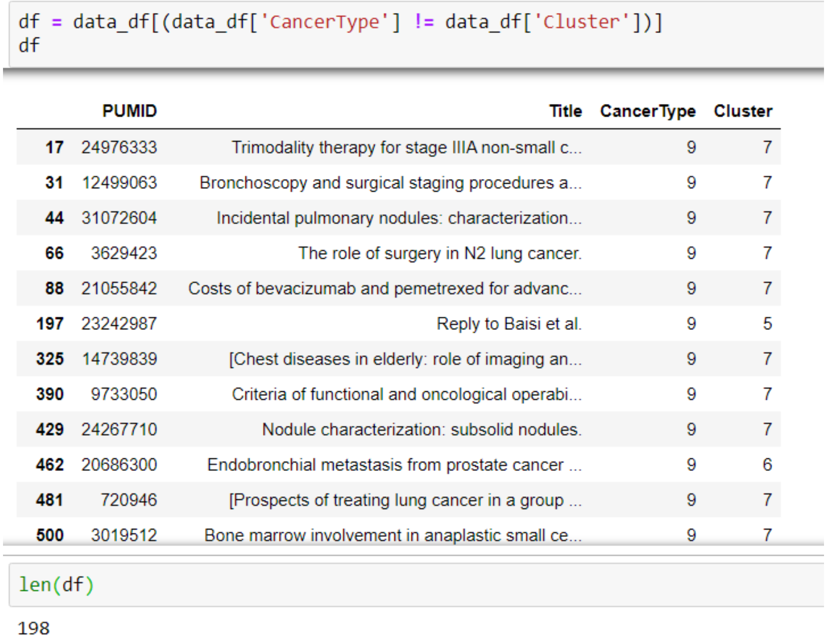


	Coclust_Accuracy	Silhouette_Score
K-Means	0.9802	0.0319
Affinity Propagation	NaN	0.0422
Ward	0.8721	0.1652

5. Misclassified Articles

a. K-Means

There are 198 misclassified articles.



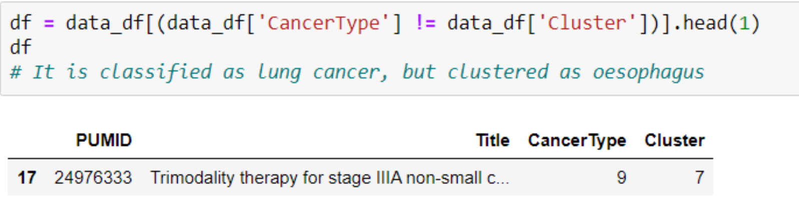
```
df = data_df[(data_df['CancerType'] != data_df['Cluster'])]
df
```

```
        PUMID                                              Title  CancerType  Cluster
17   24976333   Trimodality therapy for stage IIIA non-small c...           9        7
31   12499063  Bronchoscopy and surgical staging procedures a...           9        7
44   31072604   Incidental pulmonary nodules: characterization...           9        7
66    3629423                   The role of surgery in N2 lung cancer.      9        7
88   21055842   Costs of bevacizumab and pemetrexed for advanc...           9        7
197  23242987                                 Reply to Baisi et al.      9        5
325  14739839     [Chest diseases in elderly: role of imaging an...         9        7
390   9733050     Criteria of functional and oncological operabi...         9        7
429  24267710            Nodule characterization: subsolid nodules.       9        7
462  20686300    Endobronchial metastasis from prostate cancer ...          9        6
481    720946      [Prospects of treating lung cancer in a group ...        9        7
500   3019512   Bone marrow involvement in anaplastic small ce...           9        7
```

```
len(df)
```

```
198
```

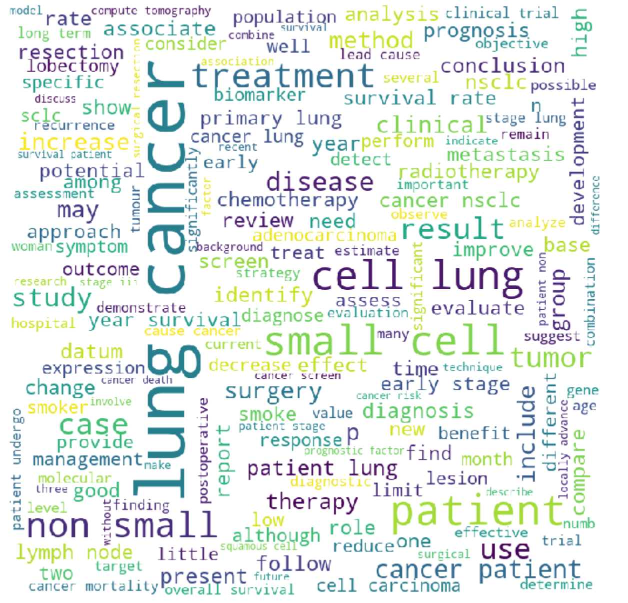
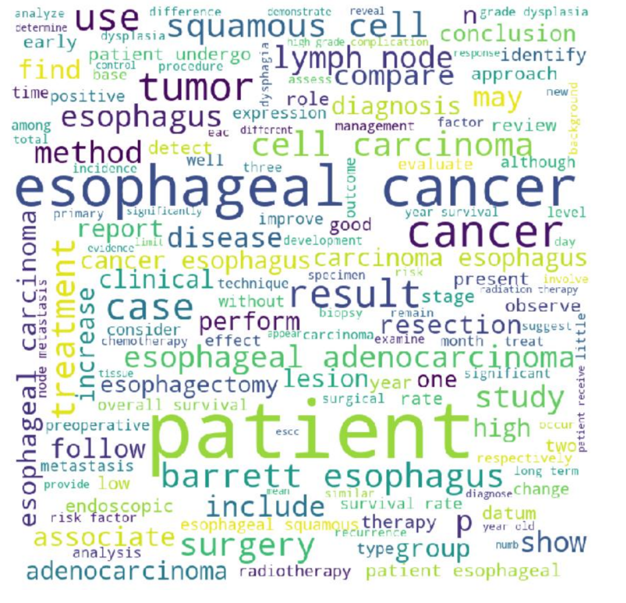
I use this article.



```
df = data_df[(data_df['CancerType'] != data_df['Cluster'])].head(1)
df
# It is classified as lung cancer, but clustered as oesophagus
```

```
       PUMID                                          Title  CancerType  Cluster
17  24976333  Trimodality therapy for stage IIIA non-small c...           9        7
```

Then I use word cloud to evaluate from the two cluster.

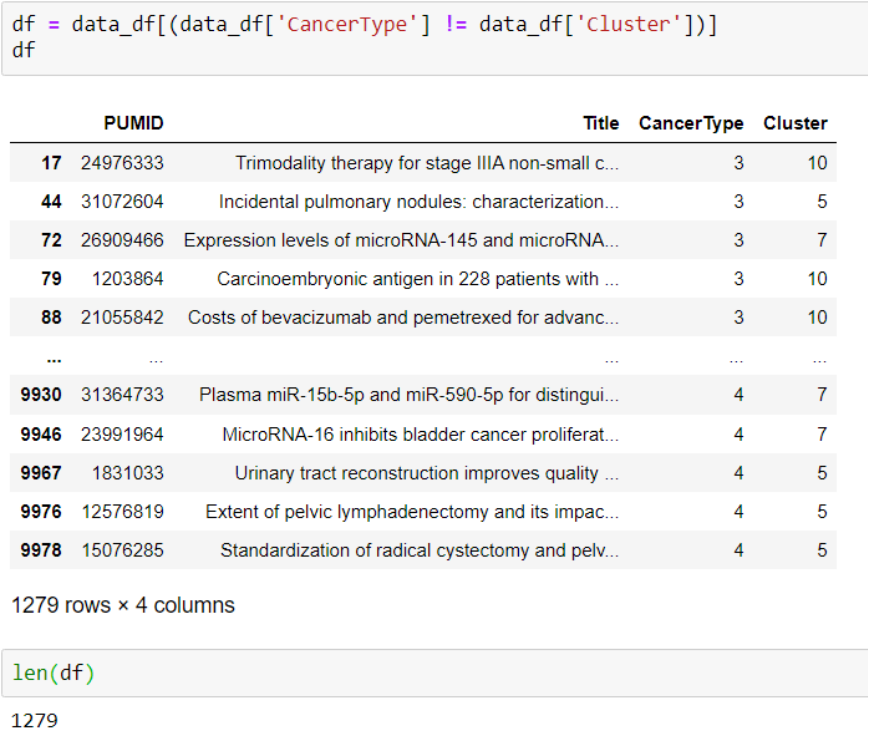
Word cloud of classification 9	Word cloud of cluster 7
	

b. Affinity Propagation

I cannot specify the misclassified articles. Because the articles are classified to 10 classes, while the cluster result of Affinity Propagation has 425 clusters.

c. Ward

There are 1279 misclassified articles.



```
df = data_df[(data_df['CancerType'] != data_df['Cluster'])]
df
```

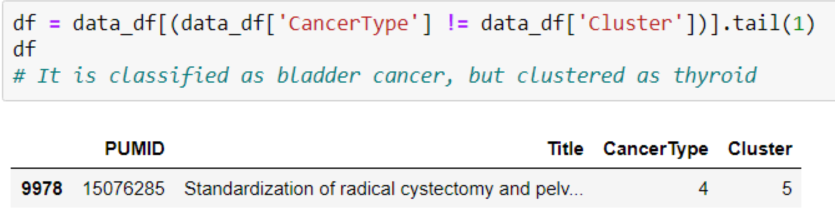
	PUMID	Title	CancerType	Cluster
17	24976333	Trimodality therapy for stage IIIA non-small c...	3	10
44	31072604	Incidental pulmonary nodules: characterization...	3	5
72	26909466	Expression levels of microRNA-145 and microRNA...	3	7
79	1203864	Carcinoembryonic antigen in 228 patients with ...	3	10
88	21055842	Costs of bevacizumab and pemetrexed for advanc...	3	10
...	...	...	...	...
9930	31364733	Plasma miR-15b-5p and miR-590-5p for distingui...	4	7
9946	23991964	MicroRNA-16 inhibits bladder cancer proliferat...	4	7
9967	1831033	Urinary tract reconstruction improves quality ...	4	5
9976	12576819	Extent of pelvic lymphadenectomy and its impac...	4	5
9978	15076285	Standardization of radical cystectomy and pelv...	4	5

1279 rows × 4 columns

```
len(df)
```

```
1279
```

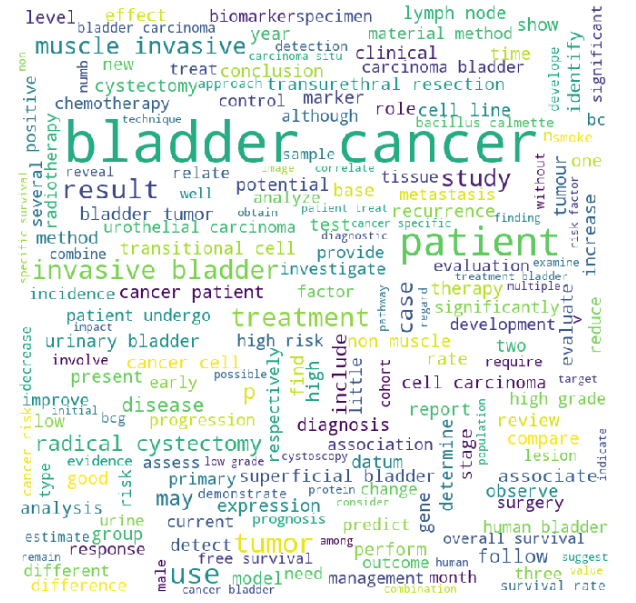
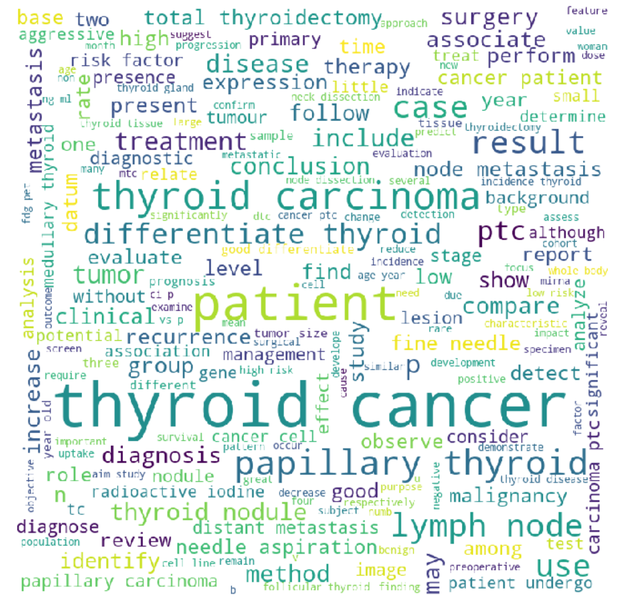
I use this article.



```
df = data_df[(data_df['CancerType'] != data_df['Cluster'])].tail(1)
df
# It is classified as bladder cancer, but clustered as thyroid
```

	PUMID	Title	CancerType	Cluster
9978	15076285	Standardization of radical cystectomy and pelv...	4	5

Then I use word cloud to evaluate from the two cluster.

Word cloud of classification 4	Word cloud of cluster 5
	

CONCLUSION

Based on the observation, text mining can be used to cluster data from PubMed. There are different results comes from different methods of clustering. Based on the observation, K-Means and Ward Clustering Algorithm performs best for clustering data with specific number of clusters. While Affinity Propagation performs best when we have no idea about the number of clusters.

K-Means is one of clustering method that can handle big data, but within the medium number of clusters. The number of clusters can be specified by using elbow method, or we can specify it ourselves.

Affinity propagation is a clustering method that can handle uneven cluster size. The algorithm creates clusters by sending messages between pairs of samples until convergence. Because of that, the execution time of this algorithm is slow.

Ward is a clustering method that can handle large data with large number of clusters. The algorithm uses dendogram as the method to find the optimum number of clusters, or we can specify it ourselves. The metric used is the distance between points.

In the future, using other text clustering methods are supposed to be conducted to compare which method performs best for clustering.

LINK :

The full code of program can be found in

https://github.com/kholishotula/TextMining_Homework2

The Youtube link can be found in

https://youtu.be/x-h5lyb9ndg