Homework I Report

Kholishotul Amaliah (ID : 109711008)

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

INTRODUCTION

Text analytics, also known as text mining, is defined as the methodology and process followed to derive quality and actionable information and insights from textual data. This involves using natural language processing, information retrieval, and machine learning techniques to parse unstructured text data into more structured forms and deriving patterns and insights from this data that would be helpful to the end user.

PubMed is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. The United States National Library of Medicine (NLM) at the National Institutes of Health maintain the database as part of the Entrez system of information retrieval. PubMed comprises more than 30 million citations for biomedical literature from MEDLINE, life science journals, and online books. Since PubMed is one of big data, text mining will be useful to get easier information retrieval.

In this report, I use Top 10 cancer type journals from PubMed. From the journal, text mining will be used is common words. Different methods of text preprocessing will be observed in this report.

METHODS

A. Getting the Data Needed

The data are medical publication about cancer. The top 10 global cancer incidence based on https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data. The top 10 cancer types are :

1. Lung
2. Breast
3. Colorectal
4. Prostate
5. Stomach
6. Liver
7. Oesophagus
8. Cervix uteri
9. Thyroid
10. Bladder

The medical publications are downloaded from https://pubmed.ncbi.nlm.nih.gov/. I download the PubMed type (.txt) and CSV type. But the website allowed to download up to 10,000 publications only. Here is the list of the files I downloaded.

csv-Top1-lungcancer-set	12/10/2020 22:03	Microsoft Excel Co...	2.794 KB
csv-Top2-Breastcanc-set	12/10/2020 21:56	Microsoft Excel Co...	2.677 KB
csv-Top3-Colorectal-set	12/10/2020 21:55	Microsoft Excel Co...	3.034 KB
csv-Top4-prostateca-set	13/10/2020 10:29	Microsoft Excel Co...	2.942 KB
csv-Top5-stomachcan-set	13/10/2020 10:32	Microsoft Excel Co...	2.657 KB
csv-Top6-livercance-set	13/10/2020 10:34	Microsoft Excel Co...	3.091 KB
csv-Top7-oesophagus-set	13/10/2020 10:35	Microsoft Excel Co...	2.871 KB
csv-Top8-Cervixuter-set	13/10/2020 10:36	Microsoft Excel Co...	2.714 KB
csv-Top9-Thyroidcan-set	13/10/2020 10:37	Microsoft Excel Co...	3.115 KB
csv-Top10-Bladdercan-set	13/10/2020 10:38	Microsoft Excel Co...	3.077 KB
pubmed-Top1-lungcancer-set	12/10/2020 22:08	Text Document	22.707 KB
pubmed-Top2-Breastcanc-set	12/10/2020 22:10	Text Document	20.917 KB
pubmed-Top3-Colorectal-set	12/10/2020 22:12	Text Document	27.713 KB
pubmed-Top4-prostateca-set	13/10/2020 10:29	Text Document	26.144 KB
pubmed-Top5-stomachcan-set	13/10/2020 10:32	Text Document	19.525 KB
pubmed-Top6-livercance-set	13/10/2020 10:34	Text Document	29.170 KB
pubmed-Top7-oesophagus-set	13/10/2020 10:35	Text Document	23.140 KB
pubmed-Top8-Cervixuter-set	13/10/2020 10:36	Text Document	19.483 KB
pubmed-Top9-Thyroidcan-set	13/10/2020 10:37	Text Document	29.684 KB
pubmed-Top10-Bladdercan-set	13/10/2020 10:38	Text Document	28.189 KB

B. Data Preparation

Before processing the text, I need to process the downloaded files. I take the PMID, Title, and Abstract from the publication details and add the CancerType attribute. After that, put into the excel file for each cancer type within name ‘pubmed-Top<number_in_list>-<cancer_type>-set_Output.xlsx’. Then, I gather all the files into 1 file within name ‘pubmed-CancerType_Top1-10-set_Output.xlsx’.

C. Text Preprocessing

a. Tokenization

First thing I do is tokenization. Tokenization is a process of breaking down or splitting textual data into smaller and more meaningful components called tokens. Tokens are independent and minimal textual components that have some definite syntax and semantics. In this case, I use split methods to get the words of a sentence.

b. Remove Non-Alphabetical Expression

Then, to make sure that it is a word, I use regular expressions. Regular expressions or regexes are specific patterns often denoted using the raw string notation. These patterns match a specific set of strings based on the rules expressed by the patterns. Regular expressions can be compiled into pattern objects and then used with a variety of methods for pattern search and substitution in strings.

In this case, to remove non-alphabetical expressions, I use the substitute methods. The substitute method will replace all the non-alphabetical expression (written in regex : [^A-Za-z]) with empty string ('').

c. Remove Stopwords

Then I remove the stopwords from the words I collected before. Stopwords are words that have little or no significance and are usually removed from text when processing it so to retain words having maximum significance and context. Stopwords usually occur most frequently if you aggregate a corpus of text based on singular tokens and checked their frequencies. Words like “a,” “the,” “and,” and so on are stopwords.

NLTK is one of python package which provide stopwords. By using it, eliminating words from stopwords would be easy. But before, I make a case conversion to lower case, so that the word will be exactly same and match case.

d. Lemmatization

The process of lemmatization is removing word affixes to get to a base form of the word. This process will not always change into the root form, but only remove the affixes until get the base word which is still present in the dictionary. These are why the process of lemmatization is slow.

Spacy is one of python package which provide lemmatization service. By loading some language styles, it will help to match with the dictionary.

D. Text Mining

The method used for mining the text is most common words. This method calculates the frequency of word in a document and arrange it in order.

RESULTS AND OBSERVATION

A. Most common 20 words (No Stopwords) in the Title (single-line)

```
[67] import pandas as pd

     file = '/content/drive/My Drive/Colab Notebooks/PubMed_Cancer_Maya/Output/Single_Line/PubMed_Top1-10_Title_NumCommonTokens_20_NoStopword.csv'

     PubMed_Title_20_NoStopword = pd.read_csv(file, delimiter=',')
     PubMed_Title_20_NoStopword
```

[67]

	Tokens	Frequency
0	cancer	58981
1	breast	9511
2	treatment	8742
3	lung	8615
4	prostate	8369
5	colorectal	8278
6	thyroid	7878
7	patients	7791
8	carcinoma	7727
9	bladder	7365
10	gastric	5252
11	liver	5213
12	diagnosis	4350
13	therapy	4349
14	cervical	4241
15	esophageal	3992
16	cell	3807
17	clinical	3755
18	screening	3522
19	stomach	3504

Observation :

- The 1st most common word is cancer, which is the main topic of the PubMed.
- In the most common 20 words above, all the cancer type words are included. Although some of types are written in different words. Those are oesophagus is written in ‘esophagus’ and cervix uteri is written in ‘cervical’ and ‘cervix’.
- The word ‘treatment’ is placed in 3rd most common words although it is not a cancer type. It means that word ‘treatment’ used in much publication’s title.
- ‘Patients’ is also placed in 8th most common words. It means that patient is one of the focus in the cancer medical publication.

B. Most common 50 words (No Stopwords) in Abstract (single-line)

```
[56] import pandas as pd

     file = '/content/drive/My Drive/Colab Notebooks/PubMed_Cancer_Maya/Output/Single_Line/PubMed_Top1-10_Abstract_NumCommonTokens_20_NoStopword.csv'

     PubMed_Abstract_20_NoStopword = pd.read_csv(file, delimiter=',')
     PubMed_Abstract_20_NoStopword
```

[56]

	Tokens	Frequency
0	cancer	40338
1	background	13118
2	patients	11520
3	study	7011
4	objective	6665
5	breast	6520
6	purpose	6325
7	carcinoma	6136
8	lung	5994
9	thyroid	5666
10	prostate	5613
11	treatment	5386
12	colorectal	5181
13	common	4654
14	gastric	4364
15	bladder	4320
16	aim	3396
17	cervical	3335
18	liver	3234
19	incidence	3117

Observation :

- The 1st most common word is cancer, which is the main topic of the PubMed.
- Word ‘background’ placed in the 2nd most common words. It is because usually abstract contains the background of writing the publication.
- Word ‘patients’ placed in the 3rd most common words in abstract. It means that patient is an important concern for medical publications.
- In the most common 20 words above, not all the cancer type words are included. Stomach and oesophagus are not included in the most common words above. It means that the word ‘stomach’ and ‘oesophagus’ are mentioned less than 3117 times in the document.

C. Comparison of the most common 50 words (No Stopwords) in Title (single-line) and Abstract (single-line)

[74]

	Title_Tokens	Title_Frequency	Abstract_Tokens	Abstract_Frequency
0	cancer	58981	cancer	40338
1	breast	9511	background	13118
2	treatment	8742	patients	11520
3	lung	8615	study	7011
4	prostate	8369	objective	6665
5	colorectal	8278	breast	6520
6	thyroid	7878	purpose	6325
7	patients	7791	carcinoma	6136
8	carcinoma	7727	lung	5994
9	bladder	7365	thyroid	5666
10	gastric	5252	prostate	5613
11	liver	5213	treatment	5386
12	diagnosis	4350	colorectal	5181
13	therapy	4349	common	4654
14	cervical	4241	gastric	4364
15	esophageal	3992	bladder	4320
16	cell	3807	aim	3396
17	clinical	3755	cervical	3335
18	screening	3522	liver	3234
19	stomach	3504	incidence	3117

Observation :

- Word ‘cancer’ is the 1st most common words in both, title and abstract.
- In the title tokens, all the cancer type words are included in the most common 20 words. But in the abstract tokens, stomach and oesophagus cancer are not included in the most common 20 words.
- Many of title included ‘treatment’ word (2nd most common word), which means many of title focused to the treatment of cancer. While in abstract, word ‘patient’ is more frequently appear.
- The words that are included in the most common 20 words in Title but not in Abstract are diagnosis, therapy, esophageal, cell, clinical, screening, and stomach.
- Besides, the words that are not included in most common 20 words in Title but included in Abstract are background, study, objective, purpose, common, aim, and incidence.
- There is a word (patients) that are not in base form.

D. Most common 20 words (No Stopwords) in Title (multi-line) and Abstract (multi-line)

```
[38] import pandas as pd

     file = '/content/drive/My Drive/Colab Notebooks/PubMed_Cancer_Maya/Output/Multi_Line/PubMed_Top1-10_Title_Abstract_NumCommonTokens_20_NoStopword.csv'

     PubMed_Title_Abstract_20_NoStopword = pd.read_csv(file, delimiter=',')
     PubMed_Title_Abstract_20_NoStopword
```

[38]

	Tokens	Frequency
0	cancer	371960
1	patients	218811
2	treatment	76510
3	survival	61037
4	breast	56685
5	results	56614
6	tumor	54674
7	prostate	54671
8	p	51908
9	study	51109
10	lung	50181
11	thyroid	49613
12	carcinoma	49487
13	bladder	46974
14	risk	45442
15	disease	44958
16	colorectal	43443
17	clinical	41157
18	cell	40293
19	cases	39900

Observation :

- Because this observation is based on the title and abstract in multi-line, so the total number of words are greater than before.
- And also, the words are more diverse than before.
- In the table above, the word ‘cancer’ still placed in the 1st most common words.
- Not all cancer types are included in the most common 20 words. The stomach, liver, oesophagus, cervix uteri, and thyroid are not included.
- There are some words that are not in the base form, which are patients, results, and cases.

E. Most common 20 words (No Stopwords and Lemmatization) in Title (multi-line) and Abstract (multi-line)

Out[13]:

	Tokens	Frequency
0	cancer	393778
1	patient	242009
2	study	82971
3	treatment	81606
4	tumor	79065
5	cell	73531
6	result	65202
7	survival	61510
8	use	60708
9	stage	58352
10	breast	56893
11	carcinoma	56567
12	prostate	54785
13	year	54162
14	case	52626
15	p	51908
16	lung	50631
17	thyroid	49687
18	disease	49097
19	risk	47969

Observation :

- Because of lemmatization, the results are different compared with before. All the words are in the basic form.
- The frequency of the tokens become greater since there is additional from another word with affixes.

F. Comparison between the most common 50 words (No Stopwords) in Title (multi-line) and Abstract (multi-line) and the most common 50 words (No Stopwords + Lemmatization) in Title (multi-line) and Abstract (multi-line)

Out[31]:

	Tokens_Before_Lemma	Frequency_Before_Lemma	Tokens_After_Lemma	Frequency_After_Lemma
0	cancer	371960	cancer	393778
1	patients	218811	patient	242009
2	treatment	76510	study	82971
3	survival	61037	treatment	81606
4	breast	56685	tumor	79065
5	results	56614	cell	73531
6	tumor	54674	result	65202
7	prostate	54671	survival	61510
8	p	51908	use	60708
9	study	51109	stage	58352
10	lung	50181	breast	56893
11	thyroid	49613	carcinoma	56567
12	carcinoma	49487	prostate	54785
13	bladder	46974	year	54162
14	risk	45442	case	52626
15	disease	44958	p	51908
16	colorectal	43443	lung	50631
17	clinical	41157	thyroid	49687
18	cell	40293	disease	49097
19	cases	39900	risk	47969

Observation :

- Because of lemmatization there are some different between before and after.
- In the most common 20 words for tokens before lemmatization (no stopwords only) there are words like bladder, colorectal, and clinical which are not included in the tokens after lemmatization.
- While there are new words like use, stage, and year appear in the most common 20 words after lemmatization.
- The number of frequent also changed. It is because, there is additional number from the same word but with affixes. For example, the word ‘patients’ in token before lemmatization. There are 218811 words appear. But in the token after lemmatization, word ‘patient’ become 242009 times appear. It is because the word ‘patient’ and ‘patients’ (and maybe another word ‘patient’ with affixes) are counted together.

G. Comparison between 60%, 70%, 80%, and 90%

In the 60% and 70%, the tokens come from a single line title or abstract. So, the total number of words are not greater than 80% or 90%. Because of multi lines, the word diversity becomes more varies. After lemmatization, the most common words become more accurate, because the tokens are in the basic form.

H. (Additional) Most common 30 words (No Stopwords) in Title (multi-line) separated with Abstract (multi-line) from each cancer type

1. Lung

Title

[38]

Lung

	Tokens	Frequency
0	lung	9964
1	cancer	9523
2	cell	2070
3	nonsmall	1546
4	patients	1315
5	treatment	1138
6	diagnosis	635
7	therapy	602
8	stage	571
9	nonsmallcell	520
10	clinical	517
11	screening	469
12	study	449
13	chemotherapy	438
14	surgical	426
15	survival	425
16	small	421
17	surgery	395
18	advanced	382
19	radiotherapy	376
20	primary	370
21	early	367
22	staging	347
23	management	307
24	risk	302
25	prognostic	299
26	pulmonary	294
27	smoking	281
28	role	279
29	resection	278
30	review	258

Abstract

[41]

Lung

	Tokens	Frequency
0	cancer	36835
1	lung	36725
2	patients	22722
3	survival	7968
4	treatment	7629
5	cell	7311
6	stage	5934
7	disease	5191
8	nsclc	5046
9	results	4788
10	study	4297
11	clinical	4211
12	chemotherapy	4071
13	tumor	3970
14	p	3895
15	diagnosis	3549
16	nonsmall	3498
17	therapy	3493
18	surgery	3401
19	cases	3299
20	risk	3259
21	years	3187
22	resection	3110
23	studies	2963
24	methods	2937
25	may	2892
26	group	2818
27	mortality	2550
28	screening	2511
29	smoking	2502
30	early	2458

Observation :

The words that may indicates as a lung cancer publication are ‘lung’, ‘pulmonary’, ‘smoking’, ‘nsclc’.

2. Breast

Title

[38]

```
==================================
Breast
==================================
```

	Tokens	Frequency
0	breast	10357
1	cancer	9687
2	treatment	1266
3	patients	959
4	women	838
5	therapy	721
6	early	594
7	screening	577
8	diagnosis	519
9	risk	510
10	clinical	470
11	management	452
12	surgery	422
13	primary	393
14	survival	378
15	study	366
16	chemotherapy	350
17	radiotherapy	330
18	factors	328
19	review	321
20	prognostic	284
21	adjuvant	274
22	new	271
23	advanced	260
24	role	258
25	surgical	251
26	prognosis	244
27	metastatic	222
28	tumor	216
29	mastectomy	216
30	detection	213

Abstract

[41]

```
==================================
Breast
==================================
```

	Tokens	Frequency
0	breast	43094
1	cancer	40489
2	patients	15506
3	women	11539
4	treatment	7945
5	risk	6110
6	survival	4811
7	disease	4711
8	therapy	4677
9	years	4450
10	clinical	4402
11	results	4328
12	study	3920
13	tumor	3854
14	screening	3793
15	diagnosis	3758
16	age	3411
17	factors	3209
18	may	3061
19	data	2939
20	chemotherapy	2917
21	surgery	2913
22	stage	2893
23	studies	2745
24	p	2731
25	cases	2707
26	early	2692
27	cancers	2509
28	primary	2456
29	mortality	2304
30	associated	2304

Observation :

The words that may indicates as a breast cancer publication are ‘breast’, ‘women’, and ‘mastectomy’.

3. Colorectal

Title

[38]

==========
Colorectal
==========

	Tokens	Frequency
0	colorectal	9648
1	cancer	9370
2	patients	1612
3	screening	1201
4	treatment	685
5	risk	561
6	study	525
7	survival	522
8	clinical	505
9	prognostic	476
10	metastatic	462
11	surgery	441
12	expression	420
13	review	357
14	resection	347
15	diagnosis	340
16	chemotherapy	336
17	management	311
18	tumor	311
19	prognosis	309
20	analysis	305
21	stage	290
22	role	284
23	therapy	274
24	hereditary	273
25	factors	270
26	liver	270
27	early	267
28	prevention	266
29	metastases	263
30	detection	254

Abstract

[41]

==========
Colorectal
==========

	Tokens	Frequency
0	cancer	36053
1	colorectal	30567
2	patients	22604
3	crc	9868
4	survival	6829
5	p	6565
6	screening	5993
7	results	5769
8	risk	5415
9	study	5249
10	treatment	5147
11	tumor	4835
12	stage	4615
13	expression	4092
14	disease	4071
15	years	3991
16	methods	3754
17	clinical	3748
18	surgery	3449
19	associated	3367
20	studies	3268
21	group	3234
22	may	3206
23	cases	3135
24	age	3023
25	colon	3002
26	data	2969
27	analysis	2954
28	resection	2944
29	factors	2885
30	cancers	2865

Observation :

The words that may indicates as a colorectal cancer publication are ‘colorectal, ‘crc’, and ‘colon’.

4. Prostate

Title

[38]

```
====================================
Prostate
====================================
```

	Tokens	Frequency
0	prostate	10129
1	cancer	9547
2	treatment	1161
3	therapy	893
4	patients	832
5	men	660
6	localized	610
7	screening	602
8	radical	525
9	prostatectomy	522
10	risk	507
11	clinical	476
12	management	442
13	diagnosis	437
14	antigen	431
15	study	429
16	radiotherapy	404
17	role	392
18	advanced	372
19	androgen	361
20	detection	351
21	metastatic	350
22	early	310
23	imaging	304
24	radiation	302
25	review	290
26	survival	288
27	biopsy	283
28	prostatespecific	280
29	prostatic	273
30	new	269

Abstract

[41]

```
========================================
Prostate
========================================
```

	Tokens	Frequency
0	prostate	43225
1	cancer	41830
2	patients	16446
3	treatment	9504
4	men	8934
5	psa	6723
6	disease	6696
7	therapy	6038
8	results	5663
9	clinical	5337
10	risk	5063
11	survival	4165
12	study	4082
13	p	3967
14	years	3747
15	may	3742
16	screening	3560
17	biopsy	3543
18	methods	3529
19	studies	3453
20	prostatectomy	3421
21	tumor	3375
22	radical	3318
23	data	3216
24	cells	3173
25	diagnosis	3055
26	stage	2994
27	antigen	2943
28	significant	2823
29	pca	2813
30	associated	2812

Observation :

The words that may indicates as a prostate cancer publication are ‘prostate’, ‘men’, ‘prostatectomy’, ‘prostatic’, and ‘pca’.

5. Stomach

Title

[38]

```
==================================
Stomach
==================================
```

	Tokens	Frequency
0	cancer	8779
1	gastric	5856
2	stomach	3974
3	patients	1216
4	early	984
5	treatment	957
6	diagnosis	790
7	carcinoma	637
8	study	567
9	surgical	526
10	surgery	479
11	chemotherapy	462
12	case	461
13	gastrectomy	449
14	advanced	442
15	clinical	438
16	survival	324
17	therapy	319
18	results	312
19	cases	309
20	resection	309
21	prognostic	299
22	risk	291
23	lymph	291
24	prognosis	279
25	endoscopic	276
26	metastasis	268
27	analysis	259
28	pylori	247
29	factors	246
30	helicobacter	243

Abstract

[41]

```
==================================
Stomach
==================================
```

	Tokens	Frequency
0	cancer	31357
1	gastric	23995
2	patients	18822
3	stomach	10490
4	survival	6833
5	treatment	4988
6	tumor	4944
7	cases	4918
8	results	4727
9	study	4266
10	p	4043
11	stage	4037
12	group	3976
13	early	3962
14	lymph	3746
15	surgery	3620
16	risk	3547
17	chemotherapy	3535
18	resection	3372
19	rate	3322
20	gastrectomy	3116
21	years	3058
22	type	2954
23	carcinoma	2824
24	advanced	2720
25	disease	2689
26	node	2669
27	metastasis	2656
28	factors	2636
29	mortality	2601
30	analysis	2391

Observation :

The words that may indicates as a stomach cancer publication are ‘gastric’, ‘stomach’, ‘gastrectomy’, ‘lymph’, ‘endoscopic’, ‘pylori’, and ‘heliobacter’.

6. Liver

Title

```
[38] ========================================
     Liver
     ========================================
```

	Tokens	Frequency
0	liver	6138
1	cancer	4940
2	carcinoma	3032
3	hepatocellular	3028
4	primary	1336
5	patients	1138
6	hepatic	1054
7	treatment	968
8	metastases	828
9	colorectal	716
10	resection	570
11	therapy	561
12	study	554
13	cells	548
14	metastasis	452
15	tumor	428
16	metastatic	425
17	hepatitis	419
18	clinical	412
19	human	379
20	diagnosis	375
21	cell	374
22	survival	339
23	chemotherapy	326
24	tumors	322
25	surgical	320
26	expression	317
27	analysis	316
28	b	315
29	case	287
30	risk	281

Abstract

```
[41] ========================================
     Liver
     ========================================
```

	Tokens	Frequency
0	liver	28504
1	patients	20738
2	cancer	20257
3	hcc	16136
4	tumor	8176
5	survival	7477
6	carcinoma	6915
7	treatment	6890
8	hepatocellular	6869
9	cells	5981
10	p	5625
11	results	5387
12	hepatic	5011
13	resection	4813
14	study	4759
15	cell	4550
16	expression	4491
17	group	4298
18	primary	4247
19	metastases	3722
20	cases	3539
21	methods	3273
22	disease	3246
23	tumors	3244
24	risk	2987
25	therapy	2939
26	may	2914
27	clinical	2877
28	b	2731
29	stage	2677
30	analysis	2676

Observation :

The words that may indicates as a liver cancer publication are ‘liver’, ‘hepatocellular’, ‘hepatic’, ‘hepatitis’, and ‘hcc’.

7. Oesophagus

Title

```
[38] ========================================
     Oesophagus
     ========================================
```

	Tokens	Frequency
0	cancer	4960
1	esophageal	4831
2	esophagus	3791
3	carcinoma	2083
4	treatment	1157
5	cell	949
6	patients	933
7	squamous	811
8	barretts	772
9	adenocarcinoma	721
10	therapy	712
11	surgical	570
12	endoscopic	551
13	case	550
14	surgery	494
15	resection	450
16	oesophageal	443
17	thoracic	431
18	study	431
19	diagnosis	406
20	early	400
21	esophagectomy	395
22	report	373
23	clinical	360
24	radiotherapy	345
25	oesophagus	344
26	chemotherapy	320
27	cases	298
28	management	290
29	preoperative	285
30	neoadjuvant	280

Abstract

```
[41] ========================================
     Oesophagus
     ========================================
```

	Tokens	Frequency
0	patients	26247
1	esophageal	18378
2	cancer	17120
3	esophagus	11445
4	survival	7486
5	carcinoma	7480
6	treatment	6750
7	p	6578
8	tumor	6401
9	cell	5315
10	results	5145
11	adenocarcinoma	4908
12	surgery	4452
13	therapy	4258
14	squamous	4240
15	cases	4229
16	study	4155
17	barretts	4065
18	group	4004
19	resection	3967
20	endoscopic	3906
21	lymph	3493
22	esophagectomy	3460
23	disease	3361
24	chemotherapy	3323
25	stage	3258
26	risk	3144
27	methods	2979
28	rate	2927
29	surgical	2844
30	years	2814

Observation :

The words that may indicates as an oesophagus cancer publication are ‘esophageal’, ‘esophagus’, ‘squamous’, ‘barretts’, ‘adenocarcinoma’, ‘thoracic’, ‘esophagectomy’, and ‘oesophagus’.

8. Cervix uteri

Title

[38] Cervix Uteri

	Tokens	Frequency
0	cancer	5076
1	cervical	4888
2	cervix	3854
3	carcinoma	1768
4	uterine	1524
5	uteri	1207
6	treatment	1125
7	screening	850
8	human	764
9	patients	627
10	diagnosis	545
11	women	543
12	papillomavirus	528
13	study	467
14	therapy	456
15	early	451
16	results	427
17	stage	420
18	radiotherapy	400
19	hpv	397
20	cytology	393
21	clinical	383
22	neoplasia	372
23	intraepithelial	339
24	detection	335
25	invasive	330
26	radical	328
27	management	318
28	cell	299
29	cases	260
30	lesions	260

Abstract

[41] Cervix Uteri

	Tokens	Frequency
0	cervical	24545
1	cancer	17582
2	patients	13783
3	hpv	9565
4	women	9440
5	cervix	6331
6	results	5999
7	cases	5906
8	carcinoma	5710
9	treatment	5504
10	screening	5442
11	study	4976
12	stage	4661
13	cells	4368
14	cin	4242
15	p	4188
16	tumor	3718
17	years	3618
18	lesions	3592
19	cell	3570
20	disease	3170
21	cytology	3101
22	methods	3082
23	squamous	3066
24	expression	3055
25	human	3016
26	uterine	2998
27	group	2977
28	survival	2873
29	normal	2858
30	clinical	2852

Observation :

The words that may indicates as a cervical uteri cancer publication are ‘cervical’, ‘cervix’, ‘uterine’, ‘uteri’, ‘women’, ‘papillomavirus’, ‘hpv’, ‘cytology’, ‘neoplasia’, and ‘intrapithelial’.

9. Thyroid

Title

[38]

```
======================================
Thyroid
======================================
            Tokens  Frequency
0          thyroid      10229
1           cancer       6418
2        papillary       2209
3        carcinoma       2092
4         patients       1385
5   differentiated       1346
6        treatment        791
7       management        515
8            study        514
9         clinical        513
10            risk        512
11         therapy        469
12           lymph        455
13       diagnosis        442
14      metastasis        411
15         nodules        407
16            node        391
17            case        372
18          iodine        362
19      follicular        354
20          review        346
21   thyroglobulin        332
22            cell        331
23        surgical        329
24          report        314
25       medullary        311
26     radioiodine        310
27         disease        308
28      metastases        305
29            neck        297
30        analysis        291
```

Abstract

[41]

```
======================================
Thyroid
======================================
            Tokens  Frequency
0          thyroid      39042
1         patients      26301
2           cancer      23414
3              ptc       7308
4        papillary       7136
5        carcinoma       7032
6                p       6304
7          results       6147
8            study       5850
9          disease       5598
10       treatment       5595
11           tumor       5009
12           cases       4994
13            risk       4815
14           years       4636
15           lymph       4501
16   thyroidectomy       4048
17         methods       3870
18        clinical       3804
19           total       3781
20  differentiated       3722
21         nodules       3710
22      metastases       3705
23           group       3668
24       diagnosis       3664
25            node       3538
26         surgery       3499
27             age       3414
28      metastasis       3398
29      expression       3327
30         therapy       3299
```

Observation :

The words that may indicates as a thyroid cancer publication are ‘thyroid’, ‘papillary’, ‘lymph’, ‘iodine’, ‘follicular’, ‘thyroglobulin’, ‘radioiodine’, ‘neck’, and ‘thyroidectomy’.

10. Bladder

Title

[38]

Bladder

	Tokens	Frequency
0	bladder	9789
1	cancer	8494
2	patients	1334
3	cystectomy	942
4	urinary	922
5	treatment	889
6	invasive	824
7	carcinoma	768
8	radical	702
9	chemotherapy	588
10	therapy	581
11	cell	544
12	study	533
13	superficial	516
14	clinical	510
15	risk	473
16	urothelial	440
17	tumor	430
18	muscleinvasive	422
19	diagnosis	408
20	expression	401
21	prognostic	388
22	cells	376
23	human	364
24	management	356
25	survival	354
26	analysis	348
27	intravesical	345
28	recurrence	307
29	radiotherapy	299
30	detection	297

Abstract

[41]

Bladder

	Tokens	Frequency
0	bladder	36177
1	cancer	30230
2	patients	24291
3	tumor	7767
4	p	7420
5	treatment	6421
6	results	6357
7	survival	6330
8	cystectomy	5329
9	cell	5145
10	study	4720
11	expression	4665
12	risk	4649
13	disease	4622
14	cells	4466
15	tumors	4259
16	clinical	4208
17	recurrence	4056
18	chemotherapy	4024
19	urinary	3792
20	stage	3792
21	methods	3773
22	carcinoma	3743
23	invasive	3667
24	therapy	3437
25	cases	3336
26	radical	3300
27	analysis	3299
28	associated	3195
29	may	3093
30	using	3051

Observation :

The words that may indicates as a bladder cancer publication are ‘bladder’, ‘cystectomy’, ‘urinary’, ‘urothelial’, and ‘intravesical’.

CONCLUSION

Based on the observation, text mining can be used to process much data from PubMed. There are different results comes from different methods of text proprocessing. Based on the observation, use tokenization, regex, stopwords removal, and lemmatization provides the best result. Using most common words as the comparison is quite effective.

In the future, text classification, text clustering, and text summarization need to be conducted using the same data. So that it will make things easier for information retrieval.

LINK

The full code of program can be found in
https://github.com/kholishotula/TextMining_HomeWork1

The Youtube link can be found in https://youtu.be/5gAxsacvWVM

REFERENCE LIST

PubMed. (2020, October 27). Retrieved from PubMed: http://pubmed.ncbi.nlm.nih.gov

Sarkar, D. (2019). *Text Analytics with Python : A Practitioner's Guide to Natural Language.* Bangalore: Apress.