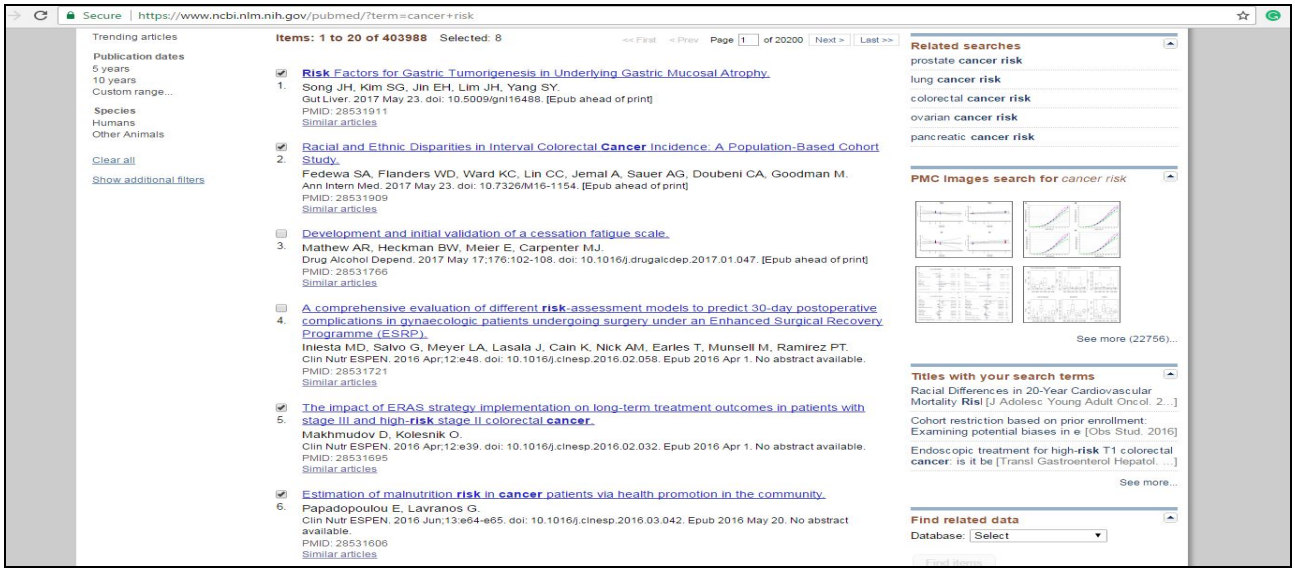
A tutorial outline on PubMed.MineR package

First I downloaded several data regarding Cancer risk from https://www.ncbi.nlm.nih.gov .



Then I opened the Rstudio and installed the pubmed.mineR package.
After installation, I loaded the pubmed.mineR library.

```
# loading pubmed.mineR package
library(pubmed.mineR)
```

And the output :

```
> library(pubmed.mineR)
>
```

At first, I need to read the abstracts in R:

```
# To read abstracts
readabs("cancer_text.txt")
```

The output :

```
> readabs("cancer_text.txt")
An object of class "Abstracts"
Slot "Journal":
[1] "1. Gut Liver. 2017 May 23. doi: 10.5009/gnl16488. [Epub ahead of print]"
[2] "2. Ann Intern Med. 2017 May 23. doi: 10.7326/M16-1154. [Epub ahead of print]"
[3] "3. Clin Nutr ESPEN. 2016 Apr;12:e39. doi: 10.1016/j.clnesp.2016.02.032. Epub 2016"
[4] "4. Clin Nutr ESPEN. 2016 Jun;13:e64-e65. doi: 10.1016/j.clnesp.2016.03.042. Epub"
[5] "5. Int J Qual Health Care. 2017 May 20:1-9. doi: 10.1093/intqhc/mzx057. [Epub ahead "
[6] "6. Clin Infect Dis. 2017 May 20. doi: 10.1093/cid/cix475. [Epub ahead of print]"
[7] "7. Carcinogenesis. 2017 May 20. doi: 10.1093/carcin/bgx046. [Epub ahead of print]"
[8] "8. J Clin Oncol. 2017 May 22:JCO2016716902. doi: 10.1200/JCO.2016.71.6902. [Epub"

Slot "Abstract":
[1] " Risk Factors for Gastric Tumorigenesis in Underlying Gastric Mucosal Atrophy.  Song JH(1), Kim SG(2), Jin EH(1), Lim JH(1)
, Yang SY(1).  Author information:  (1)Department of Internal Medicine, Healthcare Research Institute, Seoul National University
 Hospital Healthcare System Gangnam Center, Seoul, Korea. (2)Department of Internal Medicine and Liver Research Institute, Seoul
 National University College of Medicine, Seoul, Korea.  Background/Aims: Atrophic gastritis is considered a premalignant lesion
. We aimed to evaluate the risk factors for gastric tumorigenesis in underlying mucosal atrophy. Methods: A total of 10,185 subj
ects who underwent upper gastrointestinal endoscopy between 2003 and 2004 were enrolled in this retrospective cohort study. Foll
ow-up endoscopy was performed between 2005 and 2014. Atrophic gastritis and intestinal metaplasia were assessed by endoscopy usi
ng the Kimura-Takemoto classification. Helicobacter pylori infection was evaluated based on serum immuno... <truncated>
[2] " Racial and Ethnic Disparities in Interval Colorectal Cancer Incidence: A Population-Based Cohort Study.  Fedewa SA(1), Fla
nders WD(1), Ward KC(1), Lin CC(1), Jemal A(1), Sauer AG(1), Doubeni CA(1), Goodman M(1).  Author information:  (1)From Surveill
ance and Health Services Research, American Cancer Society, and Emory University, Atlanta, Georgia, and Perelman School of Medic
ine, University of Pennsylvania, Philadelphia, Pennsylvania.  Background: Interval colorectal cancer (CRC) accounts for 3% to 8%
 of all cases of CRC in the United States. Data on interval CRC by race/ethnicity are scant. Objective: To examine whether risk
for interval CRC among Medicare patients differs by race/ethnicity and whether this potential variation is accounted for by diff
erences in the quality of colonoscopy, as measured by physicians' polyp detection rate (PDR). Design: Population-based cohort st
udy. Setting: Medicare program. Participants: Patients aged 66 to 75 years who received colonoscopy betw... <truncated>
[3] "Apr 1.  The impact of ERAS strategy implementation on long-term treatment outcomes in patients with stage III and high-risk
 stage II colorectal cancer.  Makhmudov D(1), Kolesnik O(1).  Author information:  (1)Abdominal Cavity and Retroperitoneal Tumor
s, National Cancer Institute of Ukraine, Kiev, Ukraine.  DOI: 10.1016/j.clnesp.2016.02.032 "

[4] "2016 May 20.  Estimation of malnutrition risk in cancer patients via health promotion in the community.  Papadopoulou E(1),
 Lavranos G(1).  Author information:  (1)European University, Nicosia, Cyprus.  DOI: 10.1016/j.clnesp.2016.03.042 "
```

To read the abstracts in a variable:

```
# To read the abstracts in a variable
cancerabs <- readabs("cancer_text.txt")
diabetesabs <- readabs("diabetes_text.txt")
```

The output :

```
> cancerabs <- readabs("cancer_text.txt")
> diabetesabs <- readabs("diabetes_text.txt")
>
```

There were many abstracts on the downloaded file. I needed some abstracts for a specific year. For this:

```
# To retrive the abstracts for a year with the specific terms
currentabs_fn("2017", "cancer", cancerabs)
```

The output:

```
> currentabs_fn("2017", "cancer", cancerabs)
[1] "5 abstracts cancer"
An object of class "Abstracts"
Slot "Journal":
[1] "2. Ann Intern Med. 2017 May 23. doi: 10.7326/M16-1154. [Epub ahead of print]"
[2] "5. Int J Qual Health Care. 2017 May 20:1-9. doi: 10.1093/intqhc/mzx057. [Epub ahead "
[3] "6. Clin Infect Dis. 2017 May 20. doi: 10.1093/cid/cix475. [Epub ahead of print]"
[4] "7. Carcinogenesis. 2017 May 20. doi: 10.1093/carcin/bgx046. [Epub ahead of print]"
[5] "8. J Clin Oncol. 2017 May 22:JCO2016716902. doi: 10.1200/JCO.2016.71.6902. [Epub"

Slot "Abstract":
[1] " Racial and Ethnic Disparities in Interval Colorectal Cancer Incidence: A Population-Based Cohort Study.  Fedewa SA(1)
, Flanders WD(1), Ward KC(1), Lin CC(1), Jemal A(1), Sauer AG(1), Doubeni CA(1), Goodman M(1).  Author information:  (1)Fro
m Surveillance and Health Services Research, American Cancer Society, and Emory University, Atlanta, Georgia, and Perelman
School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania.  Background: Interval colorectal cancer (CRC) a
ccounts for 3% to 8% of all cases of CRC in the United States. Data on interval CRC by race/ethnicity are scant. Objective:
 To examine whether risk for interval CRC among Medicare patients differs by race/ethnicity and whether this potential vari
ation is accounted for by differences in the quality of colonoscopy, as measured by physicians' polyp detection rate (PDR).
 Design: Population-based cohort study. Setting: Medicare program. Participants: Patients aged 66 to 75 years who received
colonoscopy betw... <truncated>
[2] "of print]  A diabetes pay-for-performance program and the competing causes of death among cancer survivors with type 2
 diabetes in Taiwan.  Hsieh HM(1,)(2,)(3), Chiu HC(4,)(5), Lin YT(6), Shin SJ(7,)(8).  Author information:  (1)Department o
f Public Health, Kaohsiung Medical University, 100 Shih-Chuan 1st  Road, Kaohsiung 80708, Taiwan. (2)Department of Medical
Research, Kaohsiung Medical University Hospital, 100 Shih-Chuan 1st Road, Kaohsiung 80708, Taiwan. (3)Department of Communi
ty Medicine, Kaohsiung Medical University Hospital, 100 Shih-Chuan 1st Road, Kaohsiung 80708, Taiwan. (4)Research Education
 and Epidemiology Center, Changhua Christian Hospital, 135 Nan-Hsiao St., Changhua City 50006, Taiwan. (5)Department of Hea
lthcare Administration and Medical Informatics, Kaohsiung Medical University, 100 Shih-Chuan 1st Road, Kaohsiung 80708, Tai
wan. (6)Division of Family Medicine, Kaohsiung Medical University Hospital, 100 Shih-Chuan 1st Road, Kaohsiung 80708, Taiwa
n. (7)Graduate I... <truncated>
[3] " Age of acquiring causal human papillomavirus (HPV) infections: Leveraging simulation models to explore the natural hi
story of HPV-induced cervical cancer.  Burger EA(1,)(2), Kim JJ(1), Sy S(1), Castle PE(3,)(4).  Author information:  (1)Har
```

I need the word which mentioned most of the time in the abstracts. So, I used the *word_atomizations* **code:**

```
# To automize the words
word_atomizations(cancerabs)
```

The output :

```
> word_atomizations(cancerabs)
                words Freq
195            cancer   31
641              were   21
532              risk   19
638               was   16
41                  1   14
233               crc   14
465          patients   12
582         survivors   11
14                (hr   10
477           persons   10
58                  2    9
145               age    9
372          interval    9
422         mortality    9
651             years    9
159             among    8
248          diabetes    8
342               hpv    8
402              math    8
```

Then I felt that it will be better to find abstracts with gene:

```
# To automize with gene
gene_atomization(cancerabs)
```

The output :

```
> gene_atomization(cancerabs)
     Gene_symbol    Genes                                                Freq
[1,] "ERAS"      "ES cell expressed Ras"                                "1"
[2,] "HR"        "hair growth associated"                               "1"
[3,] "IRF5"      "interferon regulatory factor 5"                       "1"
[4,] "KRAS"      "v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog" "1"
[5,] "TP53"      "tumor protein p53"                                    "1"
>
```

To know more about the downloaded abstracts, I used some codes to find out frequency of any specific term mentioned in the abstracts:

```
# To findout a given term in each abstract
tc = c("risk", "survivor", "patients")
tdm_for_lsa(cancerabs, tc)
```

The output:

```
> tdm_for_lsa(cancerabs, tc)
         [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
risk        3    4    1    1    3    1    1   10
survivor    0    0    0    0    9    0    0    3
patients    2    1    1    1    1    0    5    0
>
```

I already got the list of genes which mentioned in the abstracts. Now, I thought that I should know about those lines which contained these genes. Then, I used codes like this:

```
# Getting sentences for a genes (ERAS)
get_gene_sentences("ERAS", cancerabs, "ERAS_sentence.txt")
```

The output :

```
Untitled1*   ERAS_sentence.txttb.txt
1  >> ERAS
2  28531695
3  Apr 1.  The impact of ERAS strategy implementation on long-term treatment outcomes in patients
   with stage III and high-risk stage II colorectal cancer.  Makhmudov D(1), Kolesnik O(1).  Author
   information:  (1)Abdominal Cavity and Retroperitoneal Tumors, National Cancer Institute of Ukraine
   , Kiev, Ukraine.  DOI: 10.1016/j.clnesp.2016.02.032
4  >> ERAS
5  28531695
6  Apr 1.  The impact of ERAS strategy implementation on long-term treatment outcomes in patients
   with stage III and high-risk stage II colorectal cancer.  Makhmudov D(1), Kolesnik O(1).  Author
   information:  (1)Abdominal Cavity and Retroperitoneal Tumors, National Cancer Institute of Ukraine
   , Kiev, Ukraine.  DOI: 10.1016/j.clnesp.2016.02.032
7
```

As above, I also got some sentences with a specific term like cancer, risk, patient, and survivor:

```
# Getting sentences for a query term (cancer)
contextSearch(cancerabs, "cancer")
```

The output :

```
Untitled1* ×   cancer.tex ×
Format ▾   Compile PDF
No TeX installation detected. Please install TeX before compiling.
1  \documentclass[11pt]{article}
2  \title { cancer search}
3  \begin{document}
4  \maketitle
5  %\section{}
6  %\subsection{}
7  Results of context search
8  \newline [ 1 ] Racial and Ethnic Disparities in Interval Colorectal Cancer Incidence: A Population
   -Based Cohort Study. Fedewa SA(1), Flanders WD(1), Ward KC(1), Lin CC(1), Jemal A(1), Sauer AG(1),
   Doubeni CA(1), Goodman M(1). Author information: (1)From Surveillance and Health Services Research,
   American Cancer Society, and Emory University, Atlanta, Georgia, and Perelman School of Medicine,
   University of Pennsylvania, Philadelphia, Pennsylvania. Background: Interval colorectal \textbf {
   cancer } (CRC) accounts for 3% to 8% of all cases of CRC in the United States.
9  \newline [ 2 ] Compared with white persons, black persons had significantly higher risk for interval
   CRC (HR, 1.31 [95% CI, 1.13 to 1.51]); the disparity was more pronounced for \textbf { cancer } of
   the rectum (HR, 1.70 [CI, 1.25 to 2.31]) and distal colon (HR, 1.45 [CI, 1.00 to 2.11]) than for cancer
   of the proximal colon (HR, 1.17 [CI, 0.96 to 1.42]).
10 \newline [ 3 ] Conclusion: Among elderly Medicare enrollees, the risk for interval CRC was higher in
   black persons than in white persons; the difference was more pronounced for \textbf { cancer } of
   the distal colon and rectum and for physicians with higher PDRs.
11 \newline [ 4 ] Apr 1. The impact of ERAS strategy implementation on long-term treatment outcomes in
   patients with stage III and high-risk stage II colorectal \textbf { cancer } . Makhmudov D(1),
   Kolesnik O(1). Author information: (1)Abdominal Cavity and Retroperitoneal Tumors, National Cancer
   Institute of Ukraine, Kiev, Ukraine. DOI: 10.1016/j.clnesp.2016.02.032
12 \newline [ 5 ] 2016 May 20. Estimation of malnutrition risk in \textbf { cancer } patients via
   health promotion in the community. Papadopoulou E(1), Lavranos G(1). Author information: (1)European
   University, Nicosia, Cyprus. DOI: 10.1016/j.clnesp.2016.03.042
13 \newline [ 6 ] of print] A diabetes pay-for-performance program and the competing causes of death
   among \textbf { cancer } survivors with type 2 diabetes in Taiwan. Hsieh HM(1,)(2,)(3), Chiu HC(4
   ,)(5), Lin YT(6), Shin SJ(7,)(8). Author information: (1)Department of Public Health, Kaohsiung Medical
   University, 100 Shih-Chuan 1st Road, Kaohsiung 80708, Taiwan. (2)Department of Medical Research,
   Kaohsiung Medical University Hospital, 100 Shih-Chuan 1st Road, Kaohsiung 80708, Taiwan. (3)Department of
   Community Medicine, Kaohsiung Medical University Hospital, 100 Shih-Chuan 1st Road, Kaohsiung 80708,
   Taiwan. (4)Research Education and Epidemiology Center, Changhua Christian Hospital, 135 Nan-Hsiao St.,
   Changhua City 50006, Taiwan. (5)Department of Healthcare Administration and Medical Informatics,
1:1                                                                                              TeX
```

The downloaded file contained many abstracts of several years. To know about their year of publication:

```
# Searching abstracts yearwise
Yearwise(cancerabs, c("2016", "2017"))
```

The output :

```
> Yearwise(cancerabs, c("2016", "2017"))
[1] "3 abstracts 2016" "3 abstracts 2017"
```

I also found the number of abstracts for a specific gene:

```
# Searching abstracts genewise
Genewise(cancerabs, "ERAS")
```

The output :

```
> Genewise(cancerabs, "ERAS")
[1] "1 abstracts ERAS"
```

Some of the abstracts were short and some were very long. So, it was time consuming and hard to learn about their information such as diseases, chemical, genes etc. To get a descriptive result on it, I used the PMID number (publication number) **with the code named** *pubtator_fucntion*:

```
# Getting information of PMID about Gene, Chemical, Mutation, Species and Diseases
pubtator_function(28531909)
```

The output :

```
> pubtator_function(28531909)
$Genes
NULL

$Diseases
[1] "Colorectal Cancer"                    "colorectal cancer"
[3] "CRC"                                  "cancer of the rectum"
[5] "cancer"                               "cancer of the distal colon and rectum"
[7] "Cancer"

$Mutations
NULL

$Chemicals
NULL

$Species
[1] "patients"     "Participants" "Patients"     "person"       "persons"

$PMID
[1] 28531909

>
```

After all of these, a question came to my mind that if I get some abstracts later, then I need to add them with previous one. Then I will be able to analyze them together. For this:

```
# Combining several abstracts in a variable
combo = combineabs(diabetesabs, cancerabs)
```

The output :

```
> combo = combineabs(diabetesabs, cancerabs)
[1] "9 combined abstracts for above terms"
>
```

When I need to remove some abstracts with a specific term:

```
# Removing abstracts for a specific term
removeabs(cancerabs, "cancer", TRUE)
```

The output :

```
> removeabs(cancerabs, "cancer", TRUE)
[1] "7 abstracts removed abstracts for term cancer"
An object of class "Abstracts"
Slot "Journal":
[1] "1. Gut Liver. 2017 May 23. doi: 10.5009/gnl16488. [Epub ahead of print]"

Slot "Abstract":
[1] " Risk Factors for Gastric Tumorigenesis in Underlying Gastric Mucosal Atrophy.  Song JH(1), Kim SG(2), Jin
EH(1), Lim JH(1), Yang SY(1).  Author information:  (1)Department of Internal Medicine, Healthcare Research Inst
itute, Seoul National University Hospital Healthcare System Gangnam Center, Seoul, Korea. (2)Department of Inter
nal Medicine and Liver Research Institute, Seoul National University College of Medicine, Seoul, Korea.  Backgro
und/Aims: Atrophic gastritis is considered a premalignant lesion. We aimed to evaluate the risk factors for gast
ric tumorigenesis in underlying mucosal atrophy. Methods: A total of 10,185 subjects who underwent upper gastroi
ntestinal endoscopy between 2003 and 2004 were enrolled in this retrospective cohort study. Follow-up endoscopy
was performed between 2005 and 2014. Atrophic gastritis and intestinal metaplasia were assessed by endoscopy usi
ng the Kimura-Takemoto classification. Helicobacter pylori infection was evaluated based on serum immuno... <tru
ncated>

Slot "PMID":
[1] 28531911
```

There are many codes to analyze the abstracts with the pubmed.mineR package. But, they are mostly similar to each other like the code of *sentence_token* and *context_search* provides similar result. So, I only showed the use of anyone from them.