Lab 8. Biomedical literature data mining

1. Objective

- To understand one of text mining techniques, co-occurrence based method
- To apply the co-occurrence based method to public data to generate structured information
- To learn python programming skills
- To learn basic text mining conceptual knowledge
- To learn brief concepts of biological data

2. Background

2.1. Biomedical literature data mining (Text mining)

2.1.1. Basis of literature data mining (Text mining)

Text mining is the use of automated method which could suggest the biological relationships by extracting information from the abundant literature. Text mining could be defined as the discovery by computer of new, previously unknown information and automatically extracting information from different written resources.

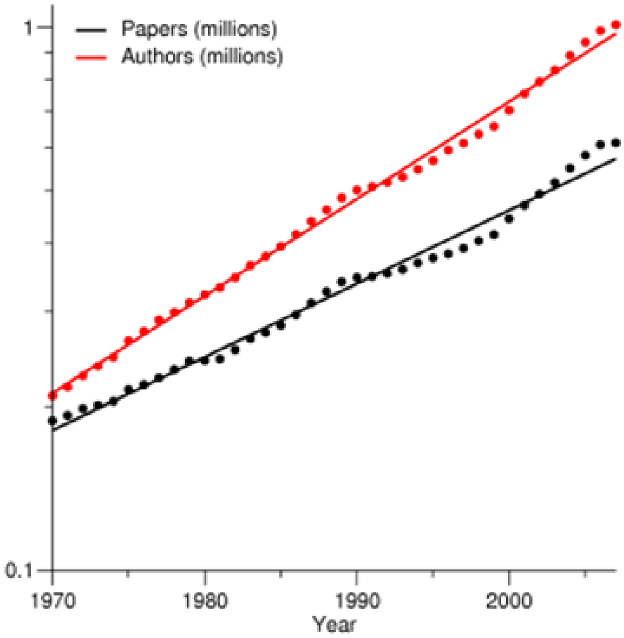


Figure 1. The number of biomedical publications per year

There are at least as many motivations for doing text mining work as there are types of bio-scientists. Model organism database curators have been heavy participants in the development of the field due to their need to process large numbers of publications in order to populate the many data fields for every gene in their species of interest. Bench scientists have built biomedical text mining applications to aid in the development of tools for interpreting the output of high-throughput assays and to improve searches of sequence databases. Bio-scientists of every stripe have built applications to deal with the two issues: dealing with the double exponential growth in the scientific literature over the past few years and searching PubMed/MEDLINE for genomics-related publications.[1]

Text mining consists of three major parts, such as information retrieval (IR), entity recognition (ER), and information extraction (IE). First, IR retrieves relevant papers which researchers are interested in. PubMed is one of the known IR systems. This system uses the Boolean model and the vector model. The Boolean model could retrieve papers which contain combinations of terms. The vector model could represent retrieved document by a frequency-based weighting scheme. Second, ER finds biological entities such as gene and protein names in literatures. Finding the entity is difficult since lack of standardization of names as well as gene and protein have several name and abbreviations. Finally, IE extracts relationships between biological entities and uses two different approaches such as co-occurrence and natural-language processing.

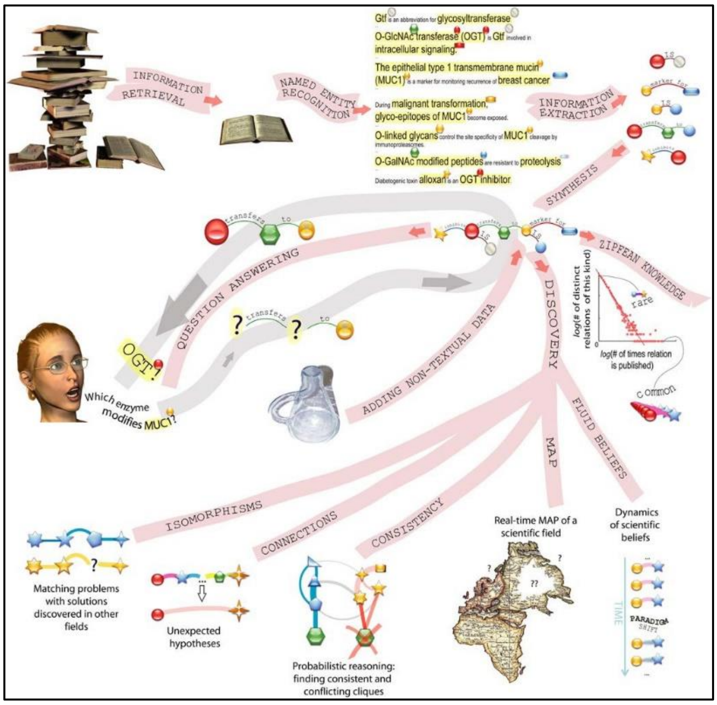
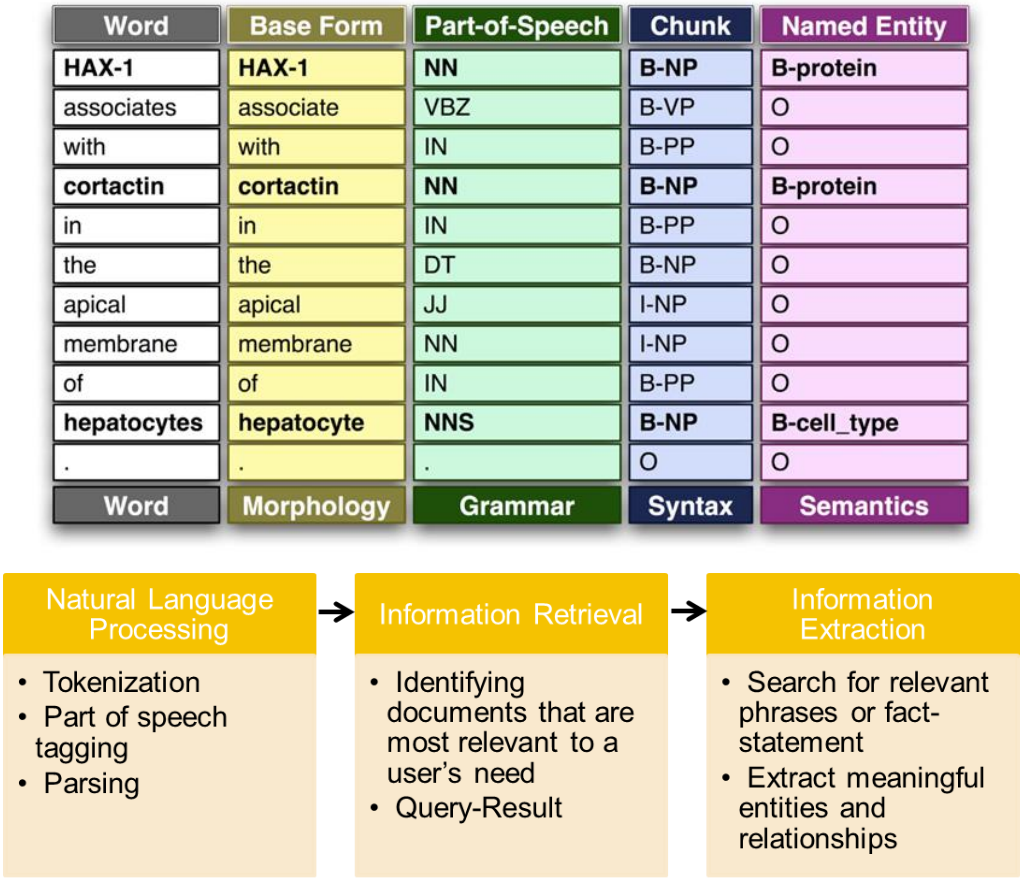

INFORMATION RETRIEVAL
NAMED ENTITY RECOGNITION
Gtf is an abbreviation for glycosyltransferase
O-GlcNAc transferase (OGT) is Gtf involved in intracellular signaling
The epithelial type 1 transmembrane mucin (MUC1) is a marker for monitoring recurrence of breast cancer
During malignant transformation, glyco-epitopes of MUC1 become exposed.
O-linked glycans control the site specificity of MUC1 cleavage by immunoproteasomes.
O-GalNAc modified peptides are resistant to proteolysis
Diabetogenic toxin alloxan is an OGT inhibitor
INFORMATION EXTRACTION
IS
marker for
IS
transfers to
inhibits
SYNTHESIS
transfers to
inhibits
transfers
to
IS
marker for
IS
ZIPFEAN KNOWLEDGE
log(# of distinct relations of this kind)
0
log(# of times relation is published)
rare
common
QUESTION ANSWERING
?
transfers
?
to
OGT?
Which enzyme modifies MUC1?
ADDING NON-TEXTUAL DATA
DISCOVERY
MAP
FLUID BELIEFS
ISOMORPHISMS
CONNECTIONS
CONSISTENCY
?
Matching problems with solutions discovered in other fields
...
Unexpected hypotheses
Probabilistic reasoning: finding consistent and conflicting cliques
Real-time MAP of a scientific field
?
?
??
Dynamics of scientific beliefs
TIME
PARADIGM SHIFT

Figure 2. Major techniques and applications of text mining

Figure 2 shows a high-level overview of the stages in text mining, with a focus on its applications. Extracted information can be further used for building systems for answering questions, fusing experimental data with literature-derived information, implementing computational creativity (discovering esoteric connections between facts, matching solutions in one field with open problems in another one, capturing cliques of internally consistent observations that are inconsistent across cliques), and analysis of large-scale dynamics of scientific fields. The ambiguity is one of direct causes which generate an error rate. However, it is the research part of literature mining to annotate genes and proteins.

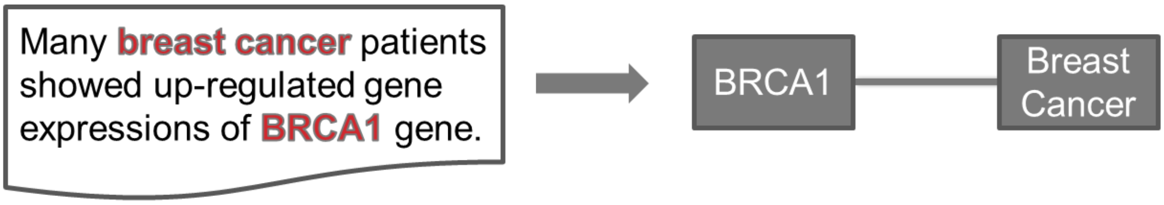


Word	Base Form	Part-of-Speech	Chunk	Named Entity
HAX-1	**HAX-1**	**NN**	**B-NP**	**B-protein**
associates	associate	VBZ	B-VP	O
with	with	IN	B-PP	O
cortactin	**cortactin**	**NN**	**B-NP**	**B-protein**
in	in	IN	B-PP	O
the	the	DT	B-NP	O
apical	apical	JJ	I-NP	O
membrane	membrane	NN	I-NP	O
of	of	IN	B-PP	O
hepatocytes	**hepatocyte**	**NNS**	**B-NP**	**B-cell_type**
.	.	.	O	O
Word	**Morphology**	**Grammar**	**Syntax**	**Semantics**

Natural Language Processing
• Tokenization
• Part of speech tagging
• Parsing
Information Retrieval
• Identifying documents that are most relevant to a user's need
• Query-Result
Information Extraction
• Search for relevant phrases or fact-statement
• Extract meaningful entities and relationships

Figure 3. literature mining process

2.1.2. Co-occurrence based method



Many breast cancer patients showed up-regulated gene expressions of BRCA1 gene.
BRCA1
Breast Cancer

Figure 4. An example of co-occurrence-based approach

Co-occurrence approach extracts entities which co-occur within same abstract or sentence. This method assumes that it is likely to relate two entities which are frequently mentioned together. Natural-language processing approach combines syntax and semantics information.[2] Co-occurrence–based methods do no more than look for concepts that occur in the same unit of text—typically a sentence, but sometimes as large as an abstract—and posit a relationship between them. For example, if such a system saw that BRCA1 and breast cancer occurred in the same sentence, it might assume a relationship between breast cancer and the BRCA1 gene. Some early biomedical text mining systems were co-occurrence–based, but such systems are highly error prone, and are not commonly built today. In fact, many text mining practitioners would not consider them to be text mining systems at all. Co-occurrence of concepts in a text is sometimes used as a simple baseline when evaluating more sophisticated systems; as such, they are nontrivial, since even a co-occurrence–based system must deal with variability in the ways that concepts are expressed in human-produced texts.

2.2. Evaluation measures

In the context of classification tasks, the terms true positives, true negatives, false positives and false negatives are used to compare the given classification of an item with the correct classification. Here is an example: Suppose there is a computer program for recognizing dogs in photographs. The program identifies 8 dogs in a picture containing 12 dogs and 10 cats. Of the 8 identified as dogs, 5 actually are dogs (true positives), while the rest are cats (false positives). The program's precision is 5/8 while its recall is 5/12. When a search engine returns 30 pages only 20 of which were relevant while failing to return 40 additional relevant pages, its precision is 20/30 = 2/3 while its recall is 20/60 = 1/3. So, in this case, precision is "how useful the search results are", and recall is "how complete the results are". This is illustrated by the table below (assuming condition positive is the class we are looking for): [4]

		Correct **result / classification**	
		Condition positive (actual dogs: 12, relevant pages: 60)	**Condition negative** (actual cats: 10, irrelevant pages: others)
Obtained **(predicted)** **result /** **classification**	**Obtained condition positive** (identified dogs: 8, returned pages: 30)	**tp, true positive** (program: 5, search: 20)	**fp, false positive** (program: 3, search: 10)
	Obtained condition negative (identified cats: 14, not returned pages: others)	**fp, false negative** (program: 7, search: 40)	**tn, true negative** (program: 7, search: others)

Precision and recall are then defined as:

$$\text{Precision} = \frac{tp}{tp + fp}, \quad \text{Recall} = \frac{tp}{tp + fn}$$

2.2.1. Recall in document retrieval case

Recall in Information Retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

For example, for text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned. It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough to evaluate the performance of text mining. One also needs to measure the rate of the number of relevant documents among the returned documents (i.e. precision).

2.2.2. Precision in document retrieval case

In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the search:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

For example for a text search on a set of documents precision is the number of correct results divided by the number of all returned results. Precision is also used with recall, the percent of all relevant documents that is returned by the search. The two measures are sometimes used together in the F1 Score (or f-measure) to provide a single measurement for a system. Note that the meaning and usage of "precision" in the field of Information Retrieval differs from the definition of accuracy and precision within other branches of science and technology.

$$F_1 = \left(\frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} \right) = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

3. Prelab activities

3.1. Prepare data

1) Download “PubMed abstracts” and “Gene dictionary” from our class board in KLMS. (http://klms.kaist.ac.kr)

   You should download abstract files by your class.

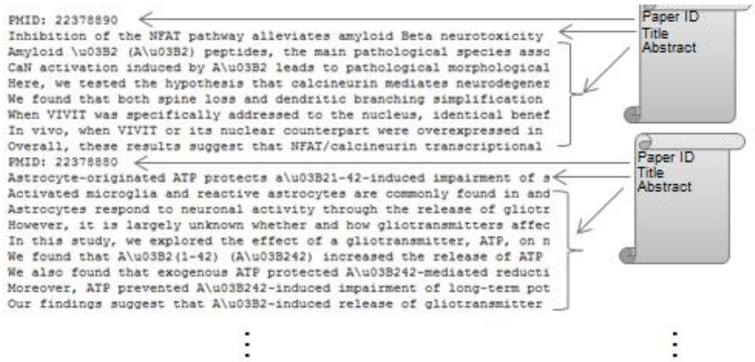
   i. Input files for making occurrence table

      1. Literature files

         A. “Disease_Class_Train_Literature.txt”

         B. “Disease_Class_Test _Literature.txt”

      One line is one sentence. The number of total abstracts: 1000

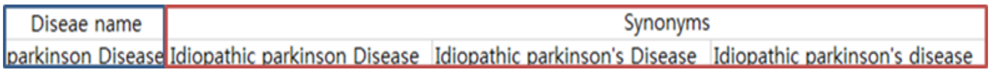


```
PMID: 22378890
Inhibition of the NFAT pathway alleviates amyloid Beta neurotoxicity
Amyloid \u03B2 (A\u03B2) peptides, the main pathological species asso
CaN activation induced by A\u03B2 leads to pathological morphological
Here, we tested the hypothesis that calcineurin mediates neurodegener
We found that both spine loss and dendritic branching simplification
When VIVIT was specifically addressed to the nucleus, identical benef
In vivo, when VIVIT or its nuclear counterpart were overexpressed in
Overall, these results suggest that NFAT/calcineurin transcriptional
PMID: 22378880
Astrocyte-originated ATP protects a\u03B21-42-induced impairment of s
Activated microglia and reactive astrocytes are commonly found in and
Astrocytes respond to neuronal activity through the release of gliotr
However, it is largely unknown whether and how gliotransmitters affec
In this study, we explored the effect of a gliotransmitter, ATP, on n
We found that A\u03B2(1-42) (A\u03B242) increased the release of ATP
We also found that exogenous ATP protected A\u03B242-mediated reducti
Moreover, ATP prevented A\u03B242-induced impairment of long-term pot
Our findings suggest that A\u03B2-induced release of gliotransmitter
```

      2. Dictionary files

         A. “Disease_Class_Train _syno_dictionary.txt”

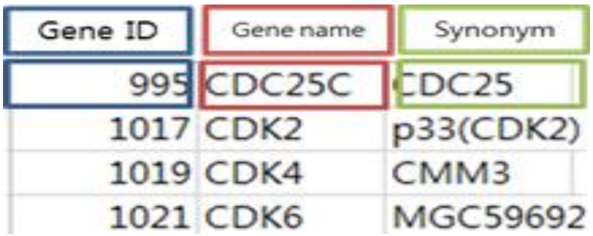
         B. “Disease_Class_Test_syno_dictionary.txt”



Diseae name	Synonyms		
parkinson Disease	Idiopathic parkinson Disease	Idiopathic parkinson's Disease	Idiopathic parkinson's disease

      3. Gene dictionary file (“Gene_dictionary_Entrez.txt”)

         A tab delimited text file.



Gene ID	Gene name	Synonym
995	CDC25C	CDC25
1017	CDK2	p33(CDK2)
1019	CDK4	CMM3
1021	CDK6	MGC59692

3.2. Preprocess uploaded data to make table

1) Goal : Generate gene and disease tagging tables for text mining. In this prelab section, you have to create gene and disease tagging tables using ‘Literature data’ and ‘dictionary data’ described above. In this gene tagging table, information about which gene exists in which literature and at which positions, will be recorded, and same with disease,

2) Write to output file

A. Generate gene-tagging and disease-tagging tables respectively

i. In other words, you have to make total 4 output files.

B. Produce attributes which is used for your algorithm for text mining

i. Informative attributes have an advantage in your F-Score.

C. Refer to supplementary file

3) Output files.

A. Gene-tagging tables for train disease and test disease.

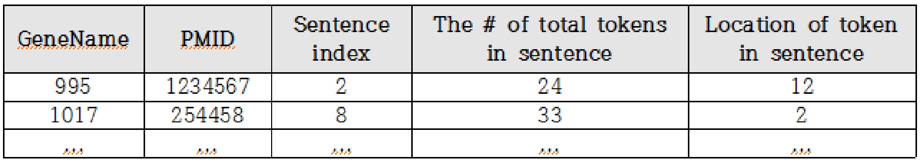
i. Gene_tagging_table_Train_literature.txt.

Train_Literature.txt and Gene_dictionary_Entrez.txt are used.

ii. Gene_tagging_table_Test_literature.txt.

Test_Literature.txt, Gene_dictionary_Entrez.txt are used.

iii. Example



GeneName	PMID	Sentence index	The # of total tokens in sentence	Location of token in sentence
995	1234567	2	24	12
1017	254458	8	33	2
...	...	...	...	...

B. Disease-tagging tables for train disease and test disease

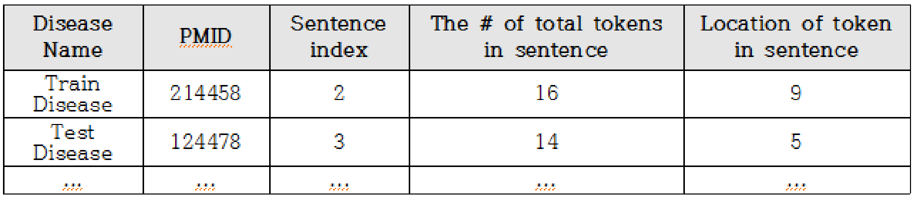
i. Disease_tagging_table_Train_literature.txt.

Train_Literature.txt, Disease_Class_Train _syno_dictionary.txt are used.

ii. Disease_tagging_table_Test_literature.txt.

Test_Literature.txt, Disease_Class_Test_syno_dictionary.txt are used.

iii. Example



Disease Name	PMID	Sentence index	The # of total tokens in sentence	Location of token in sentence
Train Disease	214458	2	16	9
Test Disease	124478	3	14	5
...	...	...	...	...

C. Description of attributes in output files

i. Gene or Disease Name : Standard gene or disease ID based on PharmGKB

ii. PMID : PMID of Abstract that include gene or disease

iii. Sentence index : Sentence order which gene or disease occurs in abstract

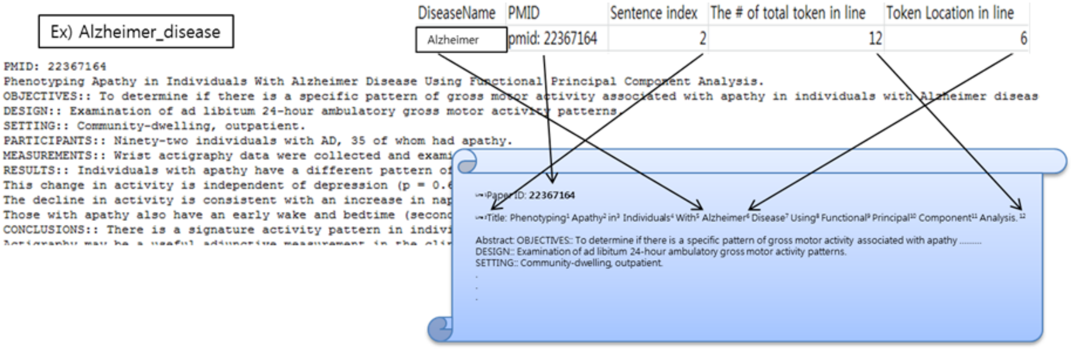
iv. The # of total tokens in sentence : The # of total tokens in gene or disease occurred sentence

v. Location of token in sentence : Position of occurred gene or disease in sentence

vi. Output files you generated will be used for your text—mining experiment. So **you**

should bring your output files in class!!

vii. Example



Ex) Alzheimer_disease

```
PMID: 22367164
Phenotyping Apathy in Individuals With Alzheimer Disease Using Functional Principal Component Analysis.
OBJECTIVES:: To determine if there is a specific pattern of gross motor activity associated with apathy in individuals with Alzheimer diseas
DESIGN:: Examination of ad libitum 24-hour ambulatory gross motor activity patterns.
SETTING:: Community-dwelling, outpatient.
PARTICIPANTS:: Ninety-two individuals with AD, 35 of whom had apathy.
MEASUREMENTS:: Wrist actigraphy data were collected and exami
RESULTS:: Individuals with apathy have a different pattern of
This change in activity is independent of depression (p = 0.6
The decline in activity is consistent with an increase in nap
Those with apathy also have an early wake and bedtime (second
CONCLUSIONS:: There is a signature activity pattern in indivi
Actigraphy may be a useful adjunctive measurement in the cli
```

DiseaseName	PMID	Sentence index	The # of total token in line	Token Location in line
Alzheimer	pmid: 22367164	2	12	6

Paper ID: 22367164

Title: Phenotyping^{1} Apathy^{2} in^{3} Individuals^{4} With^{5} Alzheimer^{6} Disease^{7} Using^{8} Functional^{9} Principal^{10} Component^{11} Analysis.^{12}

Abstract: OBJECTIVES:: To determine if there is a specific pattern of gross motor activity associated with apathy ………
DESIGN:: Examination of ad libitum 24-hour ambulatory gross motor activity patterns.
SETTING:: Community-dwelling, outpatient.
.
.
.

*You are able to add new attributes that may increase performance of your algorithm.

3.3. Devise co-occurrence based scoring function

1) Read attached references to understand co-occurrence based text-mining (getting started in text mining). Co-occurrence based methods relies on co-occurence. For example, if two words (cancer, p53) occurs in same sentence, we can assume that those two words have some kind of relationships. Reference: https://www.aabri.com/manuscripts/152265.pdf
2) List features which can be used to scoring functions (more than 5)
   A. Ex) Distance(how many words exist) between 2 words
   B. Ex) If 2 words are in same sentence
   C. …
3) Combining above features, generate your own scoring functions (more than 2)
   A. Ex) $\sum_{\text{for all co-occurrences}} \frac{a*s}{d}$ (where a=coefficient, s= if words are in same sentence, 1, else 0, d=distance between 2 words)

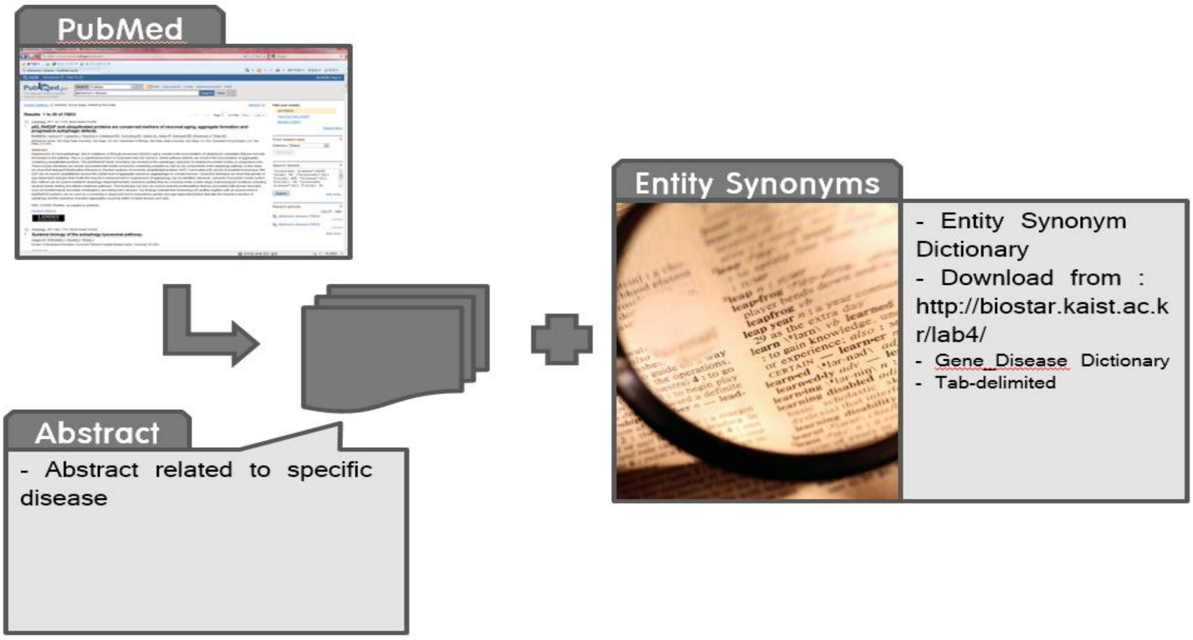
3.4. How to report

1) Basic python programming skills for text-mining
   A. Write a report including
      i. Source code with simple comments
      ii. Description of new attributes you generated
2) Devise co-occurrence-based scoring function
   A. Write a report including
      i. List of possible features with explanation (and reason) of it (more than 5)
      ii. Your own scoring functions (more than 2)
3) Bring hard copy of report and output files at the main Lab class

4. Mainlab activities

- Preprocess raw data and recognize entities in the literatures (in prelab session)
- Extract association information between diseases and genes and validate the result using F-score
- Databases : NCBI PubMed, Entity Dictionary (given)

4.1. Raw data preparation



PubMed

Abstract

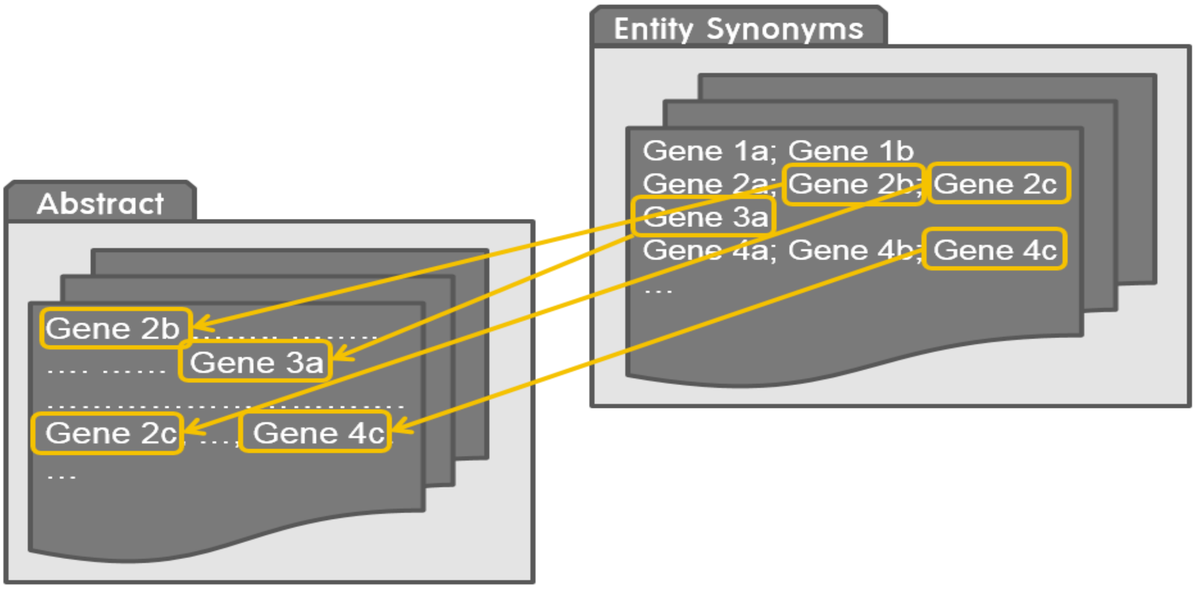
- Abstract related to specific disease

Entity Synonyms

- Entity Synonym Dictionary
- Download from : http://biostar.kaist.ac.kr/lab4/
- Gene_Disease Dictionary
- Tab-delimited

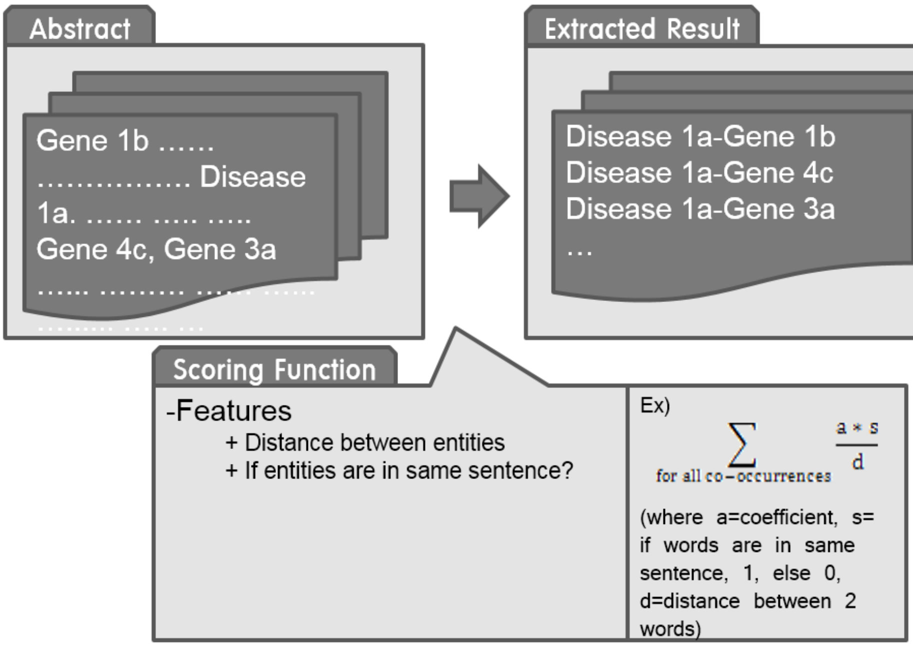
This step is prelab session. From given abstract file and given synonym dictionary, you have to parse abstracts and find for words in synonym dictionary.

4.2. Entity recognition



Also with step 1, this is prelab session. As a result of prelab activity, you'll produce a table that which entity is located on which position of which literature.

4.3. Information extraction


Abstract
Gene 1b ……
……………. Disease
1a. …… ….. …..
Gene 4c, Gene 3a
…… ……… …… ……
Extracted Result
Disease 1a-Gene 1b
Disease 1a-Gene 4c
Disease 1a-Gene 3a
…
Scoring Function
-Features
+ Distance between entities
+ If entities are in same sentence?
Ex)
$\sum_{\text{for all co-occurrences}} \frac{a*s}{d}$
(where a=coefficient, s= if words are in same sentence, 1, else 0, d=distance between 2 words)

Mainlab session starts from here. From produced files, and with your own scoring function, you have to produce a file which indicates 'co-occurance score', or 'evidence' of associations between specific disease and associatied genes. With some kind of score threshold, you'll get 'really associated' genes(gene id) with specific disease.

Final Report Question 1.

Write your source code with comments.

Final Report Question 2.

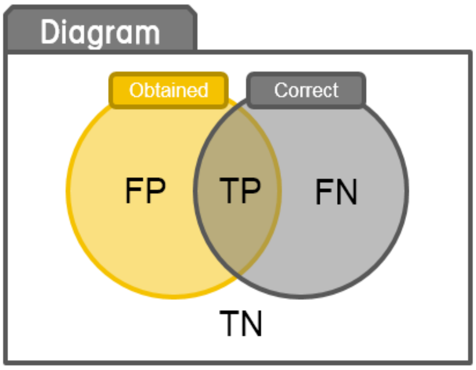
Describe your scoring function and attributes involved in it.

4.4. Validation

Classification Table

		Correct Result	
		Positive	Negative
Obtained Result	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Diagram



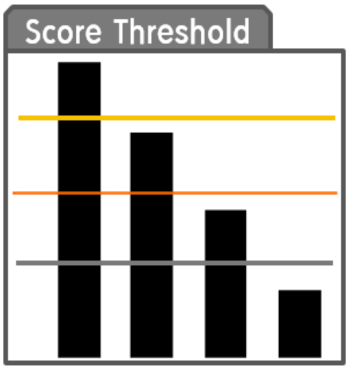
Validation Measures

$$\text{Precision}(P) = \frac{TP}{TP+FP}$$

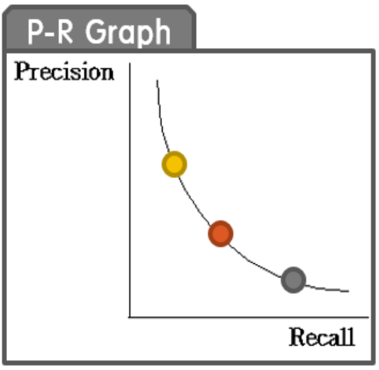
$$\text{Recall}(R) = \frac{TP}{TP+FN}$$

$$\text{F}-\text{Score} = \frac{2*P*R}{P+R}$$

Score Threshold



P-R Graph

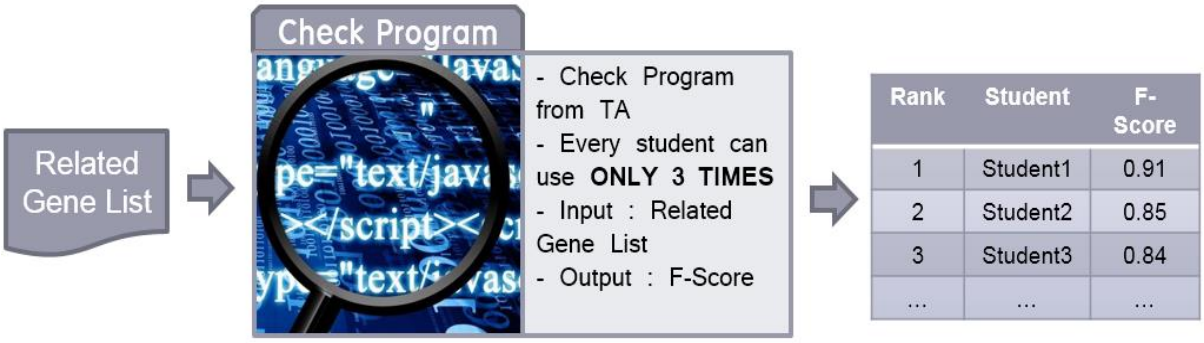


After mining those abstracts, you'll get the list of associated genes with the disease. In this case, you have to validate them. There are kinds of validation methods. Classical measures described above will be used in this lab. TAs have real 'answers' that indicates associated genes with the disease. You have to compare your result with our answer, with above metrics.

Demo Question 1.

Check your F-score for training data via web page and improve your scoring function.

4.5. Scoring results



Related Gene List

Check Program

- Check Program from TA
- Every student can use **ONLY 3 TIMES**
- Input : Related Gene List
- Output : F-Score

Rank	Student	F-Score
1	Student1	0.91
2	Student2	0.85
3	Student3	0.84
…	…	…

As a result, you'll get final score for your answers based on real answer. Those scores will be your mainlab score.

Demo Question 2.

Check your F-score for test data via web page. You have only 3 chances.

Final Report Question 3.

Write your output genes and F-score with screenshots of evaluation pages. What is the difference between two datasets?

5. Reference

[1] Rzhetsky, A., et al., Getting started in text mining: part two. PLoS computational biology, 2009. 5(7): p. e1000411.

[2] Cohen, K. and L. Hunter, Getting started in text mining. PLoS Comput Biol, 2008. 4(1): p. e20

[3] Hagit Shatkay, Ronen Feldman. Mining the Biomedical literature in the Genomic Era: An Overview. Journal of computational biology, 2003. 10(6):p821-855

[4] Wikipedia: Precision and recall, https://en.wikipedia.org/wiki/Precision_and_recall