Chapter11: Text Analysis

Asst. Prof. Dr. Supakit Nootyaskool

supakit@it.kmitl.ac.th

Learning Outcome

- Understand the process step in text analysis.
- Describe preprocessing data before text analysis.
- Have experience the usage of R getting data from Twitter.

Topics

- Text analysis overview
- Text data
- Research and work related to text analysis
- Text analysis steps
- Get data from Twitter

Text analysis overview

Text analysis

- Means

> "The discovery by computer of new, previously unknown information, by automatically extracting information from different written resources" --Marti Hearst

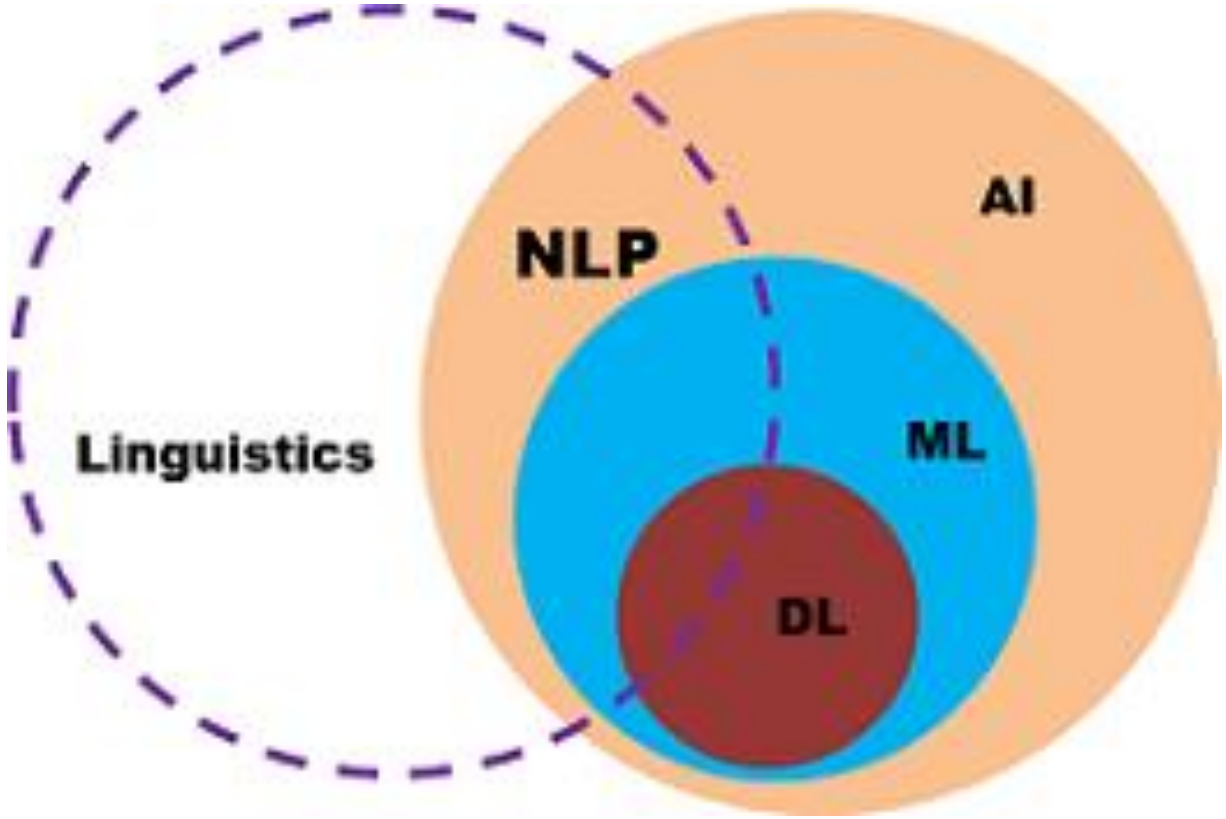
- Text analysis uses a set of linguistic ภาษาศาสตร์, statistical, and machine learning techniques to get information from text resources.

Text analysis

- Called text analytics, text data mining, and text mining.
- Text analysis is the process of delivering high-quality information from text.

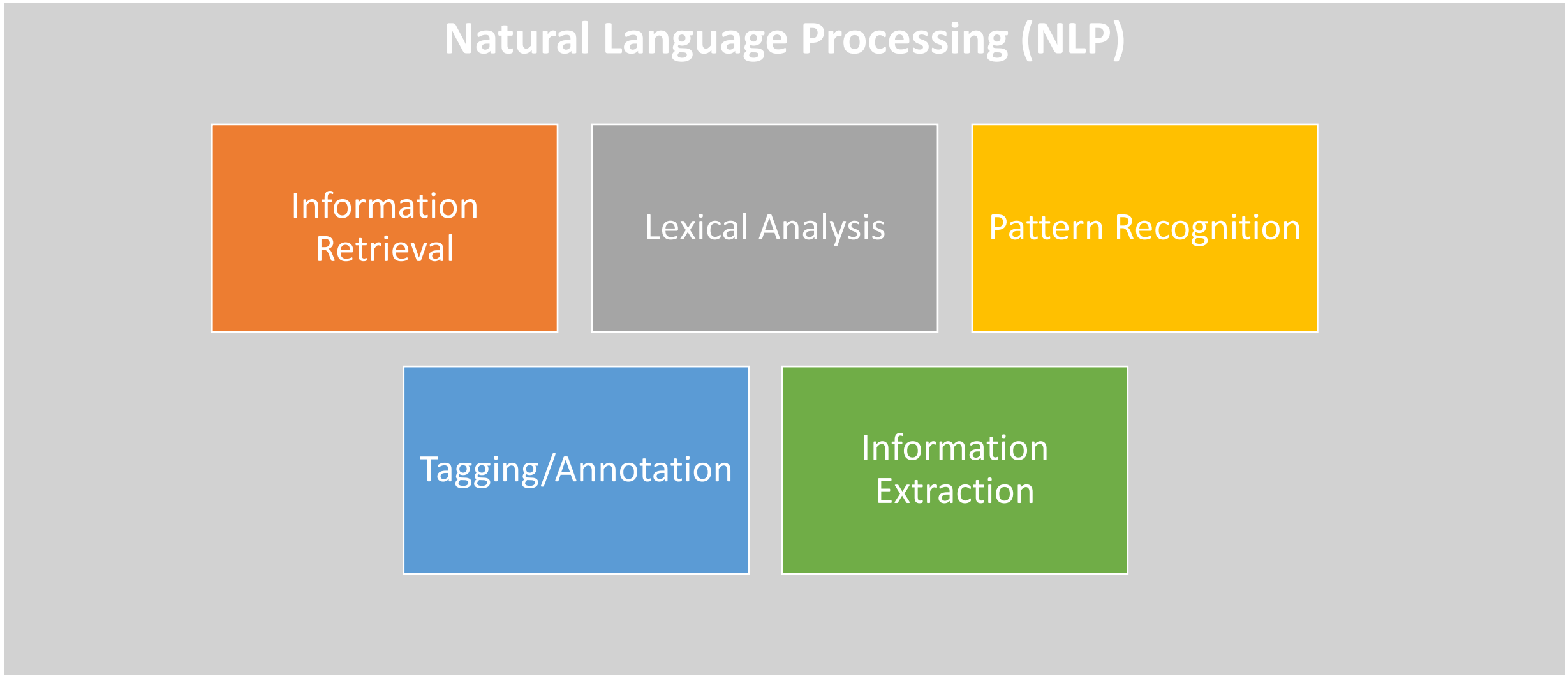
- Output from text analysis supports for
  - Business intelligent,
  - Exploratory data analysis,
  - Research or investigation.

Text Analysis Relates to NLP and Linguistics

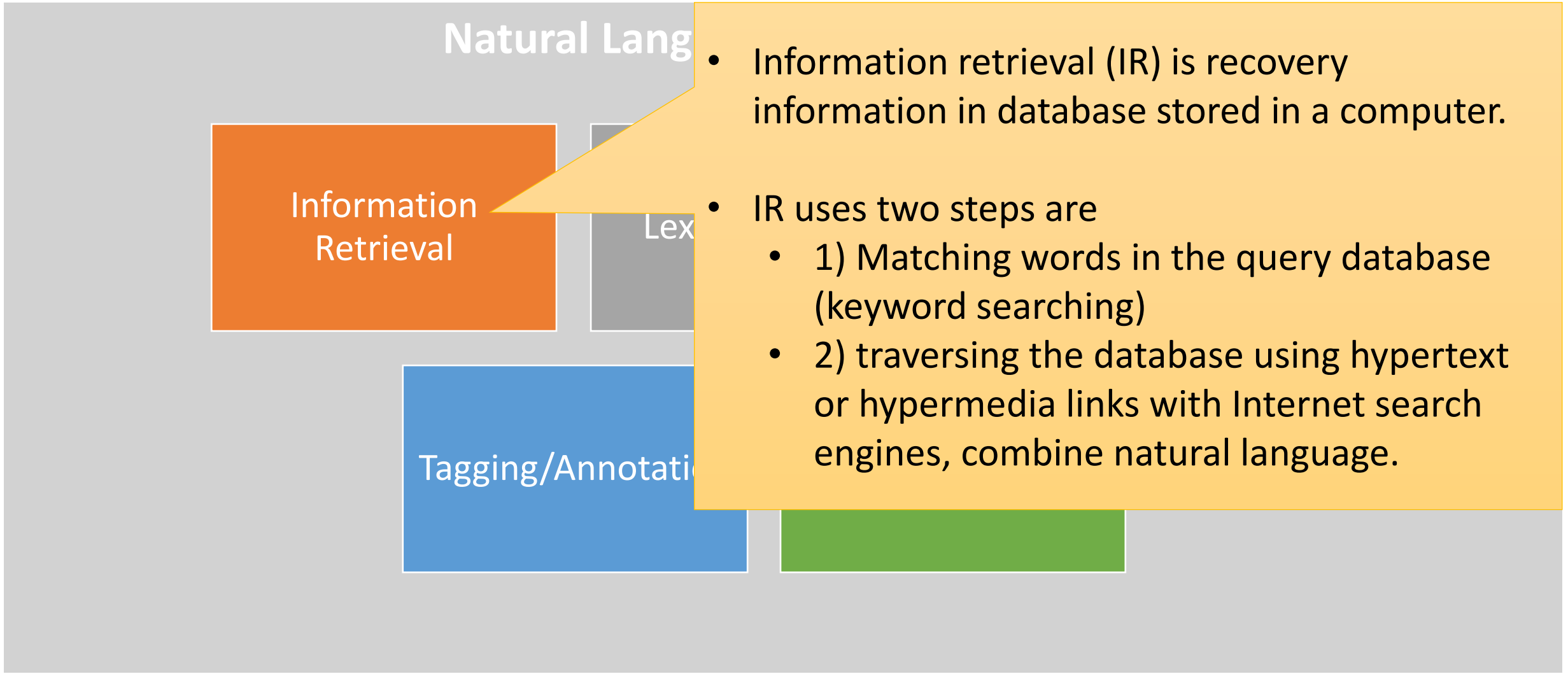


Natural Language Processing (NLP)
Linguistics

Text analysis: The inside of an NLP

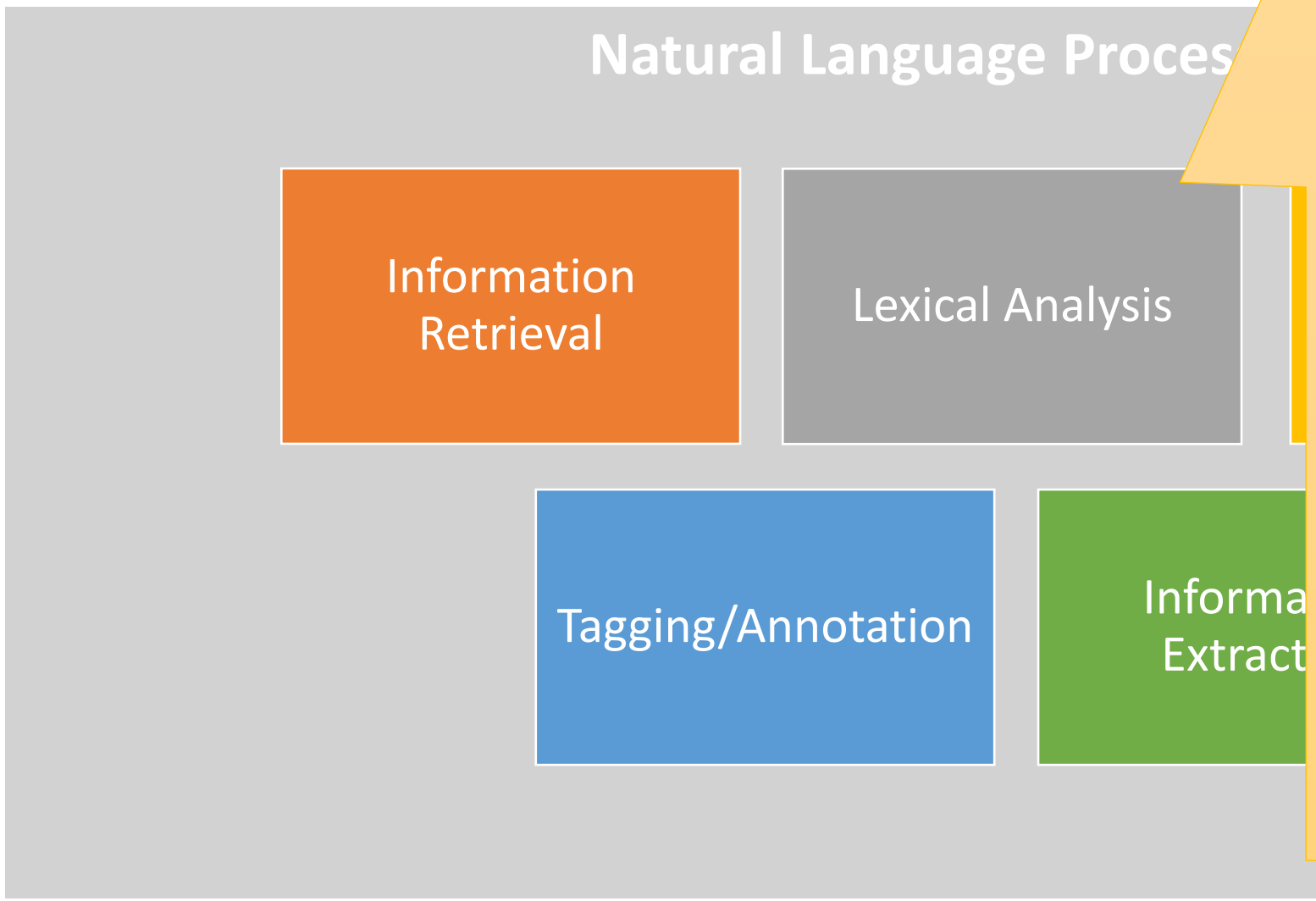


Text analysis: The inside of an NLP



- Information retrieval (IR) is recovery information in database stored in a computer.
- IR uses two steps are
  - 1) Matching words in the query database (keyword searching)
  - 2) traversing the database using hypertext or hypermedia links with Internet search engines, combine natural language.

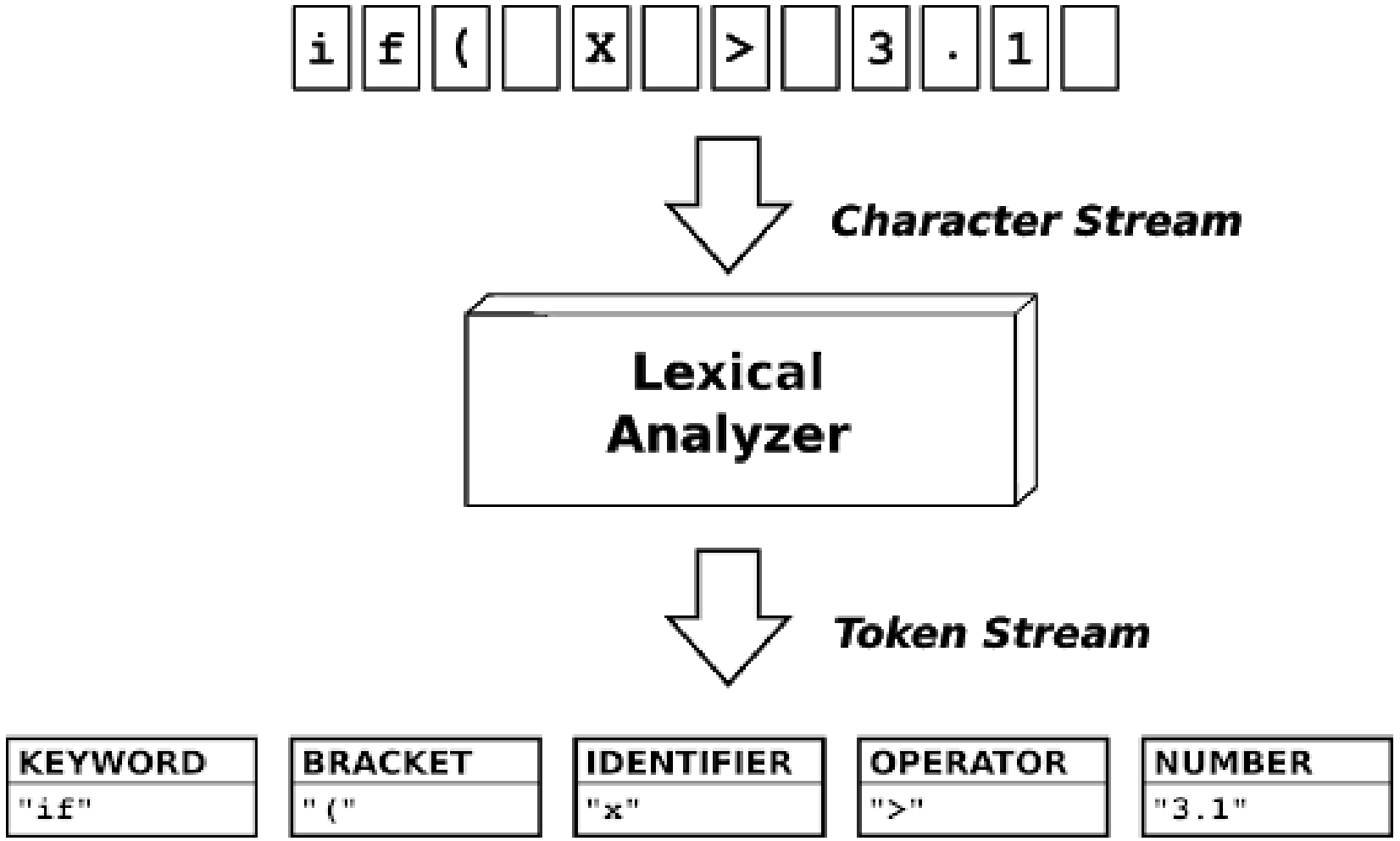
Text analysis: The inside of an



- Lexical analysis is a process converting a sequence of character to a sequence token.
-
- Lexical analysis is a first phase in the compiler design.

- Example:
  - Sequence character
    - a = 10;
    - a = a+5;
  - Sequence token
    - a (operator)
    - = (assign a value)
    - 10 (value)
    - + (increment)

Lexical analysis



Output from Text analysis

1. Indexing data
2. Document summarization
3. Sentiment analysis
4. Translation
5. Prediction and forecasting

Example research

- No decimator
  - ฉันรักคุณ| I love you.

A Hybrid Approach for Thai Word Segmentation with Crowdsourcing Feedback System

Kriangkrai Chaonithi
Computer Engineering Department,
Faculty of Engineering
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
kriangkrai.chao@mail.kmutt.ac.th

Santitham Prom-on
Computer Engineering Department,
Faculty of Engineering
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
santitham@cpe.kmutt.ac.th

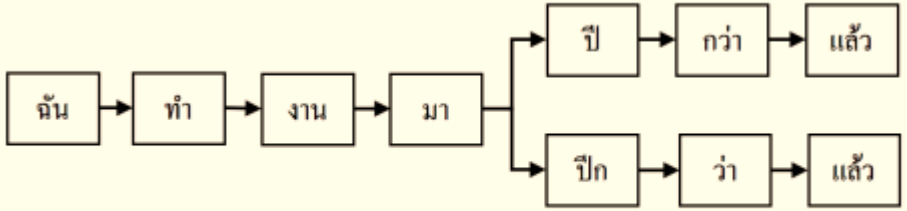


Fig. 1. Example of ambiguity problem

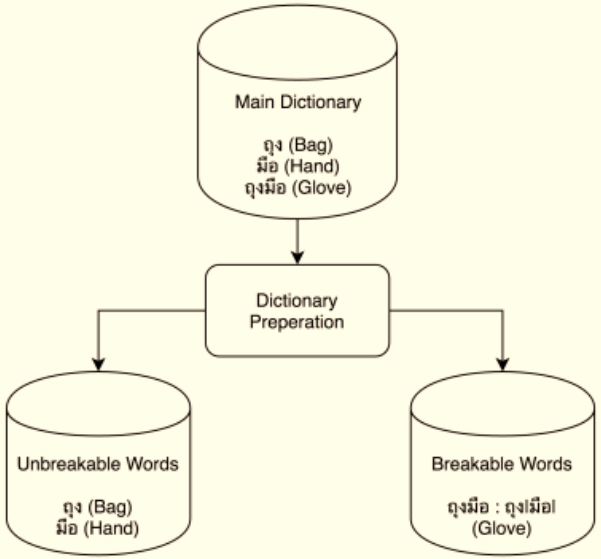


Fig. 3. Dictionary reconstruction process

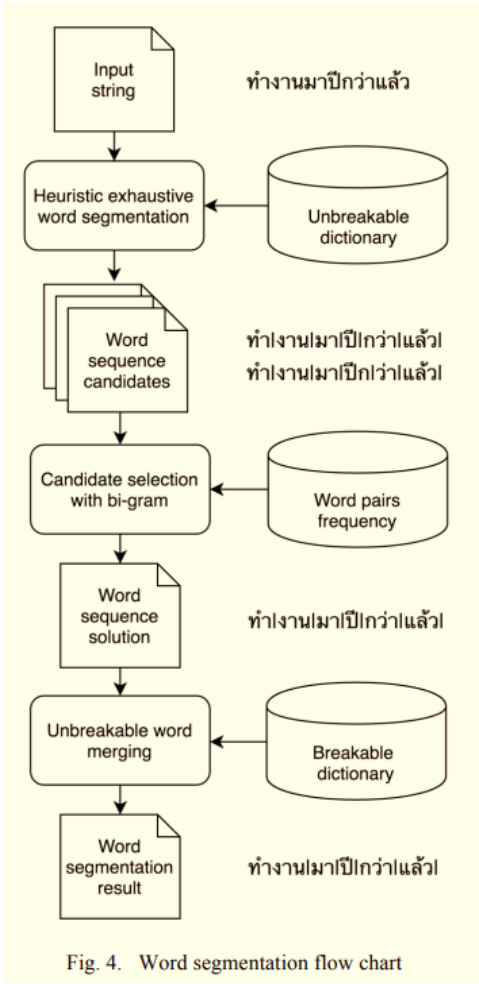
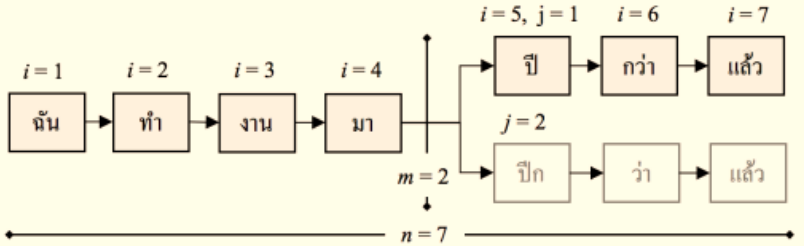


Fig. 4. Word segmentation flow chart

$$Score(candidate) = \prod_{i=1}^{n-1} \frac{f(w_i, w_{i+1})}{\sum_{j=1}^{m} f(w_i, w_j)} \quad (1)$$



Word pair frequency assumption: มา|ปี = 9 and มา|ปีก = 1

$$Score(\text{ฉัน|ทำ|งาน|มา|ปี|กว่า|แล้ว}) = 1 \cdot 1 \cdot 1 \cdot \frac{9}{10} \cdot 1 \cdot 1 = 0.9$$

$$Score(\text{ฉัน|ทำ|งาน|มา|ปีก|ว่า|แล้ว}) = 1 \cdot 1 \cdot 1 \cdot \frac{1}{10} \cdot 1 \cdot 1 = 0.1$$

Fig. 5. Example of word sequence candidate selection

Text data

Source of text Data and data types

Data Source	Data Format	Data Structure Type
News articles	TXT, HTML, or Scanned PDF	Unstructured
Literature	TXT, DOC, HTML, or PDF	Unstructured
E-mail	TXT, MSG, or EML	Unstructured
Web pages	HTML	Semi-structured
Server logs	LOG or TXT	Semi-structured or Quasi-structured
Social network API firehoses	XML, JSON, or RSS	Semi-structured
Call center transcripts	TXT	Unstructured

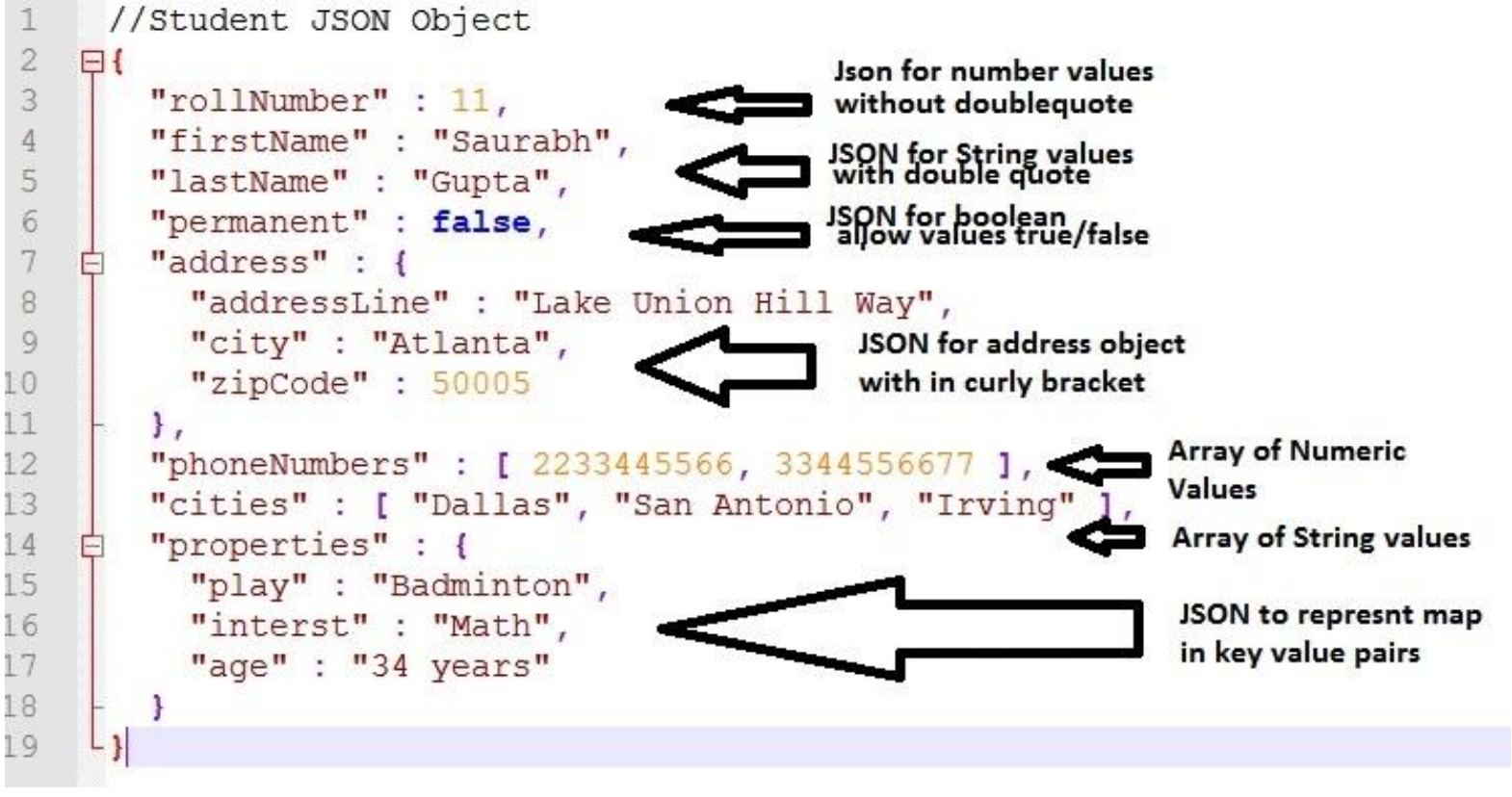
HTML / Log data

```html
<html>
<body>
    <table border="1">
        <tr>
            <td colspan="5" align="center">Temperture     </td>
        </tr>
        <tr>
            <td>CITIES </td>
            <td>DELHI  </td>
            <td>MUMBAI  </td>
            <td>KOLKATTA </td>
            <td>CHENNIA  </td>
        </tr>

        <tr>
            <td>MAXIMUM</td>
            <td>21 </td>
            <td>35 </td>
            <td>43 </td>
            <td>50  </td>
        </tr>

        <tr>
            <td>MINIMUM</td>
            <td>5  </td>
            <td>14</td>
            <td>28 </td>
            <td>32 </td>
        </tr>

    </table>
</body>
</html>
```

```
214.1.211.251 - - [15/Apr/2011:09:40:17 -0700] "GET /global.asa HTTP/1.0" 404 315 "-" "-"
214.1.211.251 - - [15/Apr/2011:09:40:17 -0700] "GET /~root HTTP/1.0" 404 310 "-" "-"
214.1.211.251 - - [15/Apr/2011:09:40:18 -0700] "GET /~apache HTTP/1.0" 404 312 "-" "-"
219.167.17.173 - - [17/Apr/2011:17:55:40 -0700] "POST /sony/mmr HTTP/1.1" 200 130 "-" "PS
218.41.54.67 - - [17/Apr/2011:18:20:18 -0700] "POST /sony/mmr HTTP/1.1" 200 130 "-" "PS3A
10.132.93.114 - - [18/Apr/2011:11:05:39 -0700] "POST /sony/mmr HTTP/1.1" 200 61 "-" "Ledi
10.132.93.114 - - [18/Apr/2011:11:07:07 -0700] "POST /sony/mmr HTTP/1.1" 200 61 "-" "Ledi
10.132.93.114 - - [18/Apr/2011:11:13:52 -0700] "POST /sony/mmr HTTP/1.1" 200 61 "-" "Ledi
218.41.54.67 - - [20/Apr/2011:17:42:37 -0700] "POST /sony/mmr HTTP/1.1" 200 100 "-" "PS3A
60.34.131.229 - - [20/Apr/2011:18:22:32 -0700] "POST /sony/mmr HTTP/1.1" 200 100 "-" "PS3
202.213.251.245 - - [21/Apr/2011:21:16:45 -0700] "POST /sony/mmr HTTP/1.1" 200 100 "-" "P
202.213.251.245 - - [21/Apr/2011:21:24:43 -0700] "POST /sony/mmr HTTP/1.1" 200 100 "-" "P
178.202.110.92 - - [22/Apr/2011:18:59:05 -0700] "GET / HTTP/1.1" 200 315 "-" "Mozilla/5.0
178.202.110.92 - - [22/Apr/2011:18:59:05 -0700] "GET /favicon.ico HTTP/1.1" 404 333 "-" "
178.202.110.92 - - [22/Apr/2011:18:59:05 -0700] "GET /favicon.ico HTTP/1.1" 404 333 "-" "
178.202.110.92 - - [22/Apr/2011:18:59:07 -0700] "GET /access-navigator-media HTTP/1.1" 20
178.202.110.92 - - [22/Apr/2011:19:05:00 -0700] "GET /admin/cdr/counter.txt HTTP/1.1" 404
178.202.110.92 - - [22/Apr/2011:19:05:41 -0700] "GET //help/readme.nsf?OpenAbout HTTP/1.1
178.202.110.92 - - [22/Apr/2011:19:05:54 -0700] "GET /catinfo?A HTTP/1.1" 404 329 "-" "Mo
178.202.110.92 - - [22/Apr/2011:19:06:08 -0700] "GET /errors-navigator-media HTTP/1.1" 20
178.202.110.92 - - [22/Apr/2011:19:27:04 -0700] "GET / HTTP/1.1" 200 315 "-" "Mozilla/5.0
```

JavaScript Object Notation (JSON)

JSON is a lightweight format for storing and transporting data that uses sending data between the sever and the client.



Activity 11.1 Convert to JSON

```
#install.packages("jsonlite")
library(jsonlite)
x = list(alpha = 1:5,
                beta = "Bravo",
            gamma = list(a=1:3,
                 b=NULL))
json <- toJSON(x)
json


fromJSON( json )
```

```
> json
{"alpha":[1,2,3,4,5],"beta":["Bravo"],"gamma":{"a":[1,2,3],"b":{}}}
>
> fromJSON( json )
$alpha
[1] 1 2 3 4 5

$beta
[1] "Bravo"

$gamma
$gamma$a
[1] 1 2 3

$gamma$b
named list()
```

eXtensible Makeup Language (XML) vs HTML

Key attribute	XML	HTML
Using	Describe the data	Display data
Tag type	User defined	Predefined
Case sensitivity	Yes	No

XML

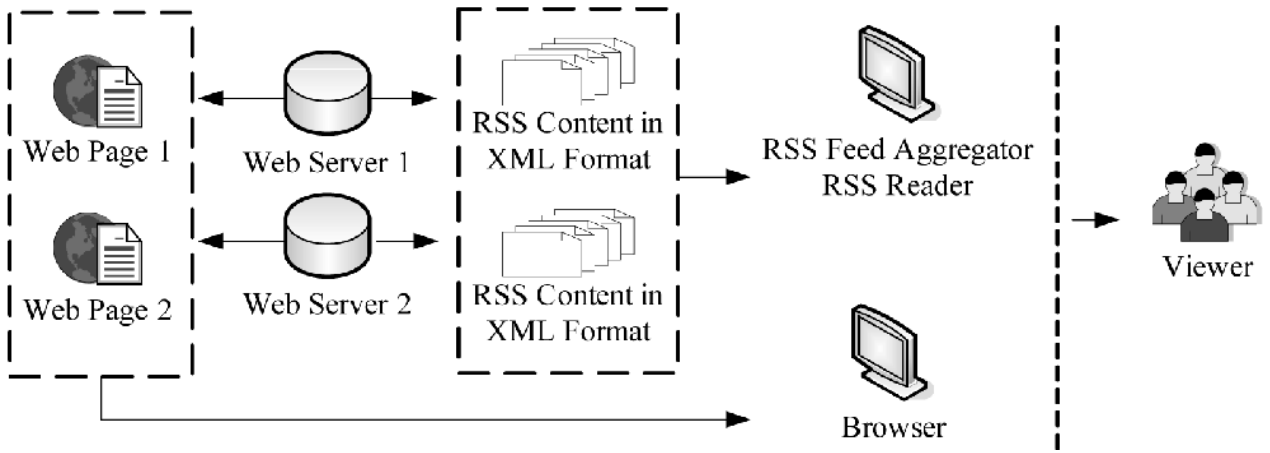
```
<firstName>Maria</firstName>
<lastName>Roberts</lastName>
<dateBirth>12-11-1942</dateBirth>
```

HTML

```
<font size="3">Maria Roberts</font>
<b>12-11-1942</b>
```

Reality Simple Syndication (RSS)

RSS is web feed that allows user and application to access update to website.




```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <rss version="2.0">
  - <channel>
      <title>CNN.com</title>
      <link>http://www.cnn.com/rssclick/?section=cnn_topstories</link>
      <description>CNN.com delivers up-to-the-minute news and information on t
        politics and more.</description>
      <language>en-us</language>
      <copyright>© 2006 Cable News Network LP, LLLP.</copyright>
      <pubDate>Fri, 07 Apr 2006 11:06:02 EDT</pubDate>
      <ttl>5</ttl>
    - <image>
        <title>CNN.com</title>
        <link>http://www.cnn.com/rssclick/?section=cnn_topstories</link>
        <url>http://i.cnn.net/cnn/.element/img/1.0/logo/cnn.logo.rss.gif</url>
        <width>144</width>
        <height>33</height>
        <description>CNN.com delivers up-to-the-minute news and information or
          entertainment, politics and more.</description>
      </image>
```

RSS File for CNN's Top Stories

Research and work related to text analysis

Real Time Sentiment Analysis of Twitter Data Using Hadoop

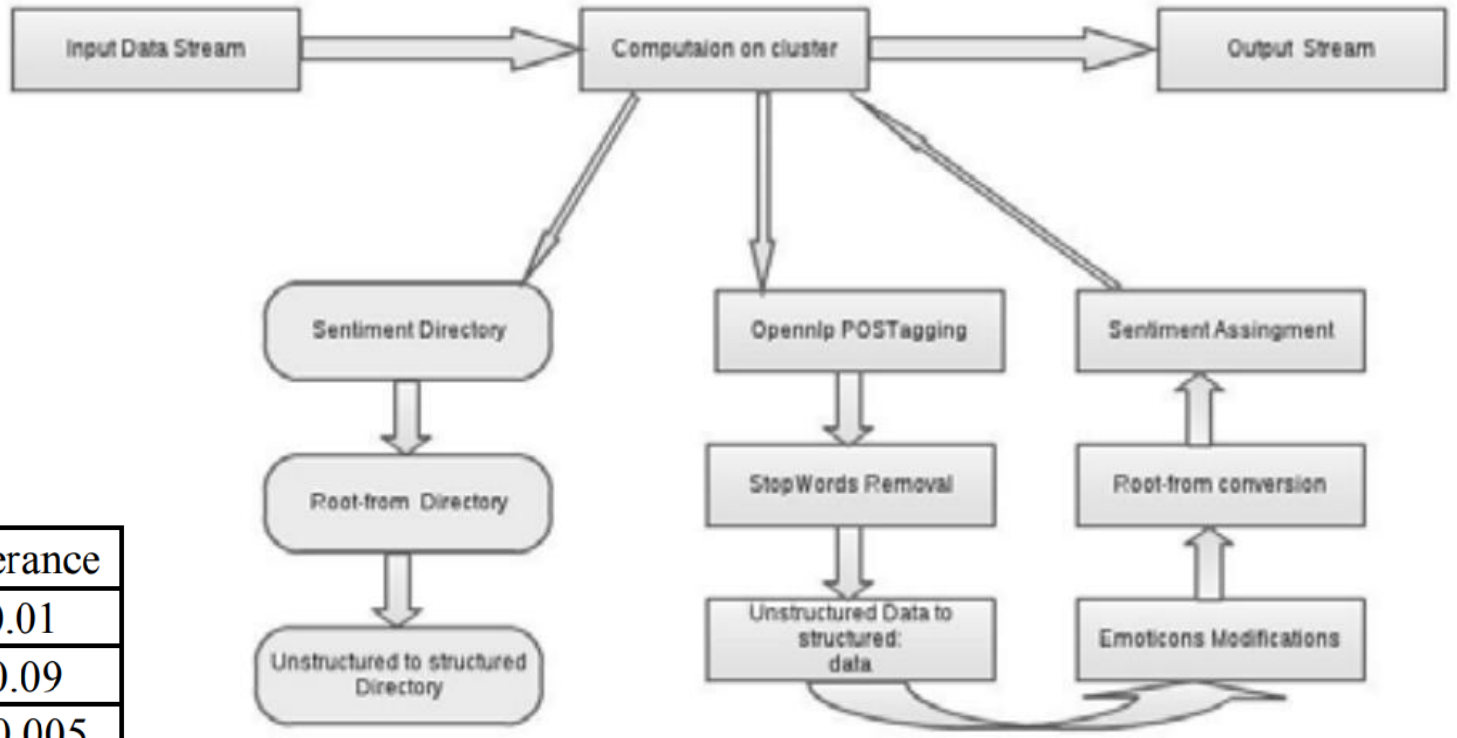
Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde

VIII. Conclusion

Sentiment analysis is a very wide branch for research. We have covered some of the important aspects. We plan ahead to improve our algorithm used for determining the sentiment value. Also the project as of now can also be expanded to other social media platform usages like movie reviews(IMDB reviews), personal blogs. The accuracy achieved is also mentioned below.[6]

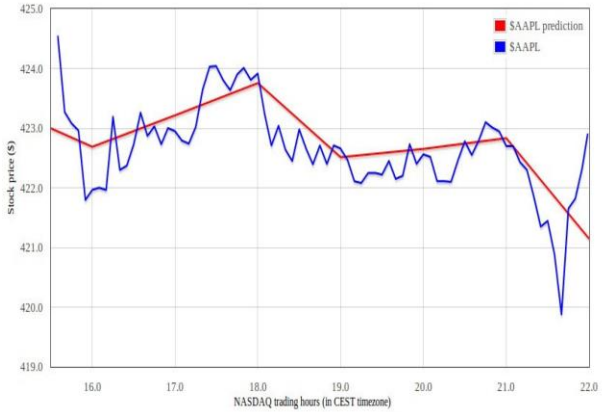
Emoticons and the use of hashtags for the sentiment evaluation is a very important inference related to sentiment analysis of social media data. Our project uses emoticons but the use of hashtags to determine the context of the tweet is not done. Hence with the current limitations the accuracy is found to be 72.27 %.

Sentiment	Count	Correct	%	Tolerance
Positive	729	542	74.34	-0.01
Negative	665	458	68.87	+0.09
Neutral	72	53	73.61	+- 0.005



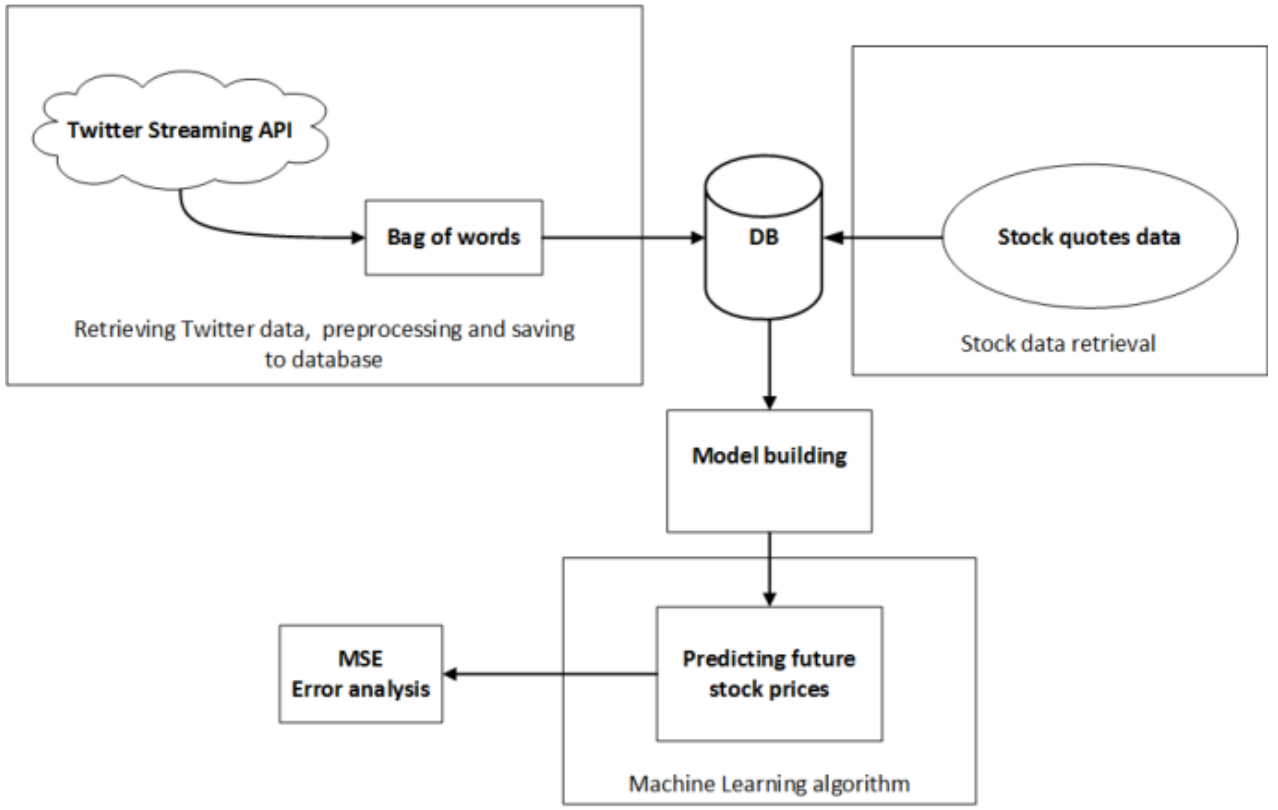
Sentiment analysis of Twitter data within big data distributed environment for stock prediction

Michal Skuza, A. Romanowski · Published 2015 · Computer Science ·
2015 Federated Conference on Computer Science and Information Systems (FedCSIS)



Abstract:

This paper covers design, implementation and evaluation of a system that may be used to predict future stock prices basing on analysis of data from social media services. The authors took advantage of large datasets available from Twitter micro blogging platform and widely available stock market records. Data was collected during three months and processed for further analysis. Machine learning was employed to conduct sentiment classification of data coming from social networks in order to estimate future stock prices. Calculations were performed in distributed environment according to Map Reduce programming model. Evaluation and discussion of results of predictions for different time intervals and input datasets proved efficiency of chosen approach is discussed here.



STOCK PRICES.

MSE	1 Hour	30 Min	15 Min	5 Min
Manual & 'AAPL'	1.5373	0.6325	0.3425	-
Auto & 'AAPL'	0.947	0.3698	0.2814	-
Manual & 'Apple'	1.9287	1.5152	0.9052	0.5764
Auto & 'Apple'	1.8475	1.4549	0.8325	0.3784

TABLE 1: MEAN SQUARE ERROR VALUES OF PREDICTED AND ACTUAL STOCK PRICES.

Solving Unbalanced Data for Thai Sentiment Analysis

Warunya Wunnasri, Thanaruk Theeramunkong
School of Information, Computer and Communication Technology
Sirindhorn International Institute of Technology
Thammasat University, Thailand
warunya.wunnasri@studentmail.siit.tu.ac.th, thanaruk@siit.tu.ac.th

Choochart Haruechaiyasak
Speech and Audio Technology (SPT) Laboratory
National Electronics and Computer Technology Center (NECTEC), Thailand
choochart.haruechaiyasak@nectec.or.th

***Abstract*— Growth of microblogging "Twitter" is dramatic among online users in Thailand. Communication on Twitter is very lively and up-to-date since users Users often express their feelings and sentiments in Twitter posts related to current topics or new growing topic. While sentiment analysis on Twitter has challenges in language related issues, such as short-length message and word usage variation, it also faces the problem of unbalanced class problem. In Twitter, people tend to make complaints more than admirations. In this paper, we propose a sampling-based method to solve data unbalanceness in Twitter sentiment analysis in Thai. Three types of sampling methods, called random, largest complete-link sampling, and largest average-link sampling are produced as preprocess before k-NN classifier. From the experimental results, the largest average-linkage sampling achieves the highest performance with the macro average F-measure of 0.57 comparing to the unbalance case.**

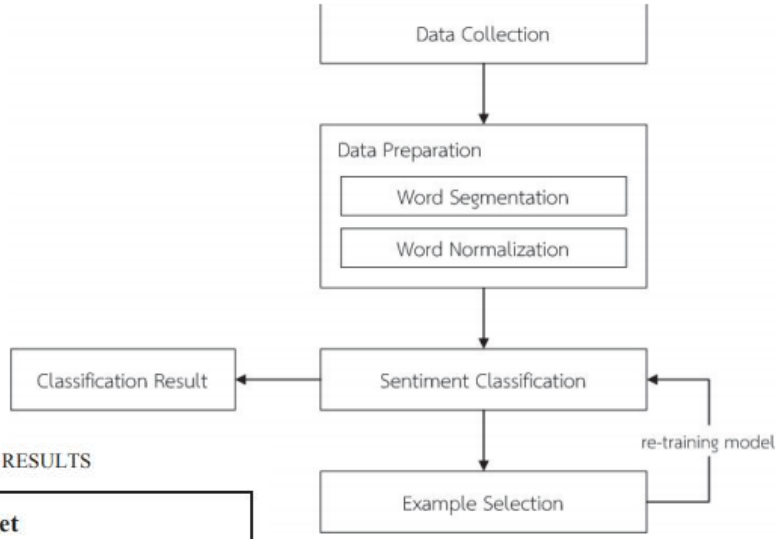


TABLE II. THE EXAMPLE OF ERROR RESULTS

Result	Polarity Class	Tweet
Negative	Neutral	(Original text) มีแต่ของแม่ที่เป็น ดีแทค เราใช้ เอไอเอส เล่นเน็ต ใช้ ทรู โทรออก // ไม่ค่อยได้โทร ไม่รู้จะโทรหาใคร
		(Translation) My mother use Dtac's cellular. I use AIS's network for play internet but use TRUE for cal. (I rarely call to someone. I have no idea to call to whomever.)
Positive	Negative	(Original text) โอเค True care ตอบตรงดี ไม่ได้ก็บอกไม่ได้ ... ไม่แถให้หงุดหงิดใจ : บริการก็ใช้ได้นี่หว่า
		(Translation) OK, 'True care' (the name of call center) answer me truthfully. If they cannot do, they told to me honestly. They do not answer indirectly so not make me moody. Their service is sufficiently
Negative	Positive	(Original text) แค่ยกเลิกเพคเกจมันยากเย็นนักหรอ1678โทรติดกี่ยาก 16789 ส่งฟรีจะไม่บ่นเลย
		(Translation) Dose the canceling the package hard? It very very hard when I call to your call center 1678. But if I sent the request message to 16789 it has a fee. If it has no fee, I will not complain about it.

Text analysis steps

Preprocessing

- Remove noise is the changing to original value.
  - WELLCOME, WeLCOME
  - อา-หารการกิน
- Remove stop words in phases
  - "I love my dog."

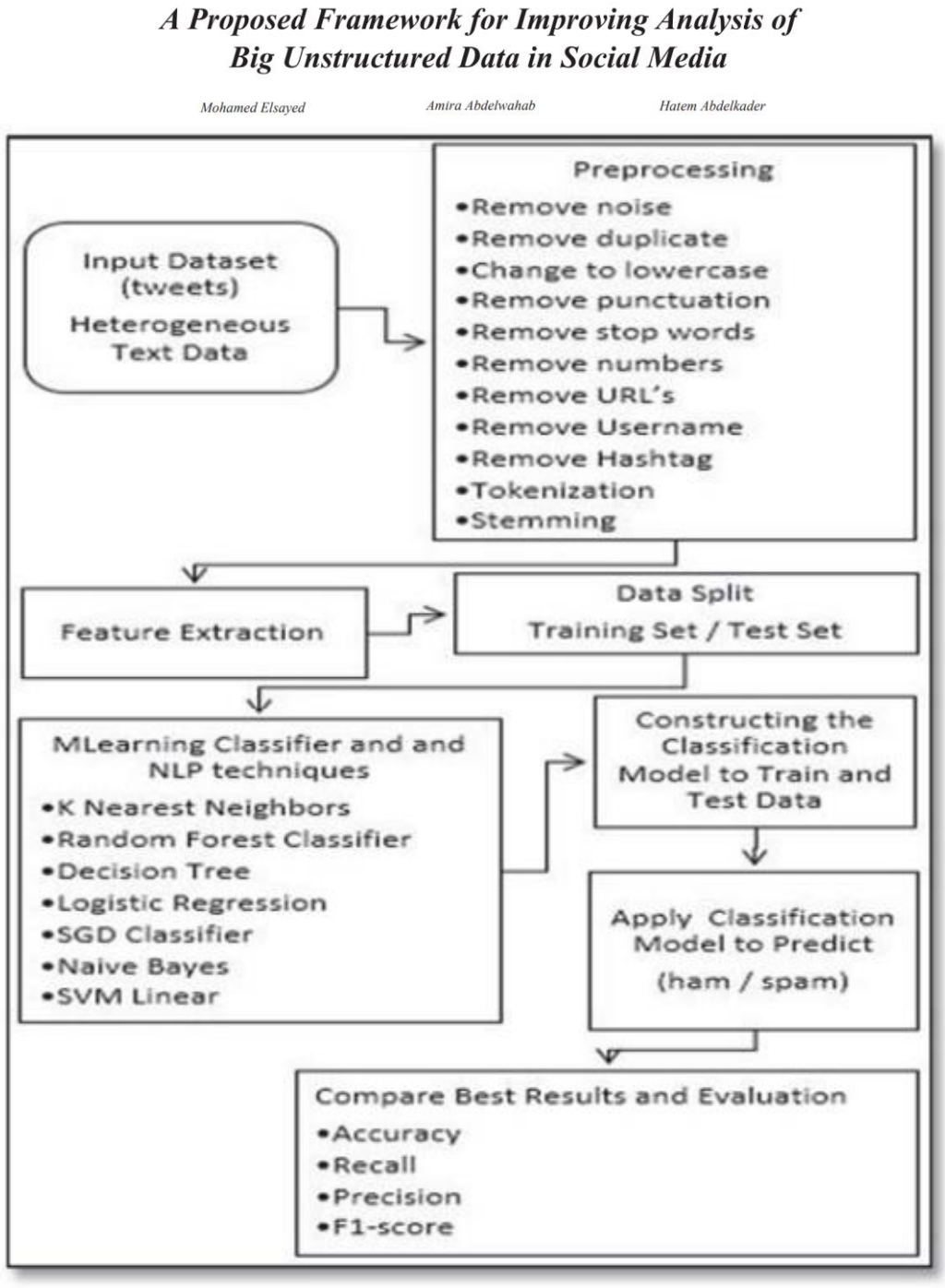
'จุดจบ' ของประโยค

Posted on July 29, 2012 by tonwachara

ปัญหาในไวยากรณ์ไทยที่ได้รับการกล่าวถึงอย่างมากคือ การไม่มีเครื่องหมายสำหรับจบประโยค เป็นที่รู้กันดีว่า ภาษาส่วนมากมักใช้เครื่องหมายมหัพภาคหรือเครื่องหมายจุด (.) เมื่อจบประโยค แต่ในภาษาไทยราชบัณฑิตยสถานได้ระบุว่า ภาษาไทยจะเว้นวรรคเมื่อจบประโยค โดยระหว่างประโยคจะเว้นวรรคมากกว่าธรรมดาเรียกว่า "วรรคใหญ่" ซึ่งต่างจากการเว้นวรรคภายในประโยคจะใช้วรรคขนาดเล็กเรียกว่า "วรรคเล็ก" แต่ด้วยระบบการพิมพ์ในปัจจุบัน การสร้างความแตกต่างระหว่างวรรคใหญ่กับวรรคเล็กแทบเป็นไปไม่ได้เลย แม้จะพูดว่า "วรรคใหญ่" ให้เคาะสองครั้ง "วรรคเล็ก" ให้เคาะครั้งเดียว หรือใช้วรรคขนาดต่าง ๆ ใน Unicode ผลลัพธ์ที่ได้ก็ยังเห็นความแตกต่างได้ยาก อย่างในบทความนี้มีทั้งวรรคใหญ่ (เคาะสองครั้ง) และวรรคน้อย (เคาะครั้งเดียว) ลองดูสิว่าคุณสามารถแยกความแตกต่างออกหรือไม่?

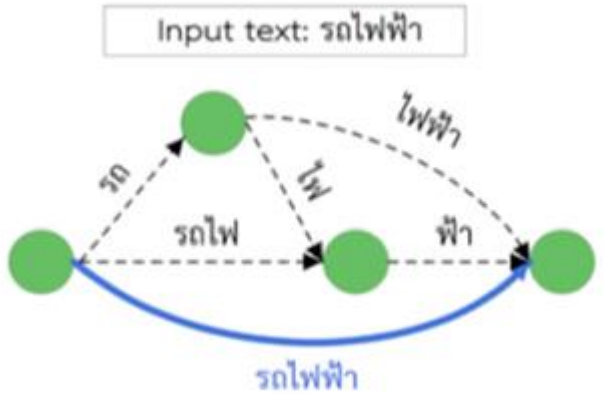
ด้วยเหตุผลที่ภาษาไทยไม่เว้นวรรคระหว่างคำ ทำให้เราแบ่งประโยคได้โดยไม่จำเป็นต้องใช้เครื่องหมายวรรคตอน อย่างไรก็ดี เครื่องหมายวรรคตอนสำหรับแบ่งความเรียง ยังปรากฏให้เห็นในภาษาไทยโบราณ โดยเฉพาะอย่างยิ่งในวรรณคดี เช่น เครื่องหมายฟองมัน (๐) เครื่องหมายไคมุดร (๛) ที่ใช้สำหรับเริ่มและจบบทกลอนตามลำดับ เพราะในสมัยโบราณ การเขียนโคลงกลอนไม่ได้เว้นวรรคตามฉันทลักษณ์แบบปัจจุบัน หากแต่เขียนต่อเนื่องกันคล้ายความเรียง เนื่องจากสมัยก่อน คนไทยบันทึกอักษรบนใบลาน หรือ "สมุดข่อย" ซึ่งมักเป็นแถบยาว ๆ การเว้นวรรคแบบปัจจุบันจึงเปลือง "กระดาษ" และไม่สอดคล้องกับรูปร่างของ "สมุดข่อย" เราจึงจำเป็นต้องใช้เครื่องหมายวรรคตอนที่ใช้เริ่มหรือจบบทกลอนในสมัยนั้น

ปัจจุบัน ผู้ใช้อินเตอร์เน็ตชาวไทยบางส่วน พยายามรณรงค์ให้ใช้เครื่องหมายวรรคตอนเมื่อจบประโยค เช่น วิกิพีเดียภาษาไทยช่วงเริ่มต้น หรือ การทดลองในบล็อกของคุณวีร์ ซึ่งวัตถุประสงค์ส่วนใหญ่คือให้ใช้เครื่องหมายจุดในการจบประโยคเช่นเดียวกับสากล ที่น่าสนใจคือ ในพระบรมราโชวาทหรือพระราโชวาทบางฉบับ ก็ปรากฏเครื่องหมายจุดเมื่อจบประโยคด้วย แต่ในความคิดของผม สาเหตุที่เครื่องหมายจุดไม่เป็นที่นิยม อาจเป็นเพราะเครื่องหมายจุดมีขนาดเล็ก กรณีศึกษาอย่างภาษาจีน จะใช้เครื่องหมายจบประโยคเป็นวงกลมขนาดเล็กที่ใหญ่กว่าจุดปกติเล็กน้อย เพื่อให้สอดคล้องกับตัวอักษรจีนที่มีขนาดใหญ่เต็มบรรทัด

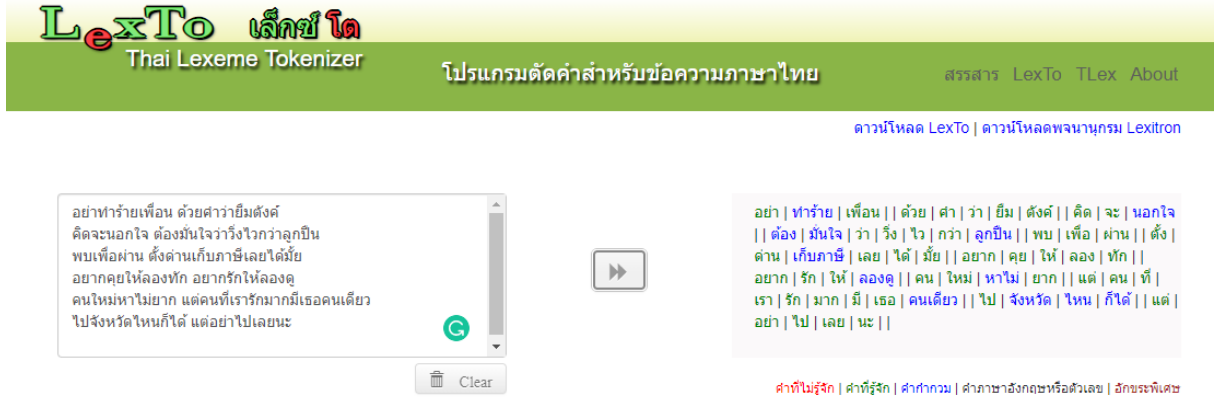


Preprocessing

- Tokenization ตัดคำ is the processing split a long sentence to a short sentence or a word.
  - Problem in Thai language
    - ตากลม
      - ตาก ลม
      - ตา กลม
    - ฉันนั่งตากลมที่หน้าบ้าน
      - ฉันนั่งตา กลมที่หน้าบ้าน
      - ฉันนั่งตาก ลมที่หน้าบ้าน
- Stemming word
  - Connect: connection, connected, connecting
  - Trident: tradition, traditional



http://www.sansarn.com/lexto/



https://www.youtube.com/watch?v=-3qG8ndG09w

Feature Extraction

- Word counting feature
  - **มีวินัยใจซื่อสัตย์ รู้ประหยัด เคร่งครัดคุณธรรม ขยัน ศึกษา ใฝ่หาความรู้ เชิดชูชาติ ศาสน์ กษัตริย์ เป็นคุณสมบัติของเด็กไทย รู้หน้าที่ ขยัน ซื่อสัตย์ ประหยัด มีวินัย และคุณธรรม รักวัฒนธรรมไทย ใฝ่ดี มีความคิด สุจริต ใจมั่น หมั่นศึกษา สามัคคี มีวินัย ใฝ่คุณธรรม นิยมไทย ใช้ประหยัด ใจซื่อสัตย์ ถือคุณธรรม**
  —รวมคำขวัญวันเด็ก เปรม ติณสูลานนท์

- Term frequency-inverse document frequency (TF-IDF)

Activity 11.2 String counting

```
str = "If you look at what you have in life, you'll always have more. If you look at what you don't have in life, you'll never have enough.- Oprah Winfrey"

nchar(str)    #number of character

str_count(str,"you")
```

```
tstr = "คำขวัญวันเด็กในปี 2516 ของ จอมพล ถนอม กิตติขจร ที่ระบุว่า เด็กดีเป็นศรีแก่ชาติ เด็กฉลาดชาติเจริญ ได้รับความชื่นชอบมากที่สุด"

nchar(tstr)    #number of character

str_count(tstr,"เด็ก")
```

```
> nchar(str)#number of character
[1] 132
> str_count(str,"you")
[1] 6
```

```
> nchar(tstr)#number of character
[1] 123
> str_count(tstr,"à´ç¡")
[1] 3
```

TF-IDF

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

- TF-IDF consists of the term frequency (TF) and the inverse document frequenc (IDF)
- Example, calculation TF

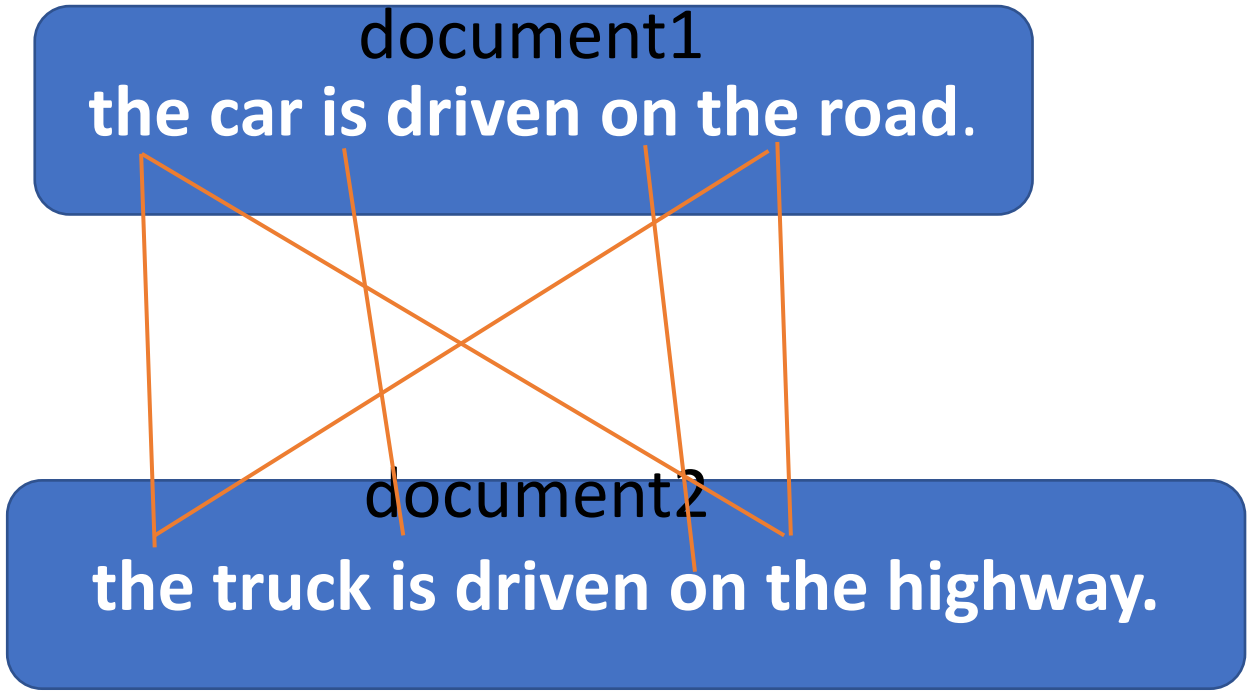
The sky is blue. The sky is beautiful.

the	= 2 / 8	= 0.25
sky	= 2 / 8	= 0.25
is	= 2 / 8	= 0.25
blue	= 1 / 8	= 0.125
beautiful	= 1 / 8	= 0.125
sum	= 8	

TF-IDF

- Example, calculation IDF

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \qquad idf(w) = log(\frac{N}{df_t})$$



Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	log(2/2) = 0	0	0
Car	1/7	0	log(2/1) = 0.3	0.043	0
Truck	0	1/7	log(2/1) = 0.3	0	0.043
Is	1/7	1/7	log(2/2) = 0	0	0
Driven	1/7	1/7	log(2/2) = 0	0	0
On	1/7	1/7	log(2/2) = 0	0	0
The	1/7	1/7	log(2/2) = 0	0	0
Road	1/7	0	log(2/1) = 0.3	0.043	0
Highway	0	1/7	log(2/1) = 0.3	0	0.043

The common words was zero, which shows they are not significant. On the other hand, the TF-IDF of “car” , “truck”, “road”, and “highway” are non-zero. These words have more significance.

Get data from Twitter with R

%>%

base

- Pipe operator is forward a result value to the next function call/expression.

```
#without pipe

a = c(1,3,2,5,6)

a=a+2

sort(a)
```

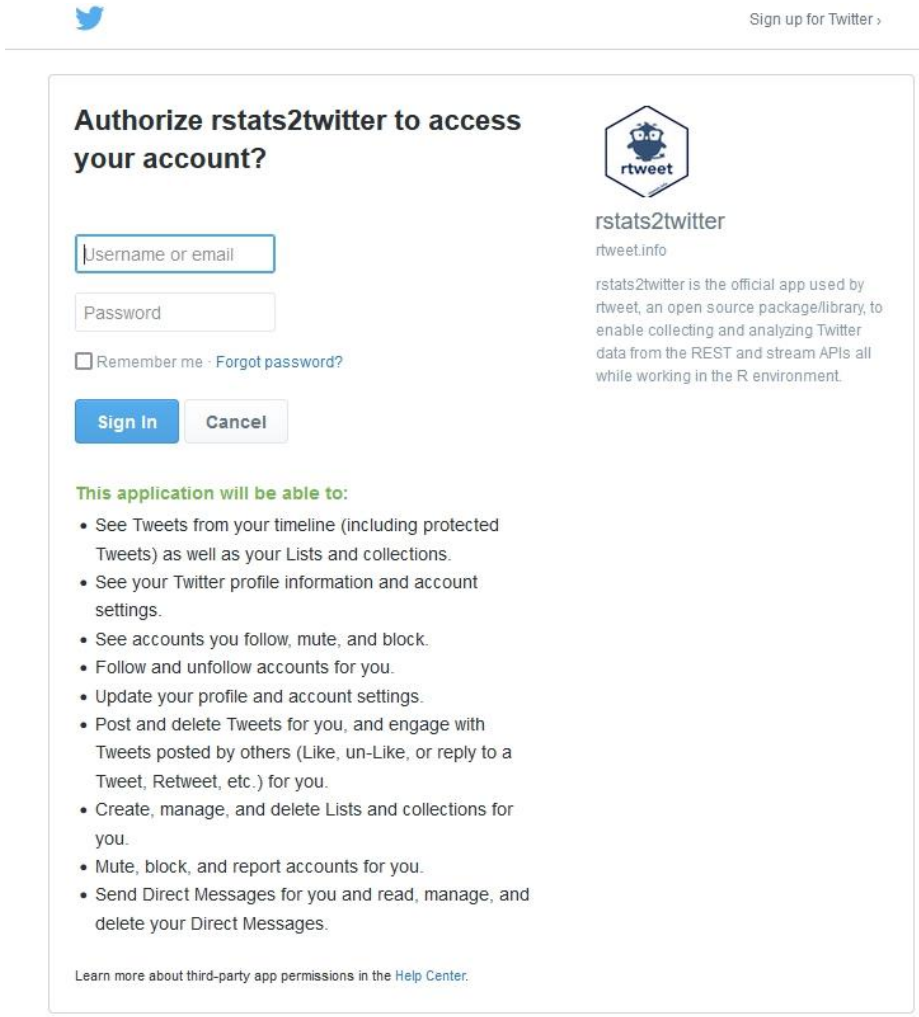
```
#two pipe

c(1,3,2,5,6) %>% +2

#three pipe

a = c(1,3,2,5,6) %>% +2 %>% sort()

print(a)
```

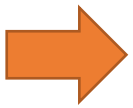
Activity 11.3 Get data from Twitter

```
install.packages("rtweet")
install.packages("ggplot2")
library(rtweet)
library(ggplot2)

rt <- search_tweets(q = "#โควิด", n=50)

names(rt)
rt$text    #Body text in twitter

head(rt, n = 2)
head(rt$screen_name)
length(unique(rt$location))
```

```
#plot location
rt %>%
    ggplot(aes(location)) +
    geom_bar() + coord_flip() +
    labs(x = "Count",
         y = "Location",
         title = "Twitter locations ")
```

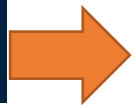
```
> names(rt)
 [1] "user_id"                "status_id"              "created_at"
 [4] "screen_name"            "text"                   "source"
 [7] "display_text_width"     "reply_to_status_id"     "reply_to_user_id"
[10] "reply_to_screen_name"   "is_quote"               "is_retweet"
[13] "favorite_count"         "retweet_count"          "quote_count"
[16] "reply_count"            "hashtags"               "symbols"
[19] "urls_url"               "urls_t.co"              "urls_expanded_url"
[22] "media_url"              "media_t.co"             "media_expanded_url"
[25] "media_type"             "ext_media_url"          "ext_media_t.co"
[28] "ext_media_expanded_url" "ext_media_type"         "mentions_user_id"
[31] "mentions_screen_name"   "lang"                   "quoted_status_id"
[34] "quoted_text"            "quoted_created_at"      "quoted_source"
[37] "quoted_favorite_count"  "quoted_retweet_count"   "quoted_user_id"
[40] "quoted_screen_name"     "quoted_name"            "quoted_followers_count"
[43] "quoted_friends_count"   "quoted_statuses_count"  "quoted_location"
[46] "quoted_description"     "quoted_verified"        "retweet_status_id"
[49] "retweet_text"           "retweet_created_at"     "retweet_source"
[52] "retweet_favorite_count" "retweet_retweet_count"  "retweet_user_id"
[55] "retweet_screen_name"    "retweet_name"           "retweet_followers_count"
[58] "retweet_friends_count"  "retweet_statuses_count" "retweet_location"
[61] "retweet_description"    "retweet_verified"       "place_url"
[64] "place_name"             "place_full_name"        "place_type"
[67] "country"                "country_code"           "geo_coords"
[70] "coords_coords"          "bbox_coords"            "status_url"
[73] "name"                   "location"               "description"
[76] "url"                    "protected"              "followers_count"
[79] "friends_count"          "listed_count"           "statuses_count"
[82] "favourites_count"       "account_created_at"     "verified"
[85] "profile_url"            "profile_expanded_url"   "account_lang"
[88] "profile_banner_url"     "profile_background_url" "profile_image_url"
```

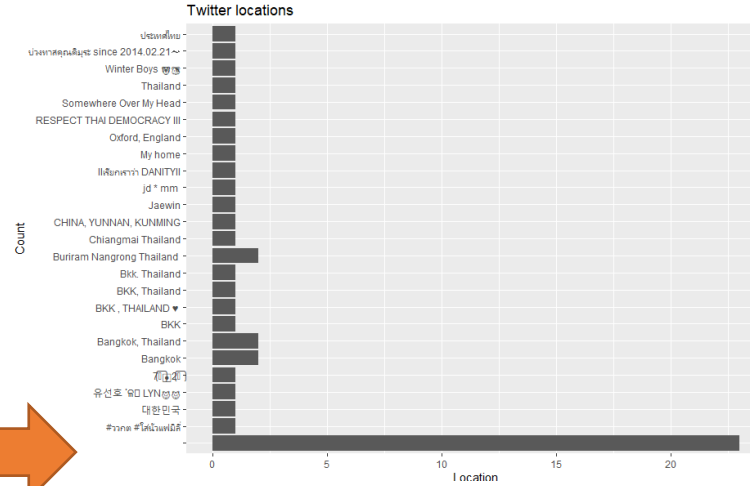




```
> rt$text   #Body text in twitter
 [1] "สถานการณ์ COVID-19 ในประเทศไทย\nวันที่ 09 เมษายน 2565\n\n#โควิด19 #COVID19 #สถานการณ์โควิด19 http
s://t.co/ThqJsCsCe2"

 [2] "อัปเดต #โควิด19 - 9 เม.ย. 65 - 08.00\n\nสถานการณ์ในประเทศ\n- พบผู้ป่วยเพิ่ม 25,298 ราย รวมสะสม 3,858,34
6 ราย\n(ในประเทศ 25,220 ราย / ตปท. 78 ราย)\n\n- หายป่วยเพิ่ม 22,003 ราย รวมหายแล้ว 3,580,920 ราย\n- เสียชีวิต
เพิ่ม 98 ราย รวมเสียชีวิตแล้ว 25,975 ราย\n\n//แก้ไขยอดหายป่วยสะสม https://t.co/ji05mXB91m"
 [3] "วันนี้ (9 เม.ย. 65) ศูนย์ข้อมูล #COVID19 รายงานจำนวนผู้ติดเชื้อโควิด-19 รายใหม่เพิ่ม 25,298 คน หายป่วยกลับบ้านเพิ่ม
 22,003 คน\n\n• รวมผู้ป่วยสะสมระลอกใหม่ 1,634,911 คน (ตั้งแต่ 1 ม.ค. 65)\n• ยอดหายป่วยสะสม 1,412,426 คน (ตั้ง
แต่ 1 ม.ค. 65)\n• เสียชีวิตเพิ่ม 98 คน\n\n#โควิด19 #ThaiPBS https://t.co/qerakUs6Y6"
 [4] "อัปเดตสถานการณ์ #COVID19 วันที่ 9 เม.ย. 2565\nติดตามประเด็น #โควิด19 ได้ที่ &gt;&gt; https://t.co/IxCqLf
C0SW\n\n\U0001f534 ผู้ติดเชื้อรายใหม่ 25,298 ราย\n\U0001f534 ผู้ติดเชื้อเข้าข่าย ATK 22,431 ราย\n\U0001f534 เสียชีวิ
ต 98 คน\n\U0001f534 หายป่วยกลับบ้าน 22,003 ราย \n\n#โควิด #โควิดวันนี้ #kapook https://t.co/lwBkh8RsiT"

 [5] "อัปเดตสถานการณ์ #COVID19 วันที่ 9 เม.ย. 2565\nติดตามประเด็น #โควิด19 ได้ที่ &gt;&gt; https://t.co/IxCqLf
C0SW\n\n\U0001f534 ผู้ติดเชื้อรายใหม่ 25,298 ราย\n\U0001f534 ผู้ติดเชื้อเข้าข่าย ATK 22,431 ราย\n\U0001f534 เสียชีวิ
ต 98 คน\n\U0001f534 หายป่วยกลับบ้าน 22,003 ราย \n\n#โควิด #โควิดวันนี้ #kapook https://t.co/lwBkh8RsiT"

 [6] "สถานการณ์ COVID-19 ในประเทศไทย\nวันที่ 09 เมษายน 2565\n\n#โควิด19 #COVID19 #สถานการณ์โควิด19 http
s://t.co/ThqJsCsCe2"
```

```
> head(rt, n = 2)
# A tibble: 2 x 90
  user_id status_id created_at          screen_name text  source display_text_wi~
  <chr>   <chr>     <dttm>              <chr>       <chr> <chr>              <dbl>
1 142234~ 15126033~ 2022-04-09 01:28:47 iamkupao    "สถา~ Twitt~               130
2 156329~ 15126033~ 2022-04-09 01:28:47 Pimry_zaa   "อัป~ Twitt~               140
# ... with 83 more variables: reply_to_status_id <lgl>, reply_to_user_id <lgl>,
#   reply_to_screen_name <lgl>, is_quote <lgl>, is_retweet <lgl>, favorite_count <int>,
#   retweet_count <int>, quote_count <int>, reply_count <int>, hashtags <list>,
#   symbols <list>, urls_url <list>, urls_t.co <list>, urls_expanded_url <list>,
#   media_url <list>, media_t.co <list>, media_expanded_url <list>, media_type <list>,
#   ext_media_url <list>, ext_media_t.co <list>, ext_media_expanded_url <list>,
#   ext_media_type <chr>, mentions_user_id <list>, mentions_screen_name <list>, lang <chr>,
#   quoted_status_id <chr>, quoted_text <chr>, quoted_created_at <dttm>, quoted_source <chr>,
#   quoted_favorite_count <int>, quoted_retweet_count <int>, quoted_user_id <chr>,
#   quoted_screen_name <chr>, quoted_name <chr>, quoted_followers_count <int>,
#   quoted_friends_count <int>, quoted_statuses_count <int>, quoted_location <chr>,
#   quoted_description <chr>, quoted_verified <lgl>, retweet_status_id <chr>,
#   retweet_text <chr>, retweet_created_at <dttm>, retweet_source <chr>,
#   retweet_favorite_count <int>, retweet_retweet_count <int>, retweet_user_id <chr>,
#   retweet_screen_name <chr>, retweet_name <chr>, retweet_followers_count <int>,
#   retweet_friends_count <int>, retweet_statuses_count <int>, retweet_location <chr>,
#   retweet_description <chr>, retweet_verified <lgl>, place_url <chr>, place_name <chr>,
#   place_full_name <chr>, place_type <chr>, country <chr>, country_code <chr>,
#   geo_coords <list>, coords_coords <list>, bbox_coords <list>, status_url <chr>,
#   name <chr>, location <chr>, description <chr>, url <chr>, protected <lgl>,
#   followers_count <int>, friends_count <int>, listed_count <int>, statuses_count <int>,
#   favourites_count <int>, account_created_at <dttm>, verified <lgl>, profile_url <chr>,
#   profile_expanded_url <chr>, account_lang <lgl>, profile_banner_url <chr>,
#   profile_background_url <chr>, profile_image_url <chr>
> head(rt$screen_name)
[1] "iamkupao"        "Pimry_zaa"       "ChatChaiPromrat" "ChatChaiPromrat" "edulocal"
[6] "loma_Poppy"
> length(unique(rt$location))
[1] 25
```





Activity 11.4 Word count in twitter

```
install.packages("rtweet")
install.packages("ggplot2")
library(rtweet)
library(ggplot2)

rt <- search_tweets(q = "#โควิด", n=500)

names(rt)
tbody = rt$text    #Body text in twitter

#install.packages("stringr")
library(stringr)
n = str_count(tbody,"ติดเชื้อ")
sum(n)
```

```
#install.packages("stringr")
n = str_count(tbody,"ดีขึ้น")
sum(n)

#install.packages("stringr")
n = str_count(tbody,"หาย")
sum(n)
```

```
> tbody = rt$text    #Body text in twitter
> #install.packages("stringr")
> library(stringr)
> n = str_count(tbody,"ติดเชื้อ")
> sum(n)
[1] 209
> #install.packages("stringr")
> n = str_count(tbody,"ดีขึ้น")
> sum(n)
[1] 0
> #install.packages("stringr")
> n = str_count(tbody,"หาย")
> sum(n)
[1] 13
```

bind_tf_idf {tidytext} R Documentation

Bind the term frequency and inverse document frequency of a tidy text dataset to the dataset

Description

Calculate and bind the term frequency and inverse document frequency of a tidy text dataset, along with the product, tf-idf, to the dataset. Each of these values are added as columns. This function supports non-standard evaluation through the tidyeval framework.

Usage

```
bind_tf_idf(tbl, term, document, n)
```

Arguments

`tbl`	A tidy text dataset with one-row-per-term-per-document
`term`	Column containing terms as string or symbol
`document`	Column containing document IDs as string or symbol
`n`	Column containing document-term counts as string or symbol

Details

The arguments `term`, `document`, and `n` are passed by expression and support quasiquotation; you can unquote strings and symbols.

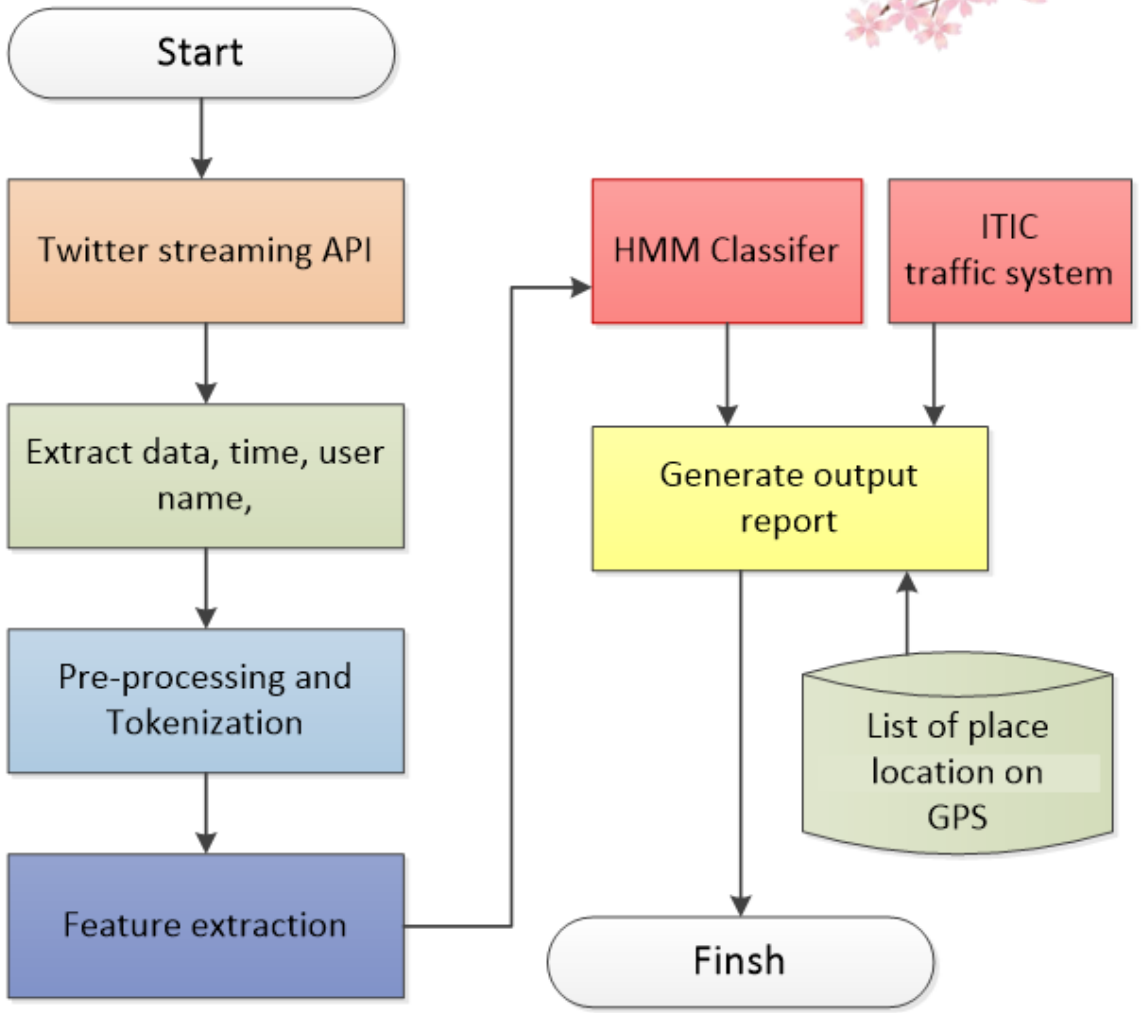
If the dataset is grouped, the groups are ignored but are retained.

The dataset must have exactly one row per document-term combination for this to work.

Summary

My research

System diagram



Preprocessing:
Represent symbols and preparing for classifier

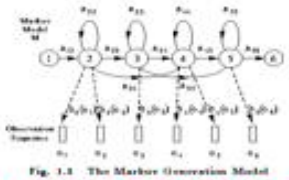
JS100

twitter

ถ.ราชพฤกษ์ ขาออก >วงเวียนบางขุนกอง เชิงทางลงสะพานข้ามคลองมหาสวัสดิ์ รถแวน เฉี่ยวชนกับ จยย. กีดขวางช่องทางซ้าย ของคู่ขนาน บาดเจ็บ

Exist from Ratchaphruek road at the circle of Bankunkong near the bridge cross Mahasawad carnal, a van crashes a bike cycle, obstacle traffic, having an injury people

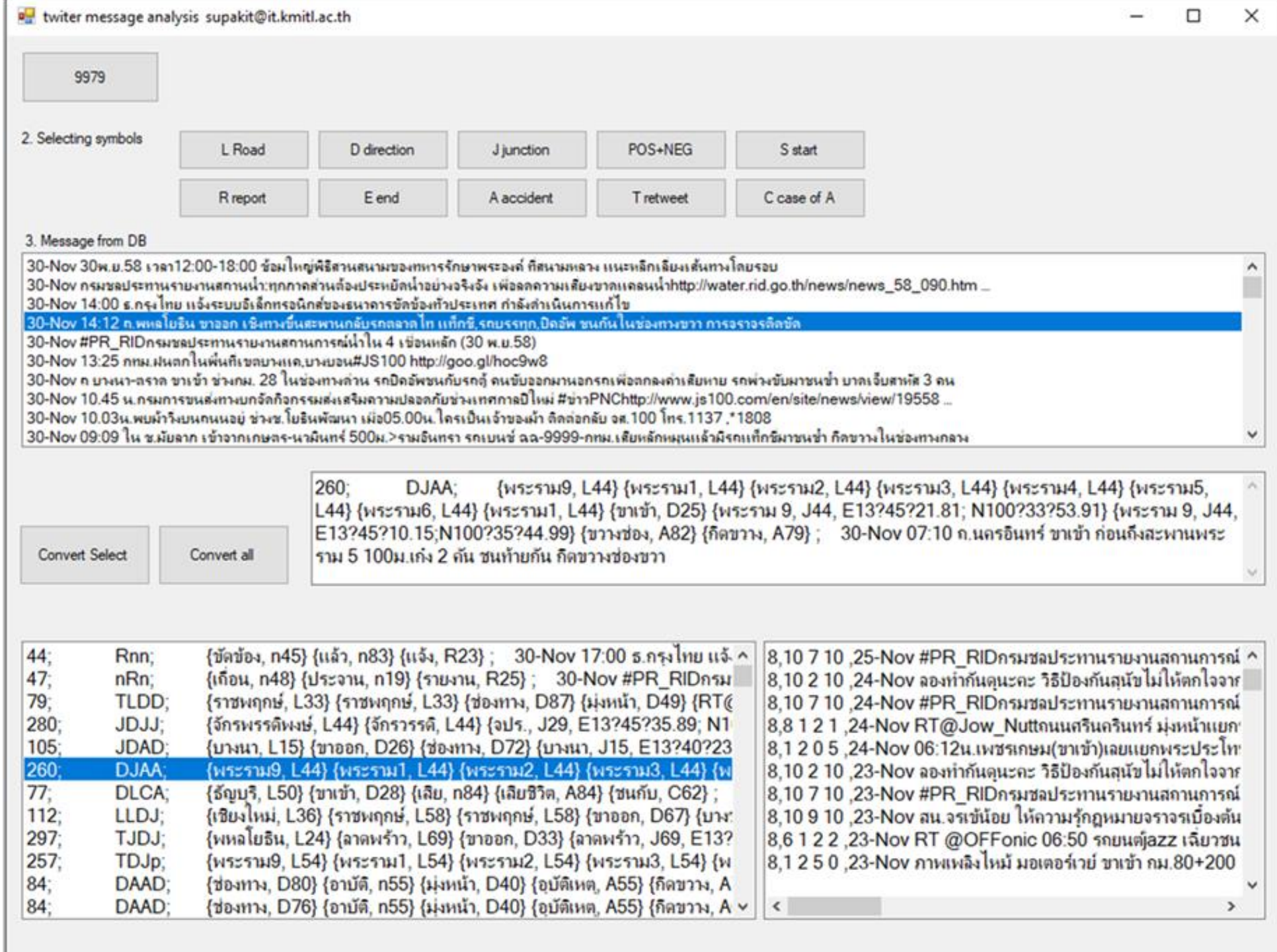
No	Symbols	Description	Example
1	R	Road name	
2	D	Direction	
3	J	Junction	
4	S	Start point	
5	E	End point	
6	P	Problem	
7	C	Cause of problem	



R2 R127 D11 E17 P91 P121 C74

N13.8229227, E100.4468671

{ราชพฤกษ์,R2} {บางนา,R127} {ขาออก,D11} {บางขุนกอง ,J25,N13.8229227,E100.4468671} {>,E17} {กีดขวาง,P91} {บาดเจ็บ,P121} {เฉี่ยวชน,C74}


twiter message analysis supakit@it.kmitl.ac.th
9979
2. Selecting symbols
L Road
D direction
J junction
POS+NEG
S start
R report
E end
A accident
T retweet
C case of A
3. Message from DB
30-Nov 30พ.ย.58 เวลา12:00-18:00 ซ้อมใหญ่พิธีสวนสนามของทหารรักษาพระองค์ ทีสนามหลวง แนะหลีกเลี่ยงเส้นทางโดยรอบ
30-Nov กรมชลประทานรายงานสถานน้ำ:ทุกภาคส่วนต้องประหยัดน้ำอย่างจริงจัง เพื่อลดความเสี่ยงขาดแคลนน้ำhttp://water.rid.go.th/news/news_58_090.htm ...
30-Nov 14:00 ธ.กรุงไทย แจ้งระบบธนาคารขัดข้องทั่วประเทศ กำลังดำเนินการแก้ไข
30-Nov 14:12 ก.พหลโยธิน ขาออก เริ่มทางขึ้นสะพานกลับรถลาดไท แท็กซี่,รถบรรทุก,บิ๊กอัพ ชนกันในช่องทางขวา การจราจรติดขัด
30-Nov #PR_RIDกรมชลประทานรายงานสถานการณ์น้ำใน 4 เขื่อนหลัก (30 พ.ย.58)
30-Nov 13:25 กทม.ฝนตกในพื้นที่เขตบางเขต,บางบอน#JS100 http://goo.gl/hoc9w8
30-Nov ก บางนา-ตราด ขาเข้า ช่วงกม. 28 ในช่องทางด่วน รถบิ๊กอัพชนกับรถตู้ คนขับออกมานอกรถเพื่อลงดูความเสียหาย รถพ่วงขับมาชนซ้ำ บาดเจ็บสาหัส 3 คน
30-Nov 10.45 น.กรมการขนส่งทางบกจัดกิจกรรมส่งเสริมความปลอดภัยช่วงเทศกาลปีใหม่ #ข่าวPNChttp://www.js100.com/en/site/news/view/19558 ...
30-Nov 10.03น.พบม้าวิ่งบนถนนอยู่ ช่วง ช.โยธินพัฒนา เมื่อ05.00น.ใครเป็นเจ้าของม้า ติดต่อกลับ จส.100 โทร.1137 ,*1808
30-Nov 09:09 ใน ช.มัยลาภ เข้าจากเกษตร-นวมินทร์ 500ม.>รามอินทรา รถเบนซ์ ฉฉ-9999-กทม.เสียหลักหมุนแล้วมีรถแท็กซี่มาชนซ้ำ กีดขวางในช่องทางกลาง
Convert Select
Convert all
260; DJAA; {พระราม9, L44} {พระราม1, L44} {พระราม2, L44} {พระราม3, L44} {พระราม4, L44} {พระราม5, L44} {พระราม6, L44} {พระราม1, L44} {ขาเข้า, D25} {พระราม 9, J44, E13?45?21.81; N100?33?53.91} {พระราม 9, J44, E13?45?10.15;N100?35?44.99} {ขวางช่อง, A82} {กีดขวาง, A79} ; 30-Nov 07:10 ก.นครอินทร์ ขาเข้า ก่อนถึงสะพานพระราม 5 100ม.เก๋ง 2 คัน ชนท้ายกัน กีดขวางช่องขวา
44; Rnn; {ขัดข้อง, n45} {แล้ว, n83} {แจ้ง, R23} ; 30-Nov 17:00 ธ.กรุงไทย แจ้
47; nRn; {เถื่อน, n48} {ประจาน, n19} {รายงาน, R25} ; 30-Nov #PR_RIDกรม
79; TLDD; {ราชพฤกษ์, L33} {ราชพฤกษ์, L33} {ช่องทาง, D87} {มุ่งหน้า, D49} {RT@
280; JDJJ; {จักรพรรดิพงษ์, L44} {จักรวรรดิ, L44} {จปร., J29, E13?45?35.89; N1
105; JDAD; {บางนา, L15} {ขาออก, D26} {ช่องทาง, D72} {บางนา, J15, E13?40?23
260; DJAA; {พระราม9, L44} {พระราม1, L44} {พระราม2, L44} {พระราม3, L44} {พ
77; DLCA; {ธัญบุรี, L50} {ขาเข้า, D28} {เลี่ยง, n84} {เลี่ยงชีวิต, A84} {ชนกับ, C62} ;
112; LLDJ; {เชียงใหม่, L36} {ราชพฤกษ์, L58} {ราชพฤกษ์, L58} {ขาออก, D67} {บาง
297; TJDJ; {พหลโยธิน, L24} {ลาดพร้าว, L69} {ขาออก, D33} {ลาดพร้าว, J69, E13?
257; TDJp; {พระราม9, L54} {พระราม1, L54} {พระราม2, L54} {พระราม3, L54} {พ
84; DAAD; {ช่องทาง, D80} {อาบัติ, n55} {มุ่งหน้า, D40} {อุบัติเหตุ, A55} {กีดขวาง, A
84; DAAD; {ช่องทาง, D76} {อาบัติ, n55} {มุ่งหน้า, D40} {อุบัติเหตุ, A55} {กีดขวาง, A
8,10 7 10 ,25-Nov #PR_RIDกรมชลประทานรายงานสถานการณ์
8,10 2 10 ,24-Nov ลองทำกันดูนะคะ วิธีป้องกันสุนัขไม่ให้ตกใจจาร
8,10 7 10 ,24-Nov #PR_RIDกรมชลประทานรายงานสถานการณ์
8,8 1 2 1 ,24-Nov RT@Jow_Nuttถนนศรีนครินทร์ มุ่งหน้าแยก
8,1 2 0 5 ,24-Nov 06:12น.เพชรเกษม(ขาเข้า)เลยแยกพระประโท
8,10 2 10 ,23-Nov ลองทำกันดูนะคะ วิธีป้องกันสุนัขไม่ให้ตกใจจาร
8,10 7 10 ,23-Nov #PR_RIDกรมชลประทานรายงานสถานการณ์
8,10 9 10 ,23-Nov สน.จรเข้น้อย ให้ความรู้กฎหมายจราจรเบื้องต้น
8,6 1 2 2 ,23-Nov RT @OFFonic 06:50 รถยนต์jazz เฉี่ยวชน
8,1 2 5 0 ,23-Nov ภาพเพลิงไหม้ มอเตอร์เวย์ ขาเข้า กม.80+200