Pre-processing Techniques in Text Mining

Ahmad Fathan Hidayatullah, S.T., M.Cs.

TeA Talk ITSC – Tuesday, October 2, 2018

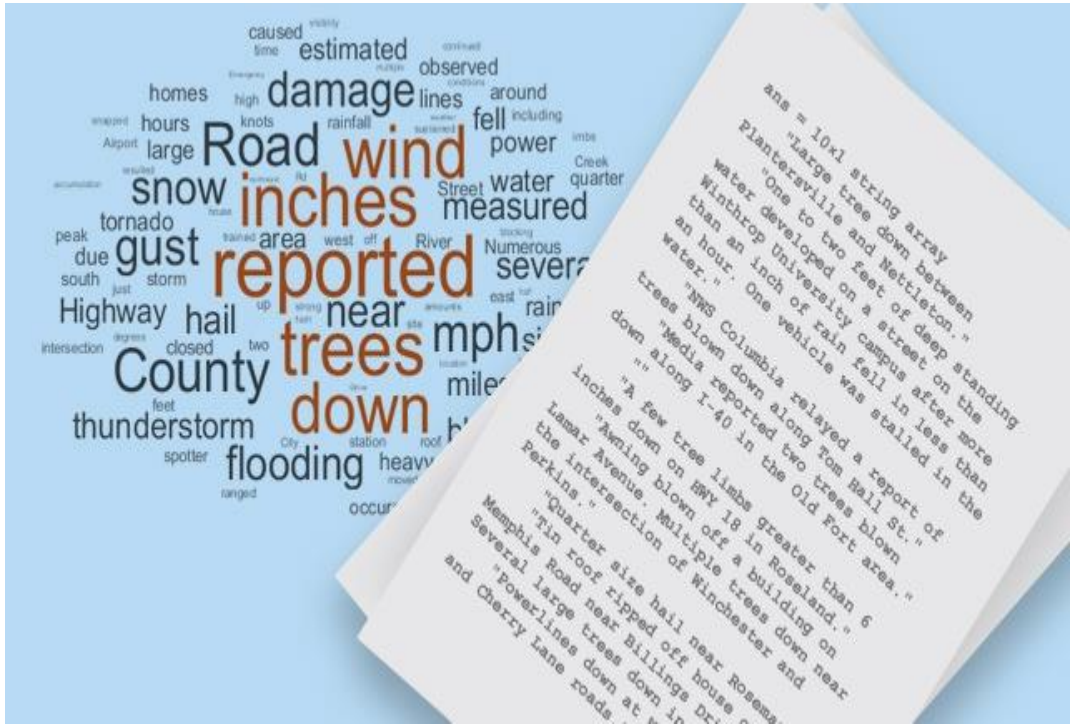
UNIVERSITAS ISLAM INDONESIA

Outline

- Introduction
- Text Mining
- Pre-processing
- Why Pre-processing?
- Text Pre-processing Techniques

Introduction

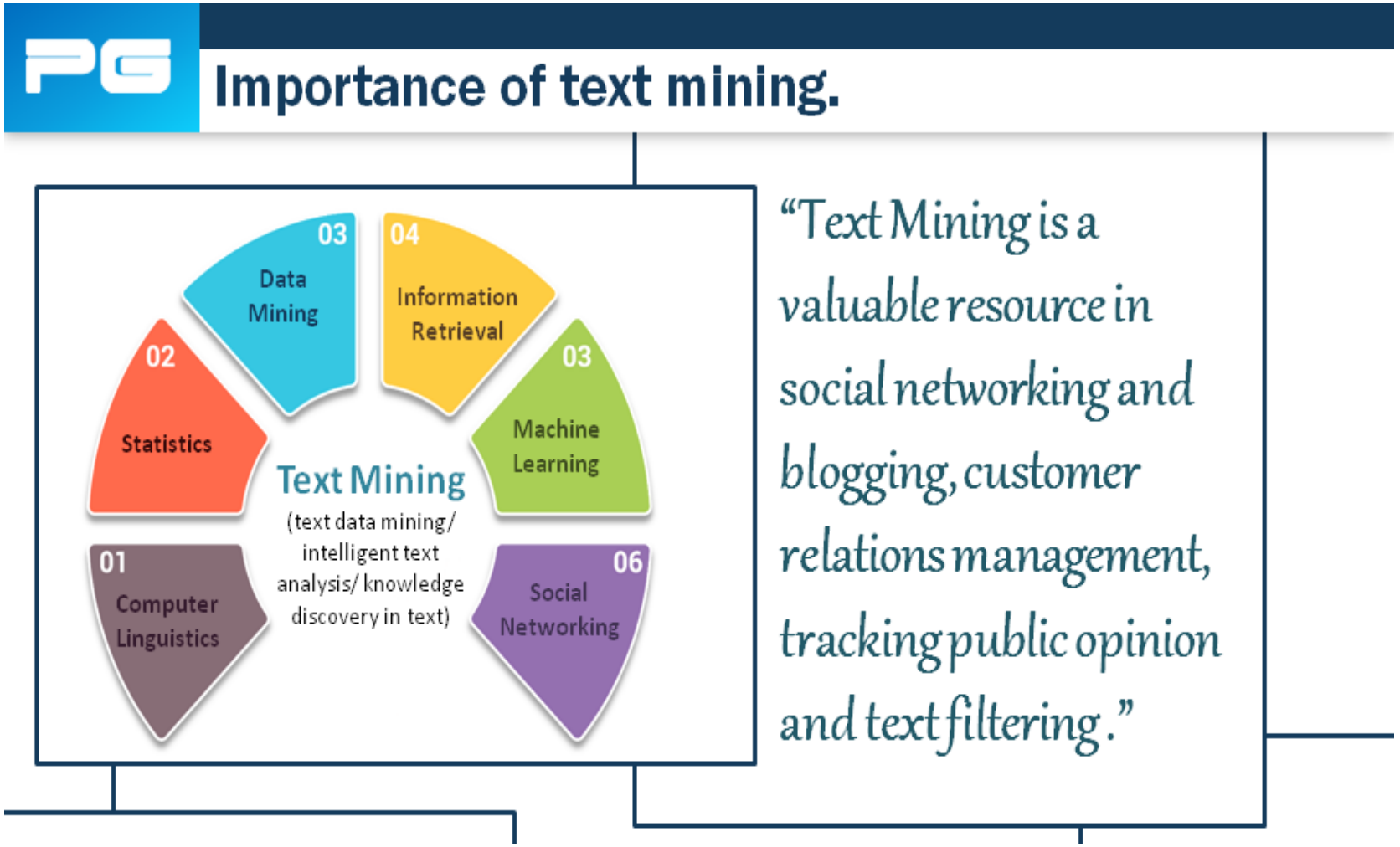
Today, text data are the most dominant data in our daily life



Text Mining

Text mining is a new and exciting area of computer science research that tries to solve the crisis of information overload by combining techniques from data mining, machine learning, natural language processing, information retrieval, and knowledge management. (Feldman & Sanger, 2007)

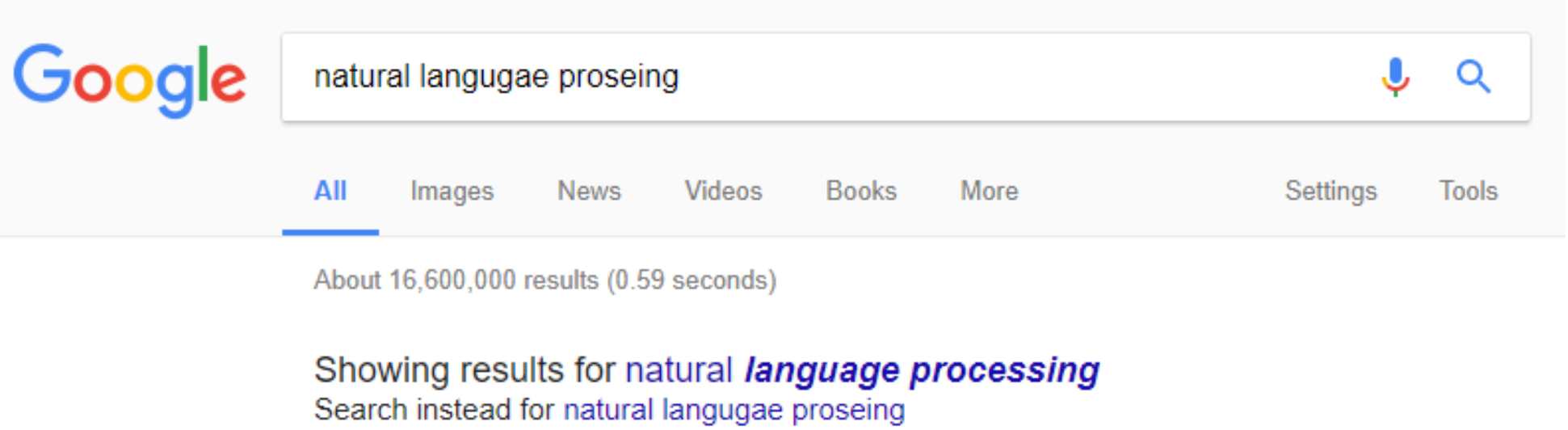
Text Mining



https://www.projectguru.in/publications/wp-content/uploads/2017/07/Slide3.png

Text Mining Application: Spell and Grammar Checking

- Checking spelling and grammar
- Suggesting alternatives for the errors

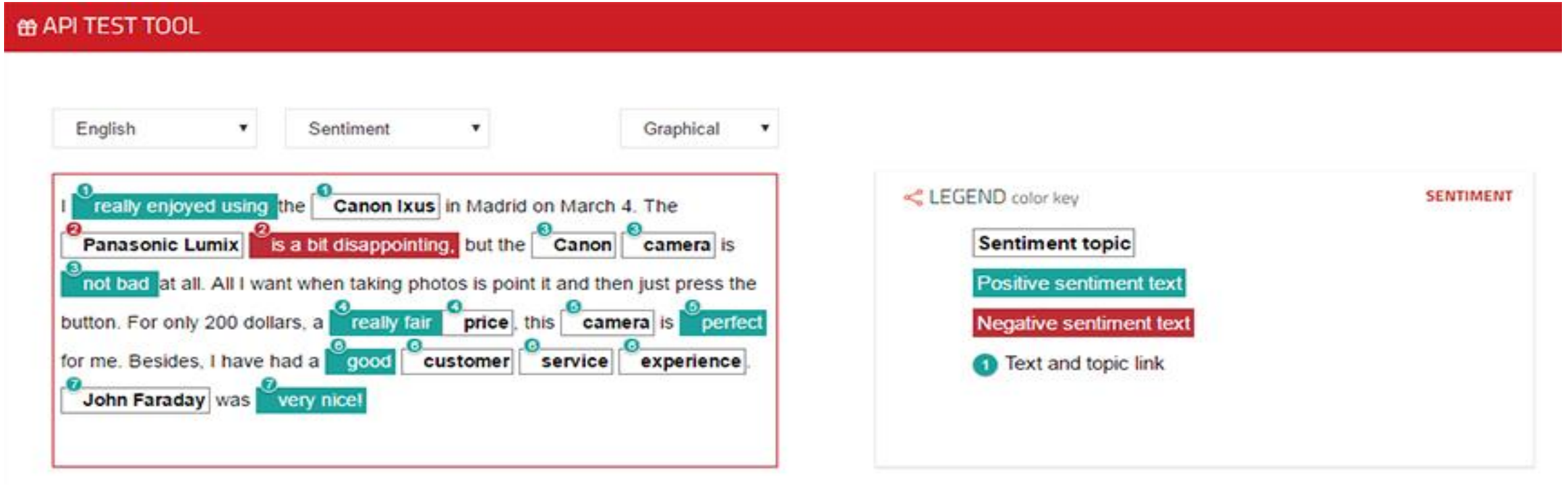


Text Mining Application: Word Prediction

- Predicting the next word that is highly probable to be typed by the user

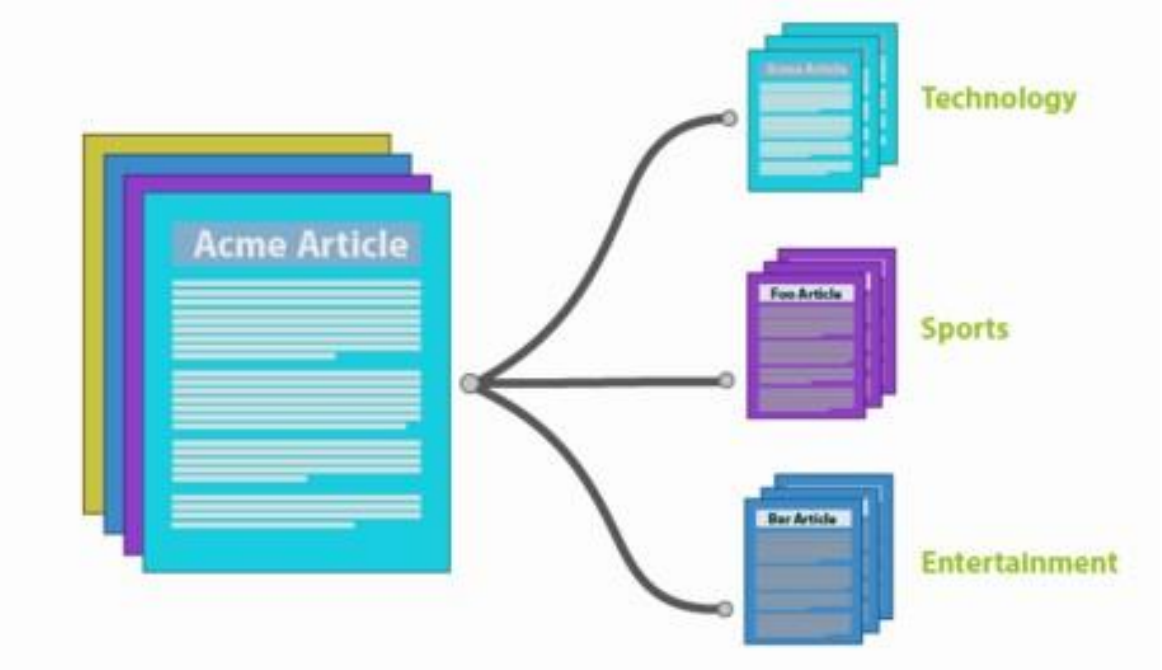
Text Mining Application: Sentiment Analysis

- Identifying sentiments and opinions stated in a text



Text Mining Application: Text Categorization

- Assigning one (or more) pre-defined category to a text



Text Mining Application: Topic Modeling

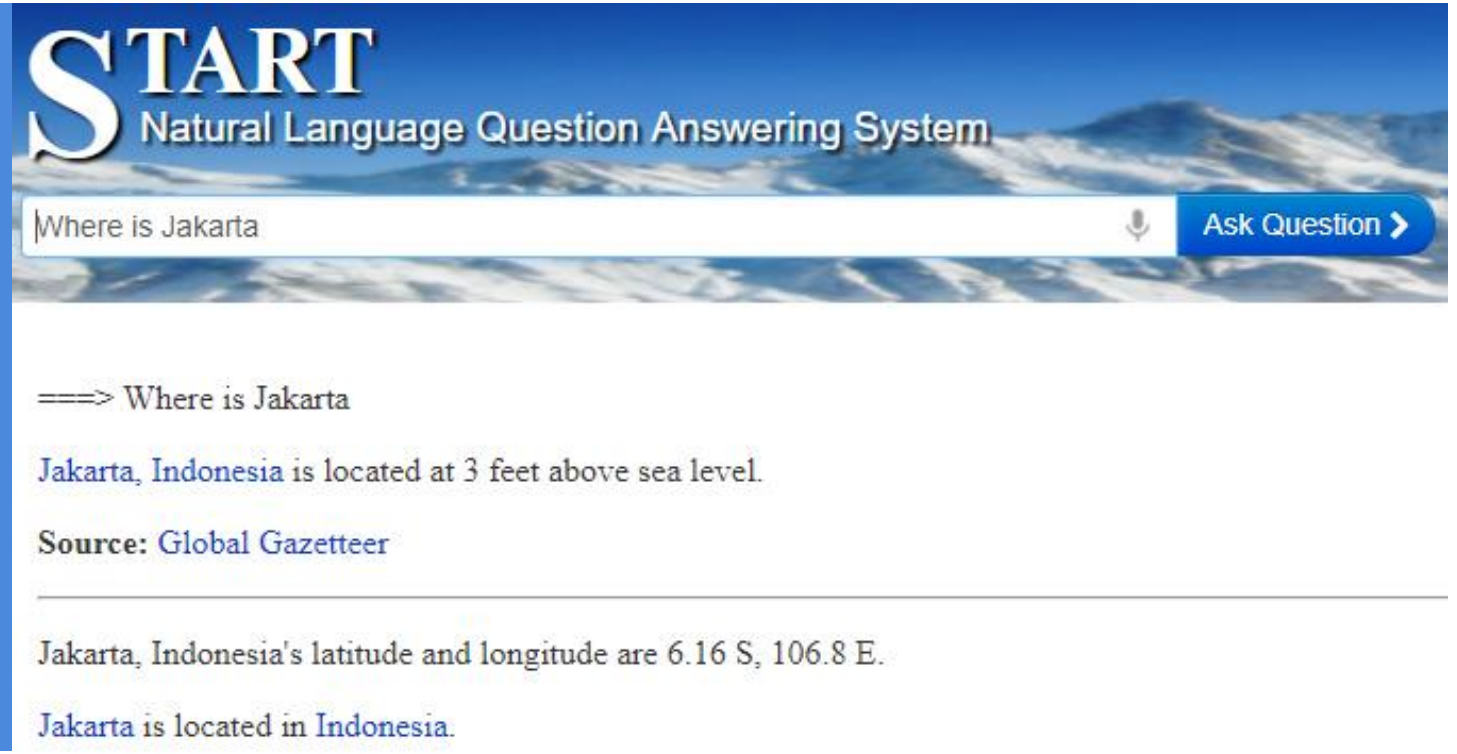
- Discovering the topics that occur in a collection of documents

Real world example: The New York Times

LDA analysis of 1.8M New York Times articles:

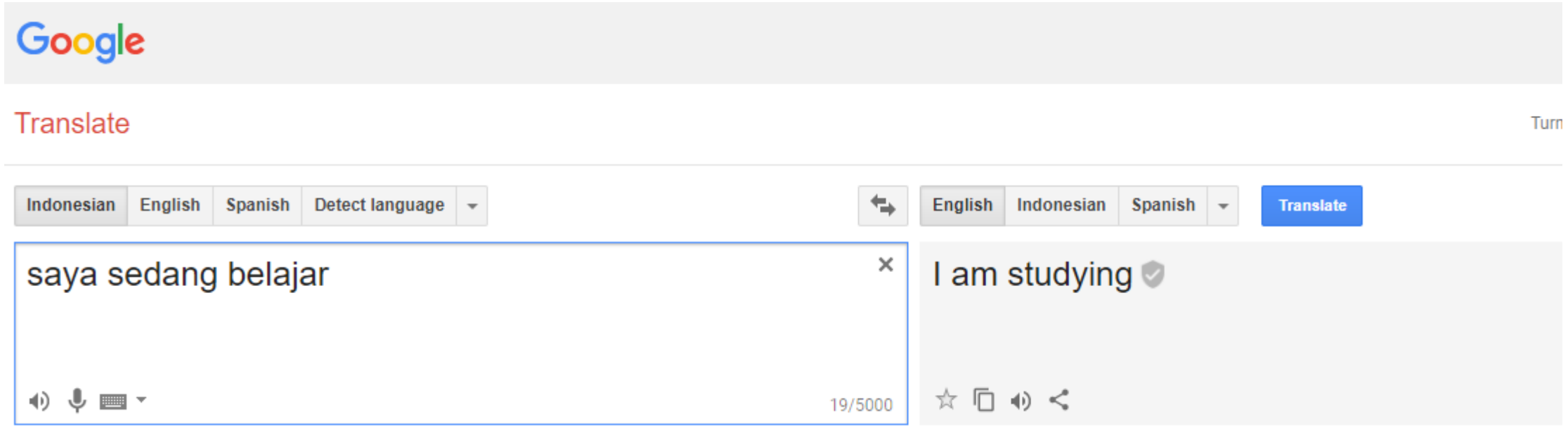
music band songs rock album jazz pop song singer night	book life novel story books man stories love children family	art museum show exhibition artist artists paintings painting century works	game knicks nets points team season play games night coach	show film television movie series says life man character know
theater play production show stage street broadway director musical directed	clinton bush campaign gore political republican dole presidential senator house	stock market percent fund investors funds companies stocks investment trading	restaurant sauce menu food dishes street dining dinner chicken served	budget tax governor county mayor billion taxes plan legislature fiscal

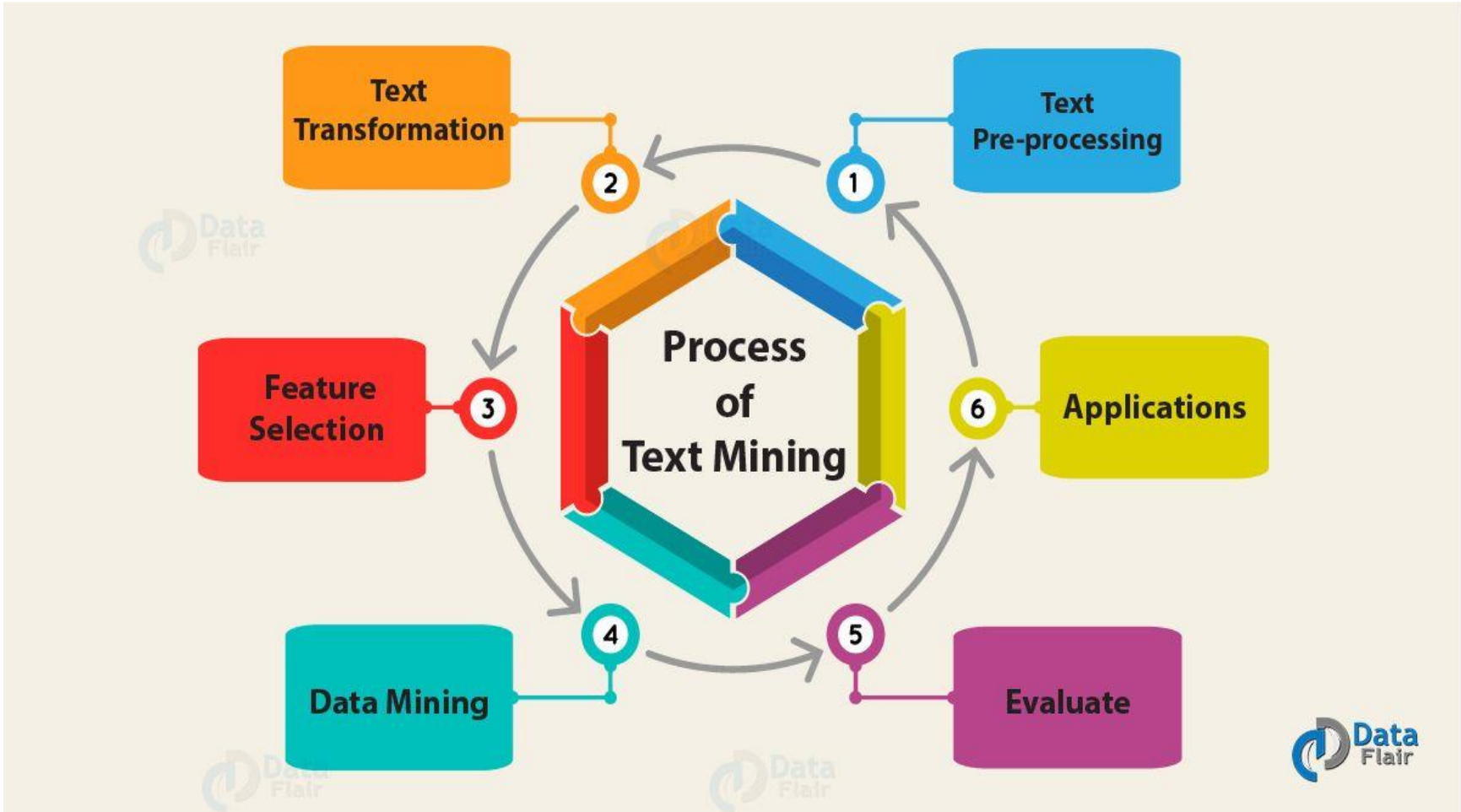
Text Mining Application: Question Answering



Text Mining Application: Machine Translation

- Translating a text from one language to another





https://data-flair.training/blogs/text-mining/

Pre-processing: Definition

- **Text preprocessing** is the task of converting a raw text file, essentially a sequence of digital bits, into a well-defined sequence of linguistically meaningful units: at the lowest level characters representing the individual graphemes in a language's written system, words consisting of one or more characters, and sentences consisting of one or more words.

Pre-processing

Raw Data

Clean Data

Why pre-processing?

- Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogenous sources (Han & Kamber, 2006)

Why pre-processing?

- Especially for social media text, SMS, chat:
  - Social media data is made up of large, noisy, and unstructured datasets
  - The texts are unstructured and are presented in many formats and written by different people in many languages and styles
  - The typographic errors and chat slang have become increasingly prevalent on social networking sites like Facebook and Twitter

Text Pre-processing Techniques

- Tokenization
- Case folding
- Stemming and Lemmatization
- Normalization
- Stopword removal
- Etc.

Tokenization

- Tokenization is the process of breaking a stream of text up into words, phrases, symbols and other meaningful elements called tokens
- Token: It’s a sequence of character that can be treated as a single logical entity
- Type of tokenization:
  - Word tokenization
  - Sentence tokenization

Tokenization: Word Tokenization

```
Example:
This is a test that isn't so simple: 1.23.
"This" "is" "a" "test" "that" "is" "n't"
"so" "simple" ":" "1.23" "."

Issues:
* Finland's capital -
  Finland Finlands Finland's
* what're, I'm, isn't -
  what 're, I 'm, is n't
* Hewlett-Packard or Hewlett Packard
* San Francisco - one token or two?
* m.p.h., PhD.
```

https://www.slideshare.net/vseloved/nlp-project-full-cycle

Tokenization: Language Issues

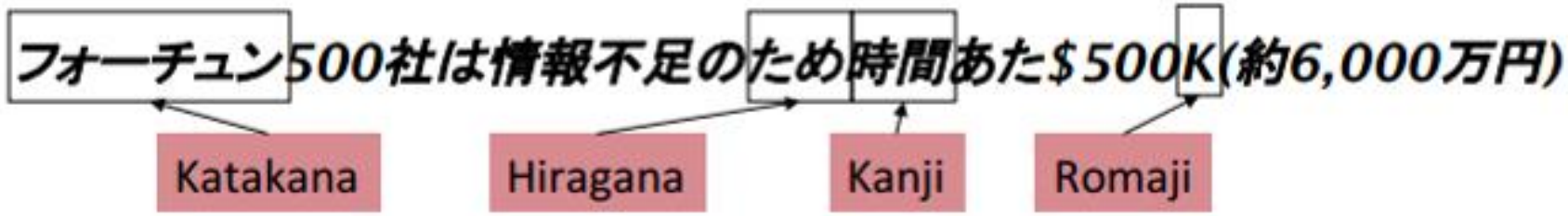
- French
  - ***L'ensemble*** → one token or two?
    - ***L*** ? ***L'*** ? ***Le*** ?
    - Want ***l'ensemble*** to match with ***un ensemble***

- German noun compounds are not segmented
  - ***Lebensversicherungsgesellschaftsangestellter***
  - 'life insurance company employee'
  - German information retrieval needs **compound splitter**

https://web.stanford.edu/class/cs124/lec/textprocessingboth.pdf

Tokenization: Language Issues

- Chinese and Japanese no spaces between words:
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
  - Sharapova now lives in US southeastern Florida
- Further complicated in Japanese, with multiple alphabets intermingled
  - Dates/amounts in multiple formats



End-user can express query entirely in hiragana!

https://web.stanford.edu/class/cs124/lec/textprocessingboth.pdf

Tokenization: Sentence Splitting

- Dividing a string of written language into its component sentences
- In English and some other languages, using punctuation, particularly the full stop/period character (.?!) is a reasonable approximation.
- Non trivial problem, since in English the full stop character also is used for abbreviations or numbers
  - Examples: “Mr.“, “4.5“

Tokenization: Sentence Splitting

Manchester United have agreed a world record deal to sign Paul Pogba for €110 million, **Goal** understands. Officials from the Premier League club met with their Juventus counterparts earlier on Wednesday to discuss a deal to bring Pogba back to Old Trafford. It is now understood that United have settled on a fee of €110m for Pogba, which eclipses the previous record set when Real Madrid paid €100m for Gareth Bale in 2013.

Manchester United have agreed a world record deal to sign Paul Pogba for €110 million, **Goal** understands.

Officials from the Premier League club met with their Juventus counterparts earlier on Wednesday to discuss a deal to bring Pogba back to Old Trafford.

It is now understood that United have settled on a fee of €110m for Pogba, which eclipses the previous record set when Real Madrid paid €100m for Gareth Bale in 2013.

Case Folding

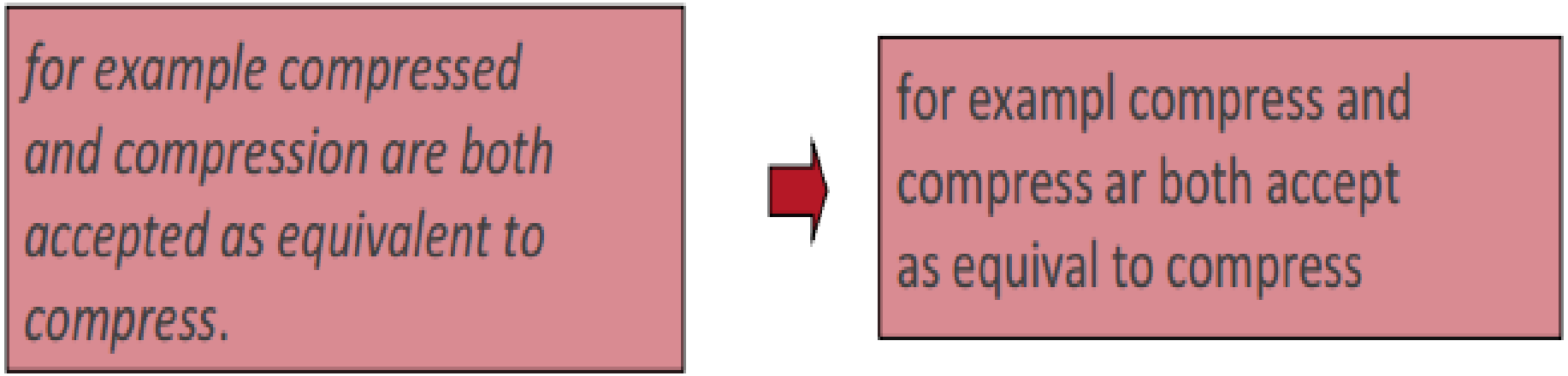
- Applications like IR: reduce all letters to lower case
  - Since users tend to use lower case
  - Possible exception: upper case in mid-sentence?
    - e.g., ***General Motors***
    - ***Fed*** vs. ***fed***
    - ***SAIL*** vs. ***sail***
- For sentiment analysis, MT, Information extraction
  - Case is helpful (***US*** versus ***us*** is important)

https://web.stanford.edu/class/cs124/lec/textprocessingboth.pdf

Stemming

- Stemming is the process of converting the words of a sentence to its non-changing portions
- Stemming tries to find the root words. A root is a word part from which other words grow, usually through the addition of prefixes and suffixes
- Stemming is crude chopping of affixes
  - Language independent
  - E.g., ***automate(s)***, ***automatic***, ***automation***, all reduced to ***automat***

Stemming


for example compressed and compression are both accepted as equivalent to compress.
for exampl compress and compress ar both accept as equival to compress

Stemming Algorithm Example

- Snowball, Lovins, Porter → English
- Nazief Adriani, Porter → Bahasa Indonesia

Lemmatization

- Lemmatization is the process of converting the words of a sentence to its dictionary form
- Lemmatization reduces inflections or variant forms to base form
  - am, are, is → be
  - car, cars, car's, cars' → car

Stemming Vs Lemmatization: Similarity

- The aim of both processes is the same: reducing the inflectional forms and derivations from each word to a common base or root

https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/

Stemming Vs Lemmatization: Difference

Stemming → Lemmatization

➡ token normalization

aka. token "regularization"

(although that is technically the wrong wording)

• Stemming	• Lemmatization
‣ produced by "**stemmers**"	‣ produced by "**lemmatizers**"
‣ produces a word's "stem"	‣ produces a word's "lemma"
‣ am → am	‣ am → be
‣ the going → the go	‣ the going → the going
‣ having → hav	‣ having → have
‣ fast and simple (pattern-based)	‣ requires: a dictionary and PoS
‣ **Snowball**; **Lovins**; **Porter**	‣ **LemmaGen**; **morpha**; **BioLemmatizer**; **geniatagger**

Normalization

- Token normalization is the process of canonicalizing tokens so that matches occur despite superficial differences in the character sequences of the tokens (Stanford IR Book)

Normalization

- Need to "normalize" terms
  - Information Retrieval: indexed text & query terms must have same form.
    - We want to match ***U.S.A.*** and ***USA***
- We implicitly define equivalence classes of terms
  - e.g., deleting periods in a term
- Alternative: asymmetric expansion:
  - Enter: ***window*** Search: ***window, windows***
  - Enter: ***windows*** Search: ***Windows, windows, window***
  - Enter: ***Windows*** Search: ***Windows***
- Potentially more powerful, but less efficient

https://web.stanford.edu/class/cs124/lec/textprocessingboth.pdf

Lexical Normalization in Social Media

User creativity on social media creates a problem for NLP Processing.

I love u -> **i love you**

tmrw -> **tomorrow**

4eva -> **forever**

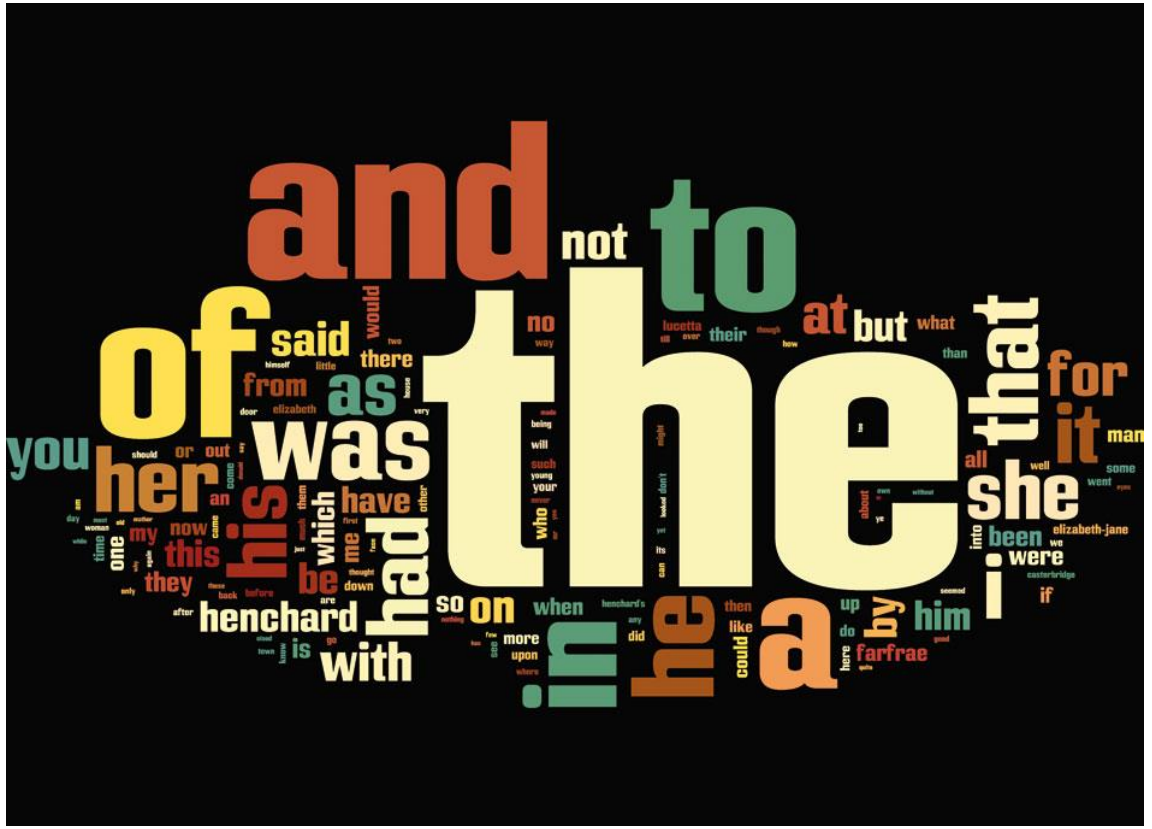
Stopword Removal

- Many of the most frequently used words are useless in IR and text mining – these words are called stopwords
  - the, of, and, to, ….
  - typically about 400 to 500 such words
  - for an application, an additional domain specific stopwords list may be constructed

English Stopword List

Stopword list

a	been	get
about	before	getting
after	being	go
again	between	goes
age	but	going
all	by	gone
almost	came	got
also	can	gotten
am	cannot	had
an	come	has
and	could	ha



Indonesian Stopword List

Peringkat frekuensi kemunculan *

#	Kompas	Wikipedia	Twitter	Kaskus
1	yang	yang	di	gan
2	di	dan	yg	ane
3	dan	di	ya	di
4	ini	pada	aku	yang
5	itu	dari	yang	yg
6	dengan	dengan	ini	ya
7	untuk	ini	itu	ada
8	dari	adalah	ada	itu
9	dalam	dalam	d	tuh
10	akan	untuk	aja	aja

#	Kompas	Wikipedia	Twitter	Kaskus
11	pada	kategori	ga	bisa
12	tidak	tahun	dan	juga
13	juga	sebagai	gak	kalo
14	ke	oleh	i	keren
15	tersebut	indonesia	mau	ga
16	ada	ke	ke	banget
17	bisa	the	udah	nya
18	saat	ia	lagi	wah
19	jakarta	tidak	kalo	nih
20	tahun	menjadi	the	jadi

* Data lengkap: https://github.com/ardwort/freq-dist-id

Why do we need to remove stopwords?

- Reduce indexing (or data) file size
  - stopwords accounts 20-30% of total word counts
- Improve efficiency and effectiveness
  - stopwords are not useful for searching or text mining
  - they may also confuse the retrieval system

Visit My Python Code on Github

- https://github.com/fathanick/text-preprocessing


Question


شكراً جزيلاً
thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
merci
obrigado
dziękuję
bedankt
спасибо
Баярлалаа
рахмат
sukriya
kop khun krap
go raibh maith agat
mochchakkeram
tapadh leat
arigatô
dakujem
grazie
terima kasih
감사합니다
ありがとう
ευχαριστώ
мерси
хвала
dank je
takk
tack
hvala
mauruuru
kiitos
dankie
enkosi
nanni
nandri
dhanyavad
faafetai lava
vinaka
спасибі
blagodaram
mersi
kia ora
barka
welalin
spas
misaotra
matondo
paldies
grazzi
mahalo
asante
manana
obrigada
tenki
murakoze
chokrane
mamnun
дякую
djiere dieuf
tau
akun
dankon
aciü
sulpáy
taiku
gràcies
gratias ago
chnorakaloutioun
köszönöm
bayarlalaa
gràçie
sobodi
dėkuji
mèsi
didi madloba
sagolun
najis tuke
kam sah hamnida
rahmat
tanemirt
rahmet
xièxie
diolch
dhanyavadagalu
shukriya
merçe
trugarez
তোমাকে ধন্যবাদ