MINING TEXT DATA

STQD6414 PERLOMBONGAN DATA

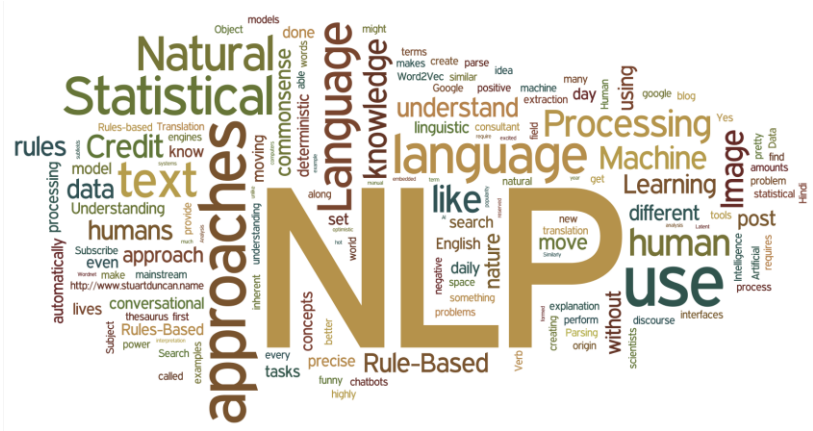
Assoc. Prof. Dr. Nurulkamal Masseran

Department of Mathematical Sciences

Universiti Kebangsaan Malaysia

INTRODUCTION:

- Text data is a type of unstructured data.
- Computer need to interpret the unstructured data to understand human languages.
- Thus, some information can extracted.
- This procedure known as a natural language processing (NLP).
- Text mining uses NLP techniques to transform unstructured data into a structured format for the purpose of identifying meaningful patterns and extracting information.



INTRODUCTION:

- Text data can be generated by many source of platform such as: emails, product reviews, social media posts, newspaper, customer compliant and feedback, document, file, and etc.
- However, common data analysis technique cannot be used to deal with these kind of data.
- Thus, here, text mining plays a major role.

- In particular, the text mining techniques are useful in:

i) identifying trends, popular topics and themes related to some particular issue.

ii) extract sentiment and people's emotions towards some particular issue.

- Example: In businesses, the feedback data from the customers help the companies to get information about the perception and opinions about their product or service.

A TEXT CORPUS:

- A plain text data need to be converted to a corpus format before a data mining analysis can be done.
- A corpus is as a collection of written texts.
- Based on corpus format, a data mining analysis, hypothesis testing, checking occurrences or validating linguistic rules can be done on a text data.

TEXT	CORPUS
Read whole	Read fragmented
Read horizontally	Read vertically
Read for content	Read for formal patterning
Read as a unique event	Read for repeated events
Read as an individual act of will	Read as a sample of social practice
Coherent communicative event	Not a coherent communicative event

DATA CLEANING ON TEXT DATA:

- Commonly, unstructured data does not organize in a proper way.
- Thus, it is difficult to analyze unstructured data directly from its original form.
- Data cleaning on text data is very important task of pre-processing prior to text mining.
- Some important steps of data cleaning on text data:

i) Remove special characters from the text, where a symbol such as; /, @ and | will be replace by a space.

ii) Convert the text to lower case.

iii) Remove numbers

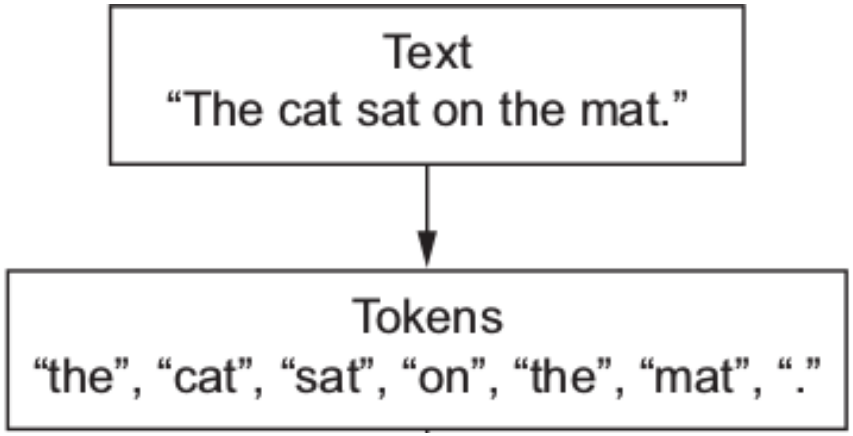
iv) Remove the stopwords in text data. Example: Stopwords in English are “the, is, at, on”. There is no single universal list of stopwords used by all NLP tools.

v) Remove punctuation.

vi) Eliminate extra unnecessary spaces in the text.

WORD TOKENIZATION:

- Tokenization is a technique used to represent text data into a numeric format that can then be used in text mining.
- Word tokenization involves splitting a text into individual words with each unique word being assigned a unique number.
- The tokens could be words, numbers or punctuation marks.
- In tokenization, smaller units are created by locating word boundaries.
- Word boundaries are the ending point of a word and the beginning of the next word.
- The tokenization is a first step for text stemming.



WORD TOKENIZATION:

Example:

- Consider the sentence, “I Love my cat”. In tokenization, unique integer value will be assign to each unique word in that sentence.
- Such that, 1 to “I”, 2 to “Love”, 3 to “my” and 4 to “cat”.
- If we have another sentence: “I Love my car”, then, the words “I Love my”, already have numbers 001 002 003.
- Thus, only a new word will be assign a new unique integer value. Here, we have integer 005 is assign for the word “car”.
- The tokens of these two sentences are: 001 002 003 004, 001 002 003 005
- Using tokenization, similarity between the sentences can be evaluated.
- Intrinsically, a computer does not understand text or language in a human sense.
- By tokenization, computer able to transform text from a human-understandable form to a statistical pattern that can be mapped by a data mining technique.

TEXT STEMMING:

- Text stemming is the process of reducing the word to its root form.
- The stemming simplifies the word to its common origin.
- Stemming is important technique used in NLP to reduces the number of computations required.
- In a similar vein, stemming is also useful in reducing the dimensionality of a text data.
- Example: the stemming process reduces the words “fishing”, “fished” and “fisher” to its stem “fish”.
- The words are different but they actually similar in term of contextually.

	original_word	stemmed_words
0	connect	connect
1	connected	connect
2	connection	connect
3	connections	connect
4	connects	connect

DOCUMENT- TERM MATRIX:

- A document-term matrix represents the relationship between terms and documents.
- The rows stand for a specific document or sentence.
- The columns represent a unique word.
- The entry in this matrix represent the number of occurrences of the term in the document.
- A document-term matrix also referred as a table containing the frequency of words.

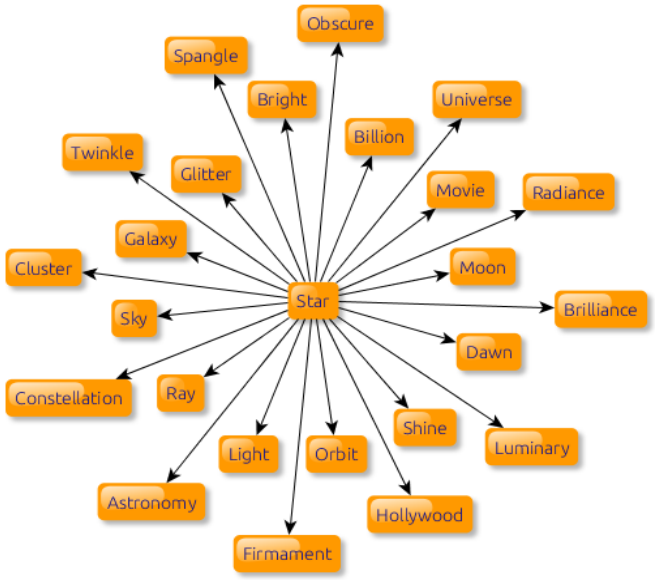
	text	mining	is	to	find	useful	information	from	text	mined	dark	came
D1	1	1	1	1	1	1	1	1	1	0	0	0
D2	0	0	1	0	0	1	1	1	1	1	0	0
D3	0	0	0	0	0	0	0	0	0	0	1	1

WORD CLOUD:

- A word cloud is a visual representation of words.
- Word cloud use the information provided by word frequency table represented by document-term matrix.
- Word cloud are useful in highlighting the popular words and phrases based on the frequency and its relevance in a text data.
- Based on word cloud visualization, more in-depth analyses can be carried out.
- The word cloud image composed of keywords found within a body of text.
- The size of each word indicates its frequency in a text data.

WORD ASSOCIATION:

- Word association is a technique to analyzing the content of text data by determining a significant relationship between terms.
- This technique compute similarity among each words in context documents, after collecting the context through a bag of words
- Correlation measure is used to determine how strong a magnitude of pairs are related.
- Word association can be illustrated using word similarity chart and word correlation graph chart.



SENTIMENT ANALYSIS:

- Sentiment Analysis is a process of extracting opinions that have different scores like positive, negative or neutral.
- Based on sentiment analysis, you can find out the nature of opinion or sentences in text.
- Sentiment Analysis is a type of classification where the data is classified into different classes like positive or negative or happy, sad, angry, etc.
- Sentiment analysis is used for many applications, especially in business intelligence.

- For example:

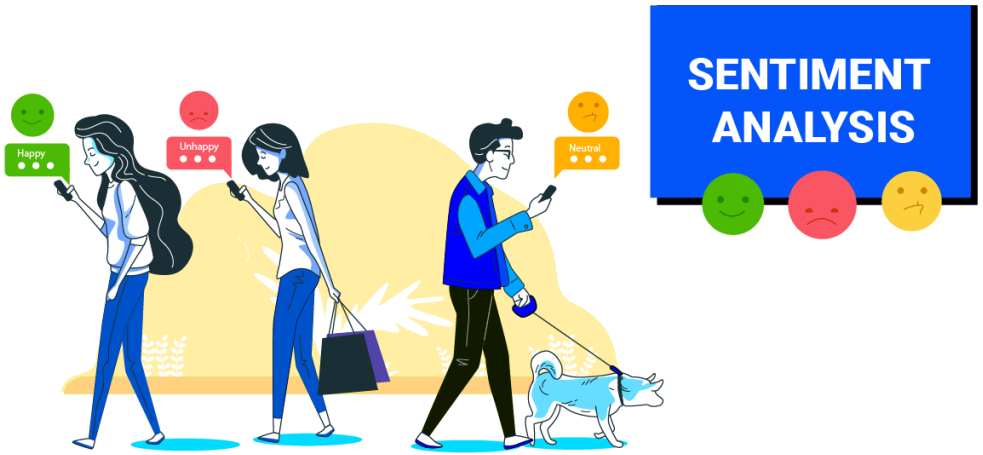
i) Analyzing the social media discussion around a certain topic.

ii) Evaluating survey responses.

iii) Determining whether product reviews are positive or negative.

SENTIMENT ANALYSIS:

- However, sentiment analysis cannot tell you why some people are feeling a certain way.
- In spite of that, sentiment analysis provides an information about the words associated with strongly positive or negative sentiment.
- This information is being derived by a count of the number of positive and negative words in the text.
- Then, the analysis will be conducted to characterize this mix of positive and negative words.



SENTIMENT ANALYSIS:

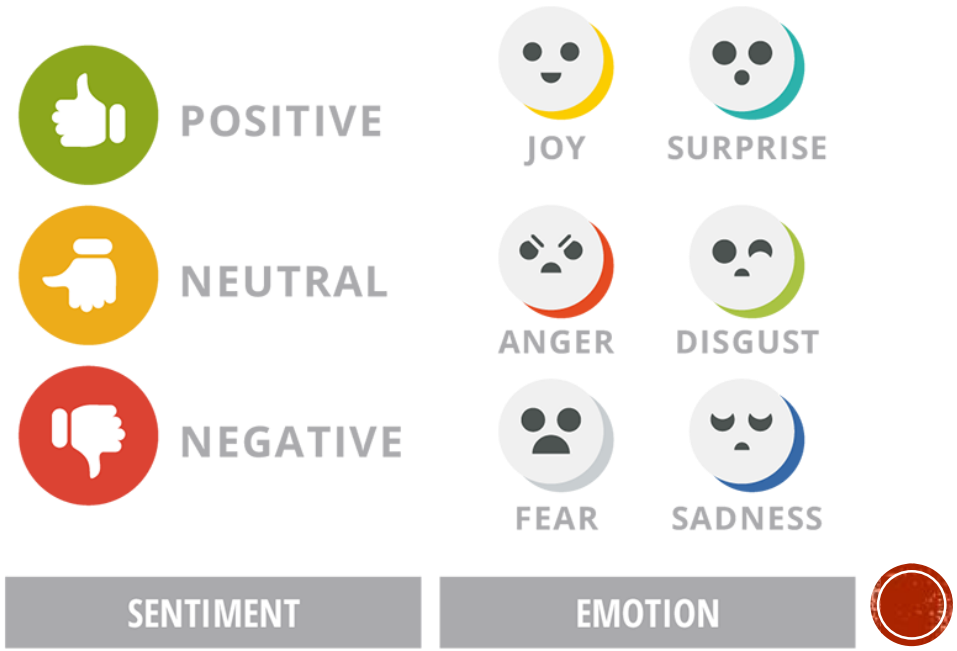
- The first step in sentiment analysis is creating a lexicon text.
- Lexicon refer to a word list.
- In default, some lexicons are already exists.
- However, if your text is having some a specific topic, your lexicon need to add to or modify in prior to sentiment analysis

Lexicon	Positive Words	Negative Words
Simplest (SM)	good	bad
Simple List (SL)	good, awesome, great, fantastic, wonderful	bad, terrible, worst, sucks, awful, dumb
Simple List Plus (SL+)	good, awesome, great, fantastic, wonderful, best, love, excellent	bad, terrible, worst, sucks, awful, dumb, waist, boring, worse
Past and Future (PF)	will, has, must, is	was, would, had, were
Past and Future Plus (PF+)	will, has, must, is, good, awesome, great, fantastic, wonderful, best, love, excellent	was, would, had, were, bad, terrible, worst, sucks, awful, dumb, waist, boring, worse
Bing Liu	2006 words	4783 words
AFINN-96	516 words	965 words
AFINN-111	878 words	1599 words
enchantedlearning.com	266 words	225 words
MPAA	2721 words	4915 words
NRC Emotion	2312 words	3324 words

EMOTION CLASSIFICATION:

- Instead of two sentiments (negative and positive), the emotion classification is also providing useful information in sentiment analysis.
- Emotion classification is built on the NRC Word-Emotion Association Lexicon
- The NRC Emotion Lexicon is a list of English correspond to eight basic emotions:

i) Anger
ii) Fear
iii) Anticipation
iv) Trust
v) Surprise
vi) Sadness
vii) Joy
viii) Disgust



REFERENCES:

- Aggarwal, C. C., Zhai, C. (2012). *Mining Text Data*. Springer.
- Kwartler, T. (2017). *Text Mining in Practice with R*. Wiley.
- Lamba, M., Madhusudhan, M. (2022). *Text Mining for Information Professionals: An Uncharted Territory*. Springer.
- Silge, J., Robinson, D. (2017). *Text Mining with R: : A Tidy Approach*. O'Reilly Media, Inc.
- Zhai, C., Massung, S. (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. ACM Books.
- Žižka, J., Darena, F., Svoboda, A. (2021). *Text Mining with Machine Learning*. CRC Press.
- Zong, C., Xia, R., Zhang, J. (2021). Text Data Mining. Springer.

NEXT TOPIC:

Mining Spatial Data