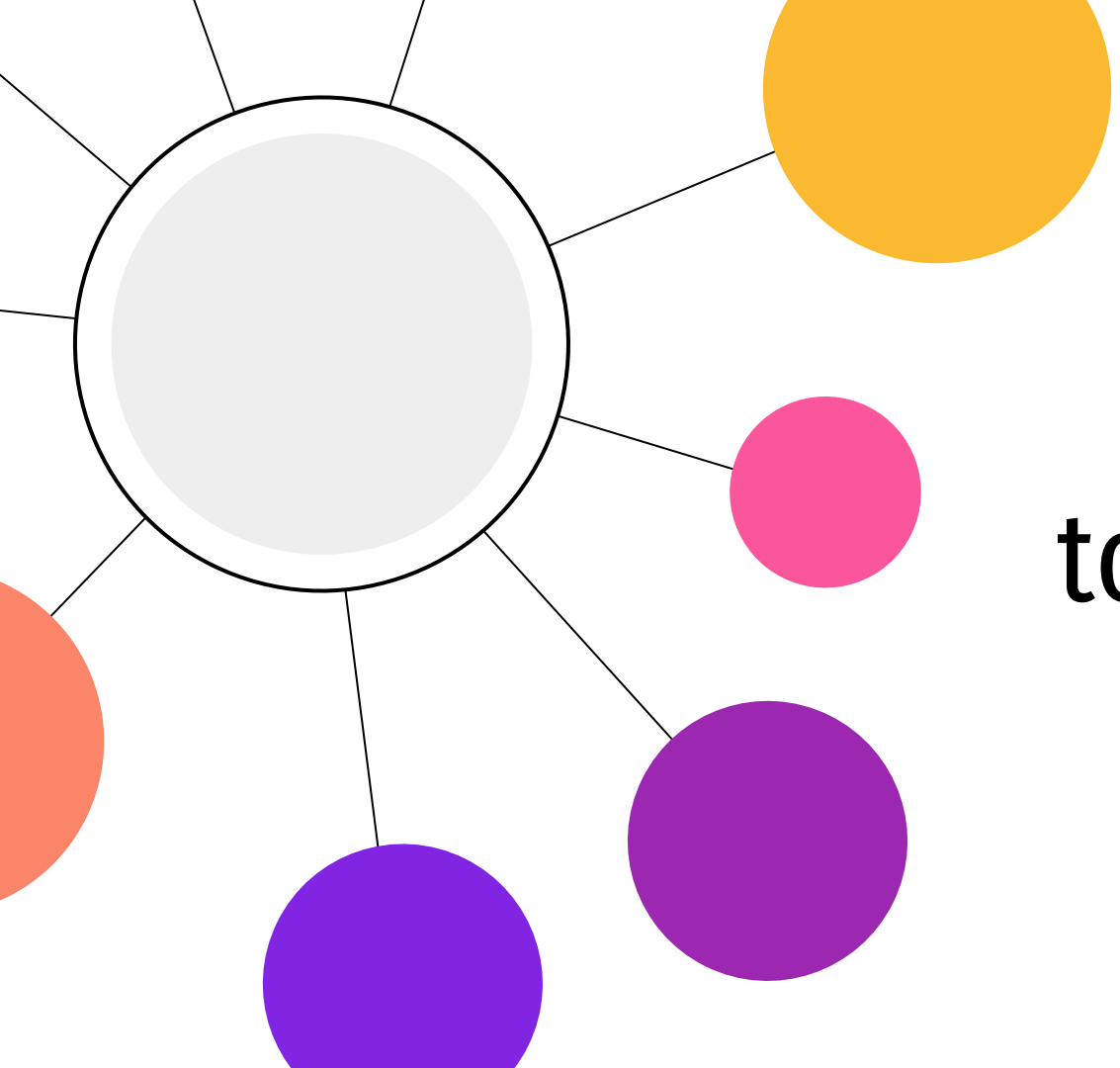
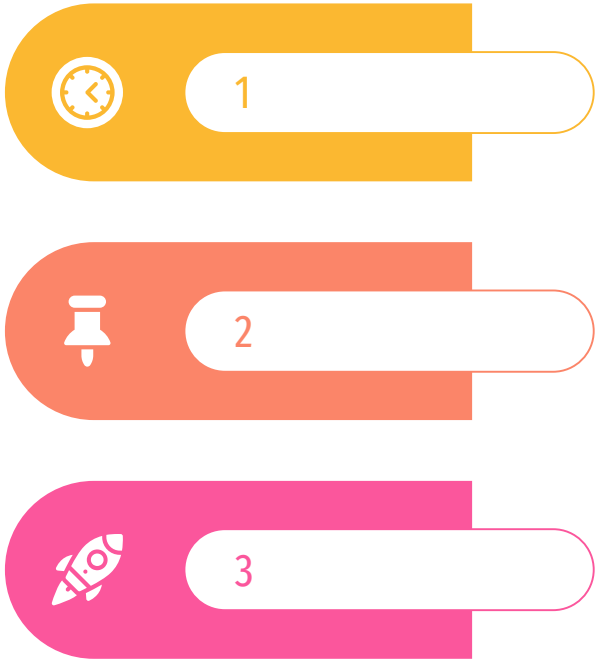


Introduction to Text Mining and NLP

Hendri Prabowo
Data Fellowship 6 IYKRA

Outline

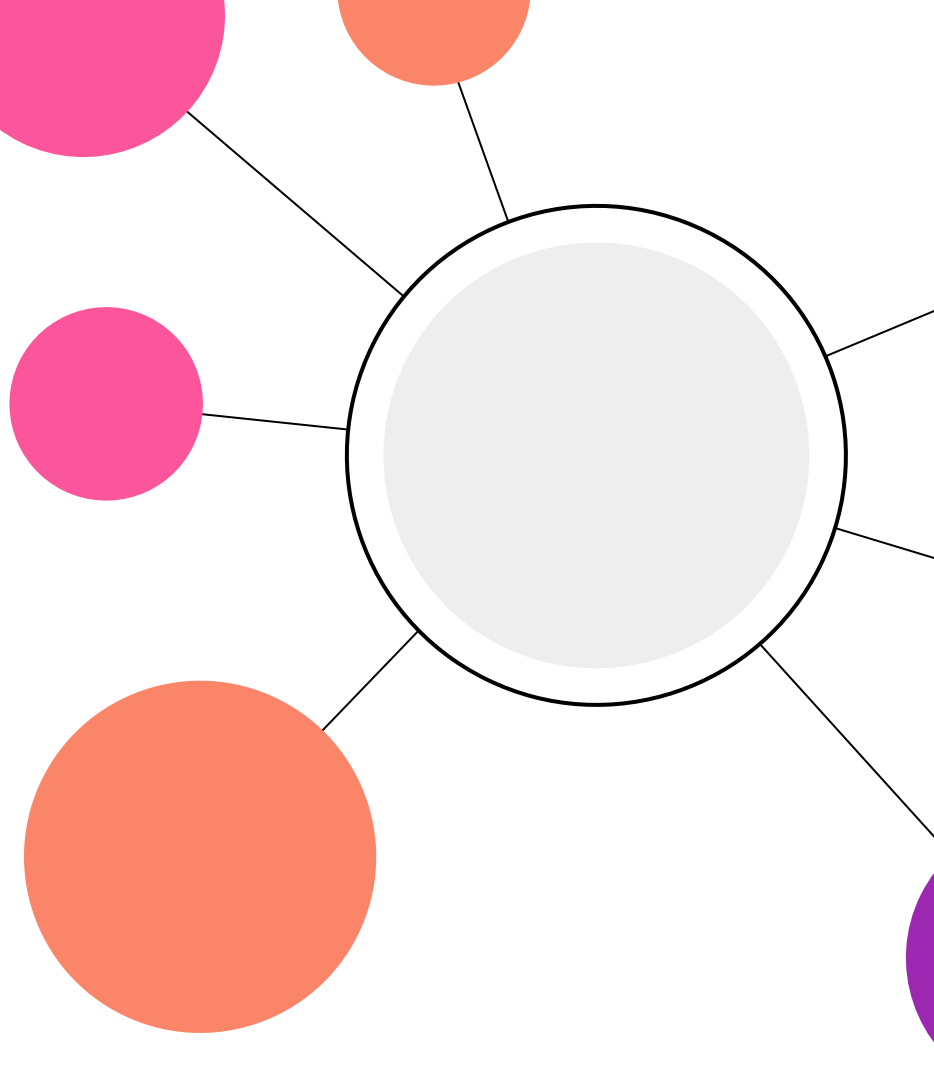


Text Mining and NLP

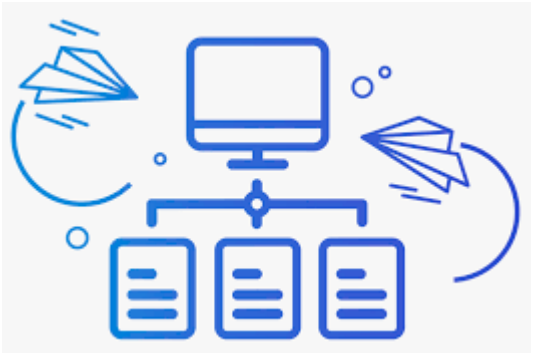
Text Pre-processing

Feature Engineering with Text

Text Mining and NLP



Source of Text Data



Sources of text data include:

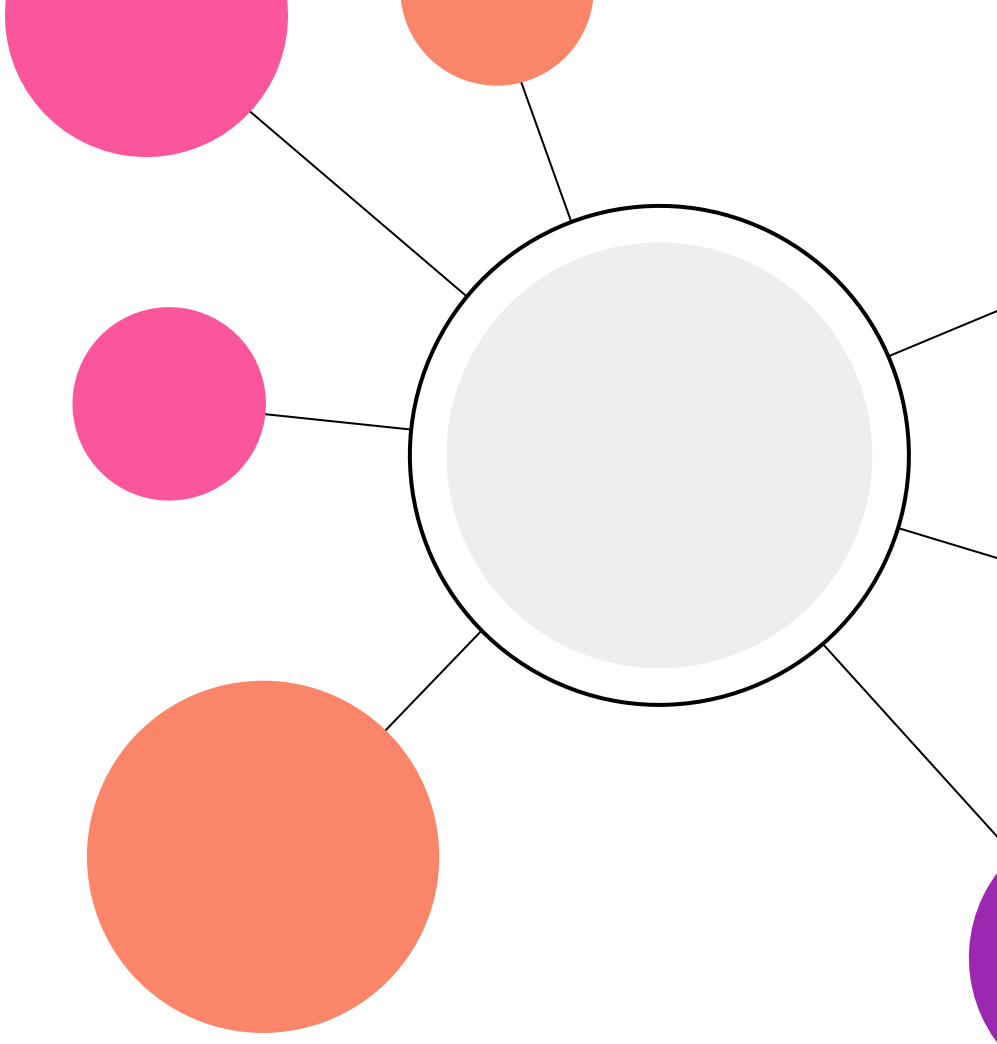
- Document
- Social media
- Web scraping
- Transcription of audio/video data
- Etc.

Text Mining vs. NLP

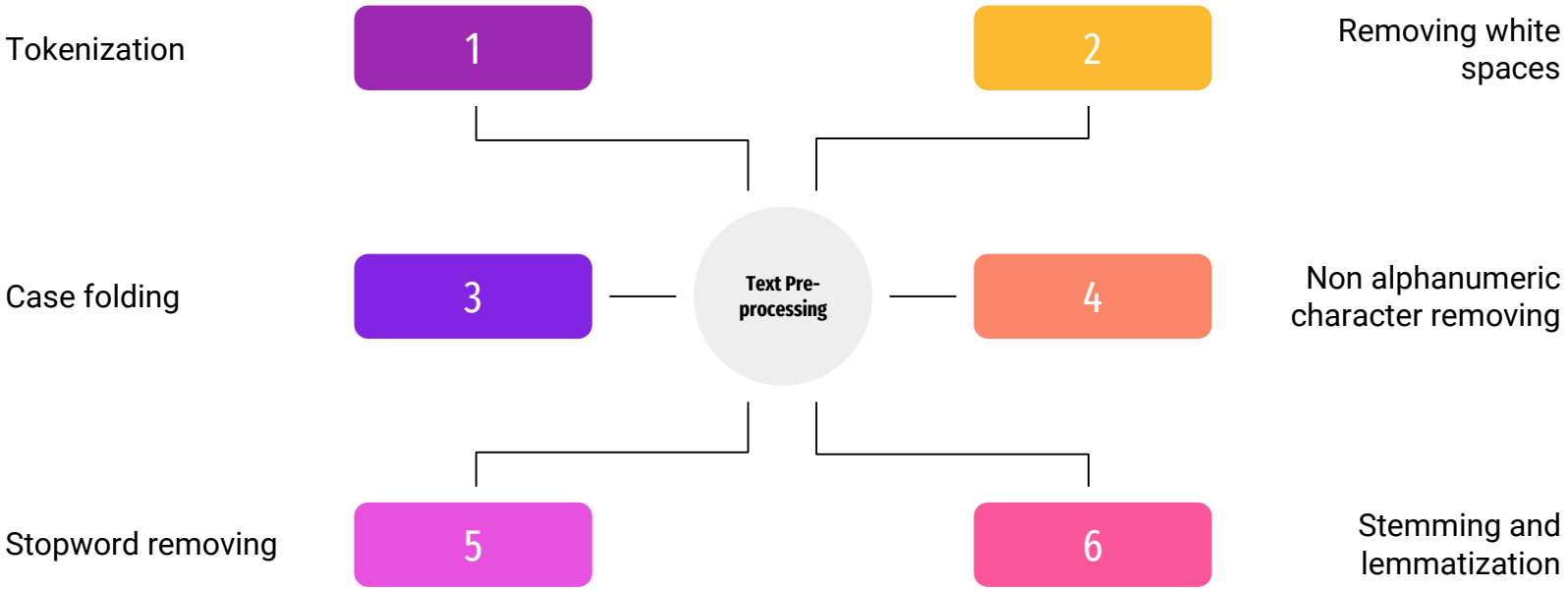
Text mining is the process or practice of examining large collections of written resources in order to generate new information. The goal of text mining is to discover relevant information in text by transforming the text into data that can be used for further analysis.

Natural language processing (**NLP**) is a component of text mining that performs a special kind of linguistic analysis that essentially helps a machine “read” text. NLP uses a variety of methodologies to decipher the ambiguities in human language.

Text Pre-processing



Text Pre-processing



Case Folding

Every character is unique, lower and uppercase character is different

Example:
Input:
'Undang @NShaniJKT48 ke hitamputih, pemenang SSK JKT48 harusnya mJKT48 ini lebih Layak di Undang karena prestasinya‘

Output:
'undang @nshanijkt48 ke hitamputih, pemenang ssk jkt48 harusnya mjkt48 ini lebih layak di undang karena prestasinya'

Remove White Spaces

Remove white spaces is a process to remove spaces in front of or back of the sentence

Example:
Input:
‘ undang @nshanijkt48 ke hitamputih, pemenang ssk jkt48 harusnya mjkt48 ini lebih layak di undang karena prestasinya'

Output:
'undang @nshanijkt48 ke hitamputih, pemenang ssk jkt48 harusnya mjkt48 ini lebih layak di undang karena prestasinya'

Tokenization

Tokenization is a process to split a sentence into words

Example:
Input:
‘undang @nshanijkt48 ke hitamputih, pemenang ssk jkt48 harusnya mjkt48 ini lebih layak di undang karena prestasinya'

Output:
['undang', '@nshanijkt48', 'ke', 'hitamputih,', 'pemenang', 'ssk', 'jkt48', 'harusnya', 'mjkt48', 'ini', 'lebih', 'layak', 'di', 'undang', 'karena', 'prestasinya']

Non Alphanumeric Character Removing

Commonly, special characters and symbols (ex : ?#@!.) will add noises into unstructured text

Example:

Input:

['undang', '@nshanijkt48', 'ke', 'hitamputih,', 'pemenang', 'ssk', 'jkt48', 'harusnya', 'mjkt48', 'ini', 'lebih', 'layak', 'di', 'undang', 'karena', 'prestasinya']

Output:

['undang', 'nshanijkt48', 'ke', 'hitamputih,', 'pemenang', 'ssk', 'jkt48', 'harusnya', 'mjkt48', 'ini', 'lebih', 'layak', 'di', 'undang', 'karena', 'prestasinya']

Stopword Removing

Stopword is a word that has appear frequently but has no significant meaning. Ex : And, The, Or, Therefore, etc. Not removing stopwords may lead to the bias

Example:

Input:

['undang', 'nshanijkt48', 'ke', 'hitamputih,', 'pemenang', 'ssk', 'jkt48', 'harusnya', 'mjkt48', 'ini', 'lebih', 'layak', 'di', 'undang', 'karena', 'prestasinya']

Output:

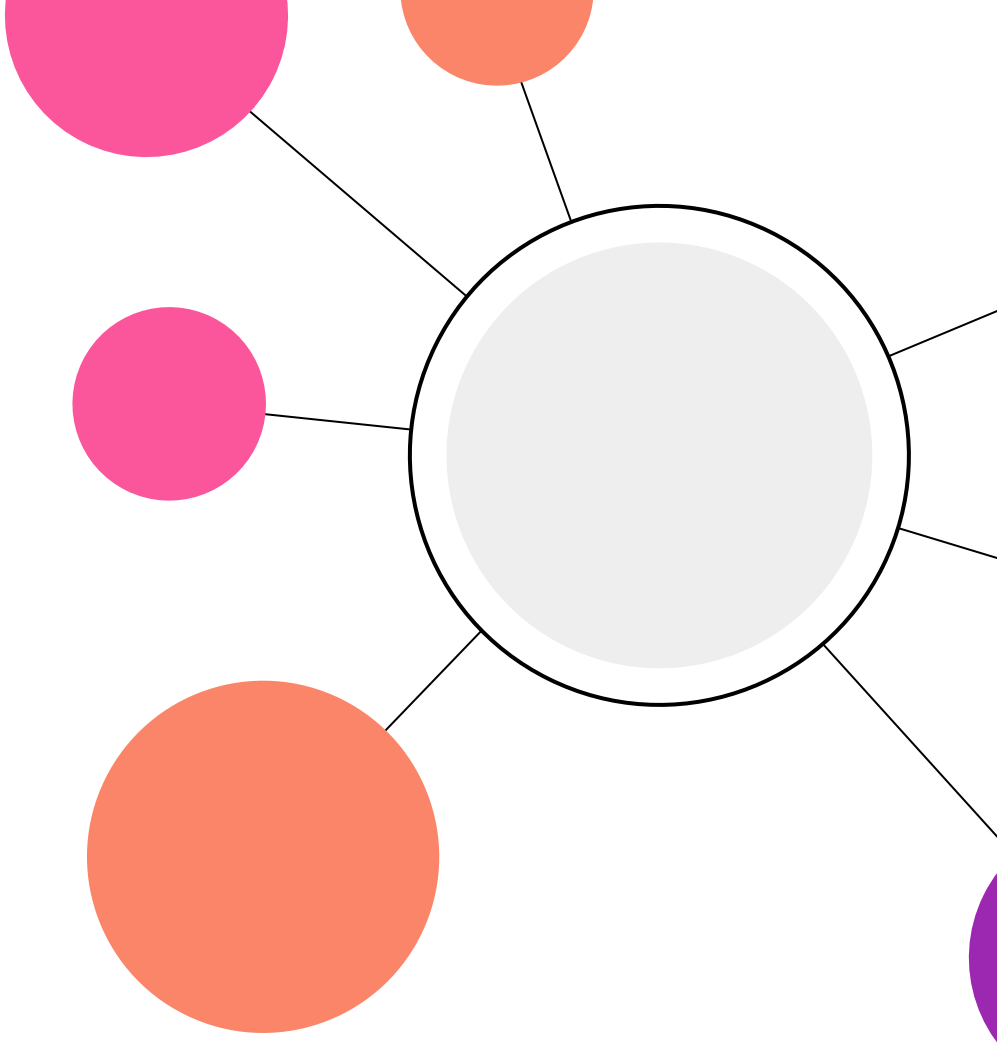
['undang', 'nshanijkt48', 'hitamputih,', 'pemenang', 'jkt48', 'mjkt48', 'layak', 'undang', 'prestasinya']

Stemming and Lemmatization

Stemming and lemmatization both generate the root form of the inflected words. The difference is that stem might not be an actual word whereas, lemma is an actual language word.

Example:
Studies (Word) -> Studi (Stemming) -> Study (Lemmatization)

Feature Engineering with Text



Feature Engineering with Text

1

TF-IDF

2

Bag of Words

3

Word Embedding

4

Named Entity Relation (NER)

5

Part of Speech - Tagging

Bag of Words

A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:

1. A vocabulary of known words.
2. A measure of the presence of known words.

Bag of Words

	16th	2019	61	9tahuntvone	aa	aagym	abang	abas	abi	abis	...	ya	yaampun	yadia	yag	yg	younglex	yuk	yukikatou2	zaitun	zhonk
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
395	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
396	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
397	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
398	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
399	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

TF-IDF

TF-IDF stands for “Term Frequency - Inverse Document Frequency”. This is a technique to quantify a word in documents, we generally compute a weight to each word which signifies the importance of the word in the document and corpus.

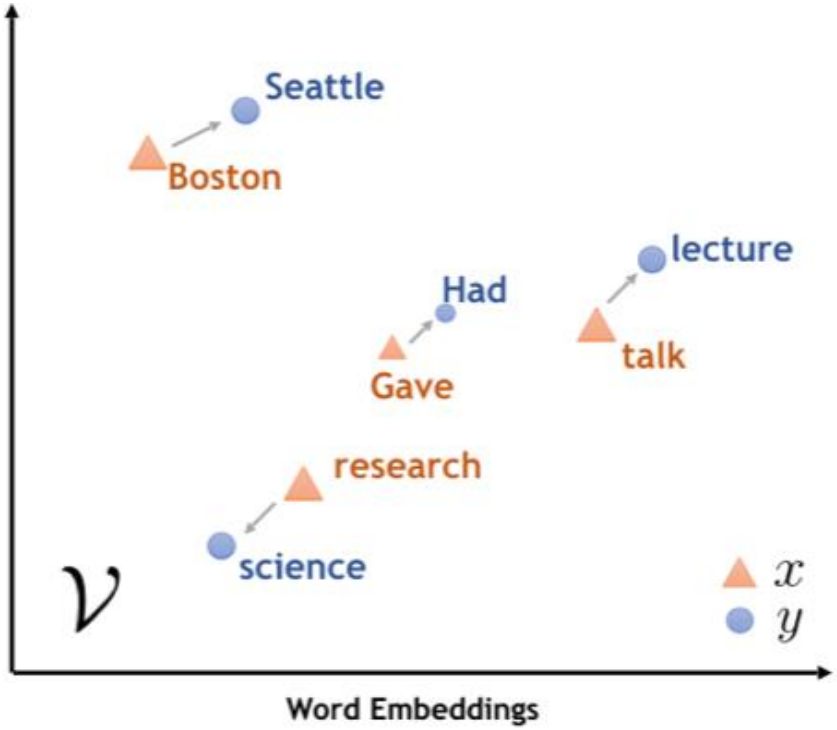
$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

TF-IDF

	16th	2019	61	9tahuntvone	aa	aagym	abang	abas	abi	abis	...	ya	yaampun	yadia	yag	yg	younglex	yuk	yukikatou2	zaitun	zhonk
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.339473	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
395	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
396	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
397	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
398	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
399	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Word Embedding



Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation.

Named Entity Relation (NER)

Task of identifying and categorizing key information (entities) in text. Classify words into its predefined categories, such as person, organizations, locations, expressions of times, quantities, percentages, etc.

Example:
Input:
"arahan menteri Desa dalam penangan Covid"

Output:
(Desa,)
Entities [('Desa', 'ORGANIZATION')]

Part of Speech - Tagging

Classify the word into their part of speech based on its definition and its context such as Subject, Noun, Object, Verb, so on.

Example:

Input:

"arahan menteri Desa dalam penangan Covid"

Output:

arahan <NOUN> menteri <NOUN> desa <NOUN> dalam <ADP> penangan <NOUN> covid <PUNCT>

Thank You