2nd November, 2020

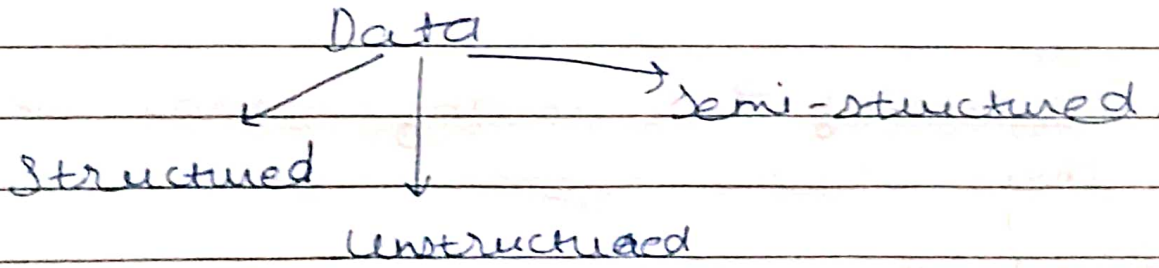
Ch - 2
Basic Feature Extraction Methods.

⇒ Objectives :

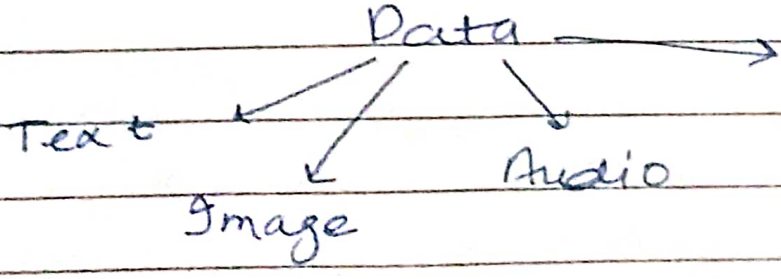
- Categorize the data based on content and structure
- Describe pre-processing steps in detail and implement them to clean data
- Describe feature engineering
- Calculate the similarity between texts.
- Visualize text using word cloud and other visualization techniques.

⇒ Types of Data

- Categoring Categorizing data by structure



- Categoring Data by Content.



⇒ Cleaning Text Data

Art of extracting meaningful portions from data by removing unnecessary portions.

- Regular expression
  Expressions which help searches for a pattern in text data.

- Types of Tokenizers:

i) Tweet tokinezer : used for tweet
ii) MWE tokenizer : multi-word expression
iii) Regular expression tokenizer : Uses RE
iv) whitespace tokenizer
v) word Punct tokenizer

⇒ Stemming → Converting words to their base form. stems.

- Reg exp stemmer
- Porter stemmer.

22-11-2020

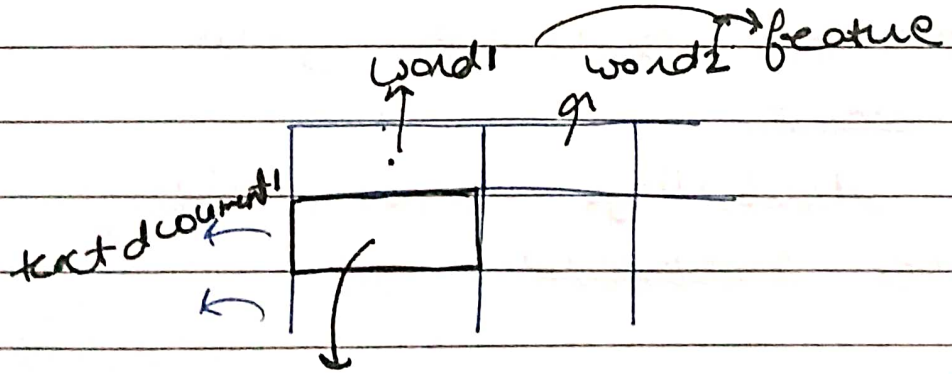
PAGE NO

Feature Extraction from text

- General feature
  - len of words
  - words occurances
  - language used
- Unique feature
  - Bag of words
  - Tf-idf

⇒ Bag of words

- Most simplest way for extracting features from raw text



1-12-2020

TF-IDF

(Term frequency - inverse document freq)

- BOW model did not conveyed how much information a given word is conveying.
- TFIDF is a method of representing text data in a matrix format using numbers that quantify how much information these terms carry in the given docs.

k for a given term j, in a document i,

tf(j) = # time j appears in i = tf

idf(j) = $\log_{10}(N/df_j)$

(df j → # document in which j appears)

11-12-2020

Feature Engineering

- Extracting new features from existing one's.

For eg: calculating how similar two given sentences are.

This can be calculated by

i) Cosine similarity →
cos of angle b/w the vector representations of 2 texts.

ii) Jaccard similarity →
common terms / # total terms