

Text Data Mining

Introduction

- ❖ Text mining is the process of deriving high-quality information from the text.
- ❖ “High Quality” in text mining usually refers to some combination of novelty and interest.



Techniques used in Text Mining

❖ **Information Extraction(IE)**

- This is the most famous text mining technique. Information extraction refers to the process of extracting meaningful information from vast chunks of textual data. This
- Text mining technique focuses on extraction of entities, attributes, and their relationships from semi-structured or unstructured texts. Whatever information is extracted is then stored in a database for future access and retrieval.
- The efficacy and relevancy of the outcomes are checked and evaluated using precision and recall processes.



Cont...

❖ **Natural Language Processing(NLP)**

- Natural Language Processing includes both Natural Language Understanding and Natural Language Generation, which simulates the human ability to create natural language text e.g. to summarize information or take part in a dialogue.



Cont...

❖ Information Retrieval

- Information Retrieval (IR) refers to the process of extracting relevant and associated pattern based on a specific set of words or phrases.
- IR systems make use of different algorithms to track and monitor user behaviors and discover relevant data accordingly. Google and Yahoo search engines are the two most renowned IR systems.



Cont...

❖ Clustering

- Clustering is one of the most crucial text mining techniques. It seeks to identify intrinsic structures in textual information and organize them into relevant subgroups or 'clusters' for further analysis.
- A significant challenge in the clustering process is to form meaningful clusters from the unlabeled textual data without having any prior information on them.
- Cluster analysis is a standard text mining tool that assists in data distribution or acts as a pre-processing step for other text mining algorithms running on detected clusters.



**TEXT
MINING
PROCESS**

Text Pre-processing

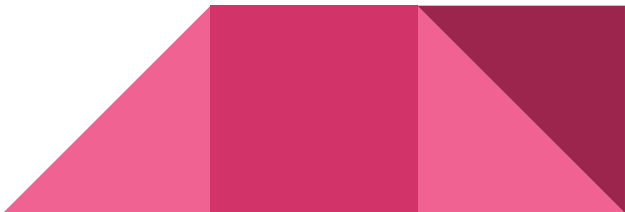
Text Transformation

Feature Selection

Data Mining

**Evaluation & Application
Results**

Text Pre-processing

- ❖ Tokenization
 - ❖ Stop word removal
 - ❖ Stemming and lemmatization
 - ❖ Part of speech (POS) tagging
 - ❖ Chunking
 - ❖ N-grams
- 

Tokenization

- ❖ Tokenization is the process by which big quantity of text is divided into smaller parts called **tokens**.
- ❖ These tokens are very useful for finding patterns and it is considered as a base step for lemmatization.



Stop word removal

- ❖ In computing, **stop words** are words which are filtered out before or after processing of natural language data.
- ❖ Though "stop words" usually refers to the most common words in a language, such as ("the", "is", "an"...).



Stemming and Lemmatization

- ❖ A **stemmer** will return the stem of a word, which needn't be identical to the morphological root of the word.
- ❖ **Lemmatization**, it will return the dictionary form of a word, which must be a valid word.

Cries - Stemmer gives “cri”
Lemmatization gives “cry”




Part of speech tagging

- ❖ **POS tagging** is the process of marking up a word in a corpus to a corresponding part of a speech tag, based on its context and definition. This task is not straightforward, as a particular word may have a different part of speech based on the context in which the word is used.
- ❖ **For example:** In the sentence “Give me your answer”, *answer* is a Noun, but in the sentence “Answer the question”, *answer* is a verb.



Different POS tagging techniques

- ❖ **Lexical Based Methods** — Assigns the POS tag the most frequently occurring with a word in the training corpus.
 - ❖ **Rule-Based Methods** — Assigns POS tags based on rules. For example, we can have a rule that says, words ending with “ed” or “ing” must be assigned to a verb. Rule-Based Techniques can be used along with Lexical Based approaches to allow POS Tagging of words that are not present in the training corpus but are there in the testing data.
 - ❖ **Probabilistic Methods** — This method assigns the POS tags based on the probability of a particular tag sequence occurring. Hidden Markov Models (HMMs) is a probabilistic approach to assign a POS Tag.
 - ❖ **Deep Learning Methods** — Recurrent Neural Networks can also be used for POS tagging.
- 

Chunking

- ❖ Chunking is a process of extracting phrases from unstructured text. Instead of just simple tokens which may not represent the actual meaning of the text, its advisable to use phrases such as “**South Africa**” as a single word instead of ‘**South**’ and ‘**Africa**’ separate words.
- ❖ Chunking works on top of POS tagging, it uses pos-tags as input and provides chunks as output. Similar to POS tags, there are a standard set of Chunk tags like Noun Phrase(NP), Verb Phrase (VP), etc.
- ❖ Chunking is very important when you want to extract information from text such as Locations, Person Names etc. In NLP called Named Entity Extraction.

Chunking

eg. - “the little yellow dog barked at the cat”



N-gram

- ❖ An N-gram means a sequence of N words. So for example, “I like” is a 2-gram (a bigram), “I like sports” is a 3-gram (trigram), and “I like sports because” is a 4-gram.
- ❖ Why we need n-gram?
 - Natural Language Processing, n-grams are used for a variety of things.
 - Examples like auto completion of sentences, auto spell check, and to a certain extent, we can check for grammar in a given sentence.



N-gram

- ❖ Thank you so much for your help.
- ❖ I really appreciate your help.
- ❖ Excuse me, do you know what time it is?
- ❖ I'm really sorry for not inviting you.
- ❖ I really like your watch.

