# JARGONS IN TEXT MINING USING NLP

These are the terms which are often confusing in Natural Language Processing(NLP).The terms I chose were based on terms we often find use on a day to day basis.

By: Mowlanica Billa- https://www.linkedin.com/in/mowlanica-billa-965467140/

## NATURAL LANGUAGE PROCESSING

A Computer Science field connected to Artificial Intelligence and Computational Linguistics which focuses on interactions between computers and human language and a machine's ability to understand, or mimic the understanding of human language.

Examples of NLP applications include Siri and Google Now.

## TEXT MINING

**Text Analytics**, also known as text mining, is the process of examining large collections of written resources to generate new information and to transform the unstructured text into structured data for use in further analysis. It identifies facts, relationships, and assertions that are present in the mass of textual data. These facts are extracted and turned into structured data, for analysis, visualization (e.g. via HTML tables, mind maps, charts), integration with structured data in databases or warehouses, and further refinement using machine learning (ML) systems.

Examples of Text Mining include Sentiment Analysis, Resume filtering, Email filtering.

## NLP and TEXT MINING- the difference

Text Mining deals with the text itself, while NLP deals with the underlying/latent metadata.

Text-mining uses NLP because it makes sense to mine the data when you understand the data semantically.

## SENTIMENT ANALYSIS

It is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

Examples of Applications of Sentiment Analysis Classifying the reviews

## TEXT CLASSIFICATION

It assigns one or more classes to a document according to their content. Classes are selected from a previously established taxonomy (a hierarchy of categories or classes). The Text Classification API takes care of all preprocessing tasks (extracting text, tokenization, stop word removal and lemmatization) required for automatic classification.

This API supports a variety of text classification scenarios like -Binary classification like spam filtering, Multiple class classifications like selecting one category among several alternatives - movie genre classification, Multilabel categorization - assigning all categories that apply to a single document.

## CORPUS/CORPORA

A corpus is a large body of natural language text used for accumulating statistics on natural language text. The plural is Corpora. Corpora often include extra information such as a tag for each word indicating its part-of-speech, and perhaps the parse tree for each sentence.

A corpus contains unstructured natural language text and is used to apply NLP tasks on an attempt to enable machines to better understand this text.

In short, Corpus is a group of documents.

## LEXICON

A word regarded as a comparatively abstract object which has more or less consistent meaning.

or

A lexicon is a collection of information about the words of a language about the lexical categories to which they belong.

Example: dogs and dog are a particular form of lexical item DOG

## BAG OF WORDS,TF-IDF, VECTOR SPACE MODEL

Bag of words and Vector space refer to the different approaches of categorizing body of a document.

In Bag of words, you can extract only the unigram words to create the unordered list of words without syntactic, semantic and POS(Parts Of Speech) tagging. This bunch of words represents the document.

In Vector space model, it is an algebraic model used for representing documents as vectors. from the given bag of words, you can create a feature document vector where each feature is a word and its value is term weight.

In TF-IDF, TF is Term Frequency which represents the number of times a word has appeared in a document, IDF is Inverse Document frequency which tries to how relevant is the word in the document. It is the term weight which is represented in the Vector space model.Thus, the entire document is a feature vector. which points to a point in vector space such that there is an axis for every term in our bag.

# TOKENIZATION

Tokenization is the task of chopping it up into pieces, called Tokens, perhaps at the same time throwing away certain characters, such as punctuation when a character sequence and a defined document unit are given.
An example of tokenization is given below:

Input: Friends, Romans, Countrymen, lend me your ears;
Output:Friends| Romans| Countrymen|lend|me|your|ears

A “token” in natural language terms is “an instance of a sequence of characters in some particular document that is grouped together as a useful semantic unit for processing

# TOPIC MODELLING

In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "**topics**" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body.
It is a process to automatically identify topics present in a text object and to derive hidden patterns exhibited by a text corpus. Thus, assisting in better decision making.
Topics can be defined as “a repeating pattern of co-occurring terms in a corpus”. A good topic model should result in – “health”, “doctor”, “patient”, “hospital” for a topic – **Healthcare**, and
“farm”, “crops”, “wheat” for a topic – “**Farming**"

# NAMED ENTITY RECOGNITION

Named Entity Recognition tries to find out whether or not a word is a named entity. Named entities are persons, locations, organizations, time expressions etc. This problem can be broken down into detection of names followed by classification of a name into the corresponding categories. So most often a word recognized by NER may be recognized as a noun by a POS tagger.
Input: Hyderabad was established in 1591 by Muhammad Quli Qutb Shah
Output: Hyderabad was established in 1591 by Muhammad Quli Qutb Shah.
NER Tagging: Location->Hyderabad,Year->1591, Person->Muhammad Quli Qutb Shah

# REGULAR EXPRESSIONS

A regular expression or a regex for short is a text string used for describing a certain search pattern.
For example:
[0-9]: will return the numbers between 0 and 9 single-valued entities
'nlp': will return the strings containing 'nlp' in them
These are useful especially in NLP because there is a lot of unstructured data, where the regular expressions are needed to be able to use these patterns to create some structure within the document.
Use cases: Searching for URL, Searching for files on your computer, Document scraping

# POS TAGGING

Part-of-speech tagging is one of the most important text analysis tasks used to classify words into their part-of-speech and label them according to the tagset which is a collection of tags used for the POS tagging. Part-of-speech tagging was also known as word classes or lexical categories. It is built on top of the word segmentation(tokenization) and sentence. An important use for POS tagging - Word Sense Disambiguation. Words often occur in different senses as different parts of speech.
She saw a bear.
Your efforts will bear fruit.
The word bear in the above sentences has completely different senses, but more importantly one is a noun and the other is a verb.

# VECTORIZATION

Vectorizing means converting a text to a numerical representation of that text, where you are actually counting the occurrences of each word in each text message using a matrix with one row per text message and one column per word in the form of a Document-term matrix.

A Document term matrix is a matrix in which each cell is a weight of how important that word is, by measuring how frequently it occurs within that text message, relative to how frequently that word occurs across all other text messages.

# LDA,LSA,ESA,NMF

- Latent Dirichlet Allocation (LDA) – A common topic modeling technique, LDA has based on the assumption that each document or piece of text is a mixture of a small number of topics and that each word in a document is attributable to one of the topics.
- Latent Semantic Analysis (LSA) – The process of analyzing relationships between a set of documents and the terms they contain. Accomplished by producing a set of concepts related to the documents and terms. It assumes that words that are close in meaning will occur in similar pieces of text.
- Explicit Semantic Analysis (ESA) – Used in Information Retrieval, Document Classification and Semantic Relatedness calculation (i.e. how similar in meaning two words or pieces of text are to each other), ESA is the process of understanding the meaning of a piece text, as a combination of the concepts found in that text.
- Non-negative Matrix Factorization (NMF) is a state of the art feature extraction algorithm. NMF is useful when there are many attributes and the attributes are ambiguous or have weak predictability. By combining attributes, NMF can produce meaningful patterns, topics, or themes.NMF is often used in text mining. In a text document, the same word can occur in different places with different meanings. For example, "hike" can be applied to the outdoors or to interest rates.:"hike" + "mountain" -> "outdoor sports" and "hike" + "interest" -> "interest rates"

# STEMMING & LEMMATIZATION-the difference

**Stemming** is the process of reducing inflected or derived words to their word stem or root. More simply put, the process of stemming means often crudely chopping off the end of a word, to leave only the base. So this means taking words with various suffixes and condensing them under the same root word. For example: stemming/stemmed=>stem, electrical/electricity=>electr,berries/berry=>berri,connection/connective=>connect,Meanness/meaning=>mean
**Lemmatization** is the process of grouping together the inflected forms of a word so they can be analyzed as a single term, identified by the word's "**lemma**". The lemma is the canonical form of a set of words.
For example, the words typed, typing, typed are reduced to "type".
The **goal** of both is to condense derived words down into their base form, to reduce the corpus of words that the model's exposed to and to explicitly correlate words with similar meaning.
The **difference** is that **Stemming** takes a more crude approach by just chopping off the ending of a word using heuristics, without any understanding of the context in which a word is used whereas Lemmatization leverages more informed analysis to create groups of words with similar meaning based on the context around the word, part of speech, and other factors. Lemmatizers will always return a dictionary word. For example:
meanness/meaning => Stemming output: mean ; Lemmatizer output: meanness,meaning