Slide-2

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The result is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

Slide-5

Language is a thing of beauty. But mastering a new language from scratch is quite a daunting prospect. If you've ever picked up a language that wasn't your mother tongue, you'll relate to this! There are so many layers to peel off and syntaxes to consider – it's quite a challenge.

And that's exactly the way with our machines. In order to get our computer to understand any text, we need to break that word down in a way that our machine can understand. That's where the concept of tokenization in Natural Language Processing (NLP) comes in.

Simply put, we can't work with text data if we don't perform tokenization. Yes, it's really that important!

And here's the intriguing thing about tokenization – it's not just about breaking down the text. Tokenization plays a significant role in dealing with text data.

Slide -6

Stopwords are the words in any language which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. For some search engines, these are some of the most common, short function words, such as the, is, at, which, and on. In this case, stop words can cause problems when searching for phrases that include them, particularly in names such as "The Who" or "Take That".

Pros:

Stop words are often removed from the text before training deep learning and machine learning models since stop words occur in abundance, hence providing little to no unique information that can be used for classification or clustering.

* On removing stopwords, dataset size decreases, and the time to train the model also decreases without a huge impact on the accuracy of the model.

* Stopword removal can potentially help in improving performance, as there are fewer and only significant tokens left. Thus, the classification accuracy could be improved

Slide-7:

stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. ... Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation.

Slide-8:

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. Lemmatization is similar to stemming but it brings context to the words. So it links words with similar meaning to one word.

One major difference with stemming is that lemmatize takes a part of speech parameter, “pos” If not supplied, the default is “noun.”

Slide-10:

Named Entity Recognition (NER) is a standard NLP problem which involves spotting named entities (people, places, organizations etc.) from a chunk of text, and classifying them into a predefined set of categories. Some of the practical applications of NER include:

Scanning news articles for the people, organizations and locations reported.

Providing concise features for search optimization: instead of searching the entire content, one may simply search for the major entities involved.

it is interesting to note that spaCy’s NER model uses capitalization as one of the cues to identify named entities.

In the output, the first column specifies the entity, the next two columns the start and end characters within the sentence/document, and the final column specifies the category.

The word “apple” no longer shows as a named entity. Therefore, it is important to use NER before the usual normalization or stemming preprocessing steps.

Slide-12:

The Bag-of-words model is an orderless document representation — only the counts of words matter. For instance, in the above example "John likes to watch movies. Mary likes movies too", the bag-of-words representation will not reveal that the verb "likes" always follows a person's name in this text.

Thus, Bag of Words (BOW) is a method to extract features from text documents. These features can be used for training machine learning algorithms. It creates a vocabulary of all the unique words occurring in all the documents in the training set.

Slide-13:

The idea is simple. Cosine similarity takes the angle between two non-zero vectors and calculates the cosine of that angle, and this value is known as the similarity between the two vectors. This similarity score ranges from 0 to 1, with 0 being the lowest (the least similar) and 1 being the highest (the most similar).

It is often used to measure document similarity in text analysis.

Slide-14:

RegEx, or Regular Expression, is a sequence of characters that forms a search pattern.

RegEx can be used to check if a string contains the specified search pattern.

Python has a built-in package called re, which can be used to work with Regular Expressions.

Slide-16:

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language. ... NLTK supports classification, tokenization, stemming, tagging, parsing, and semantic reasoning functionalities.

Slide-18:

Text classification is the process of assigning tags or categories to text according to its content. It's one of the fundamental tasks in natural language processing with broad applications such as sentiment analysis, topic labeling, spam detection, and intent detection.

Thus, it is Important for businesses these days.

Linear Support Vector Machine is widely regarded as one of the best text classification algorithms.