25-02-2021 (190) KANAV BANSAL

LECTURE-56

NATURAL LAGUAGE PROCESSING (NLP):

A branch of AI, that deals with the interaction between computers and humans using the natural language.

→ The ultimate objective of NLP is to

* Read
* Decipher
* Understand and
* make sense of human languages in a manner that is valuable.

HOW DOES NLP WORK?

As we know that NLP involves the reading and understanding of spoken or written language through the medium of a computer.

→ This includes,

For an Example:

* The automatic translation of one language into another, but also spoken word recognition or the automatic answering of the questions.

The five common techniques used for extracting information from text are:

→ NAMED ENTITY RECOGNITION

- The most basic and useful technique in NLP is extracting the entities in text.

→ SENTIMENT ANALYSIS.

→ TEXT SUMMARIZATION.

→ ASPECT MINING

→ TOPIC MODELING.

STEMMING:

A process of reducing a word to its word stem that affixes to suffixes & prefixes (or) to the roots of words known as lemma.

→ Stemming is important in Natural Language Understanding (NLU) and NLP.

→ When a new word is found, it can present new research opportunities.

* There are majorly 2 errors in stemming Algorithms. They are:

↳ OVERSTEMMING: This is when two words with different stems are stemmed to be same root.

- This is also known as a false positive

1. Universal
2. University
3. Universe

All the above 3 words are stemmed to "UNIVERS" which is wrong behavior.

→ Though these three words are etymologically related, their modern meanings are in widely different domains, so treating them as synonyms in NLP/ NLU will likely reduce the relevance of the search results.

4 UNDERSTEMMING: This is when - two words that should be stemmed to the same root or not.

- This is also known as a false negative.

1. Alumnus
2. Alumni
3. Alumnae.

→ As of now, there are lot of ways by which we can stem a word and in this, we will be focusing on 3 stemming techniques.

There are two types of stemming. they are

1. PORTER STEMMING: A process for removing the commonese morphological and inflexional endings from words in English.

→ Its main use is, as part of a term normalisation process that is usually done when setting up information retrieval systems (IRS).

2. SNOWBALL STEMMER: A stemming algorithm which is also known as the "PORTER2 STEMMING" algorithm as it is better version of the porter stemmer since some issues of it were fixed in the stemmer.

→ As in the case of the suffix 'ed' if the words are 'cared' & 'bumped' they will be stemmed as 'care' & 'bump'.

→ When compared to porterstemmer, the Snowball stemmer is having a greater computational speed.

LEMMATIZATION:

It usually refers to do things properly with the use of a vocabulary and also morphological analysis of words, normally aiming to remove inflectional endings only & return the base or dictionary form of a word, which is known as the "LEMMA".

STOP WORDS:

In NLP, useless words (data), are referred to as stop words.

→ A stop word is commonly used word such as 'the', 'a', 'an', 'in'.... that a search engine has been programmed to ignore, both when indexing entries for

searching and when retrieving them as the result of a search query.

WORDNET:

It is the lexical database i.e., dictionary for the English language, specifically designed for NLP.

→ Synset is a special kind of a simple interface that is present in NLTK to look up words in WordNet.

→ Synset instances are the groupings of synonymous words that express the same concept.

WHAT IS THE PURPOSE OF STEMMING?

→ It usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time & often includes the removal of derivational affixes.

WORDCLOUD:

The name suggests is a cloud of words.

→ It is a visualization technique for text data wherein each word is picturized with its importance in the context or its frequency.

→ This is a very handy application.

→ When it comes to understanding the crux of today's news (or) the content of any youtube channel.