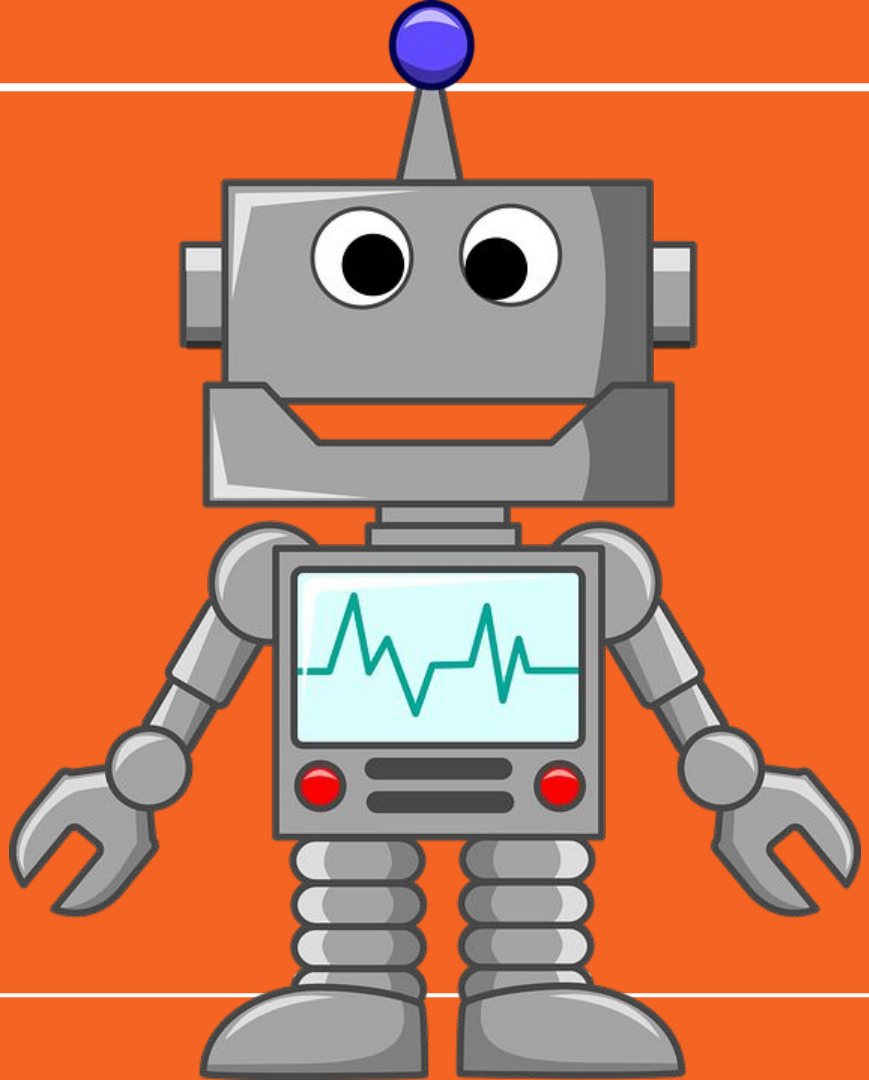
DevCon 4
Natural Language Processing

By: Farisology and Ali

The importance of language in our universe.

Whales sing, wolves howl, birds tweet and chirp, and frogs croak



Takeouts

Thing you should Learn

➔ **What is NLP?**
The essence of modeling human language in the computer.

➔ **Publishing is a challenge**
You can publish in normal journals but not in a reputable ones.

➔ **Simple**
But don't say this to your supervisor

Agenda

- Goals of NLP
- Applications of NLP
- Levels of Language Processing
- Tokenization
- Spelling Correction
- Text Classification (Optional)

Goals of NLP

- To get computers to perform useful tasks involving human language.
- Tasks like enabling human-machine communication.
- Improving human-human communication.
- Doing useful processing of text and speech.

Applications of NLP

Applications of NLP are applications that requires the knowledge of the language in their operation or in delivering the services.

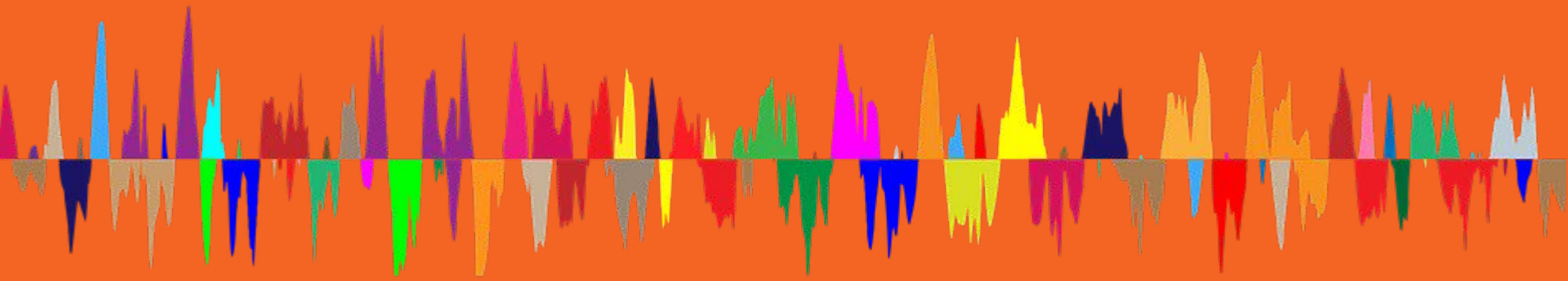
- Translation
- Plagiarism Detection
- Text Classification and emotions mining
- Chatbots

Levels of Language Processing

- Phonetics & Phonology
- Words
- Syntax
- Semantics
- Pragmatics

Phonetics & phonology:

How words are pronounced in terms of sequence of sounds and how each of these sounds realized acountically.



Words:

Lexicon: word set of a language.
Morphology: the study of the structure and forms of a word.

Words in a sentence can be tagged with their part of the speech:
The *(article)* **big** *(adjective)* **cat** *(noun)* **ate** *(verb)* **the** *(article)* **gray** *(adjective)* **mouse** *(noun)*

Syntax:

The order of words in sentence and their relationship.
Prasing: determine the structure of sentence.

Phrase structure rules:

1- A sentence consists of a noun phrase and a verb phrase.
2- A noun phrase consists of an article and a noun.
3- a verb phrase consists of a verb and a noun phrase.

#Syntax as defined by Noam Chomsky 1957

The boy hit the ball

Semantics:
Meaning of words and sentences.

Pragmatics:

The contextual interpretation; meaning of words and sentences in specific situations.

Tokenization:

The process of splitting a string into a list of pieces of tokens. A token is a piece of a whole.

A word is a token of in a sentence.
A sentence if a token in a paragraph.

Let's code Ipython

Code hints

Tokenization

```
from nltk.tokenize import sent_tokenize

para = 'Hello there, my name is faris. I am feeling sleep.
Please wash your face'

sent_tokenize(para)

spanish_tokenizer =
nltk.data.load('tokenizers/punkt/PY3/spanish.pickle' )

spanish_tokenizer.tokenize('Hola amigo. Estoy bien.')
```

```
From nltk.tokenize import
word_tokenize

word_tokenize(‘hello World’)
```

Code hints

Stop words

```
From nltk.corpus import stopwords

english_stops = set(stopwords.words('english'))

>>> words = ["Can't", 'is', 'a', 'contraction']

>>> [word for word in words if word not in english_stops]
```

```
stopwords.words('dutch')
```

Code hints

Stemming

```
from nltk.stem import PorterStemmer

stemmer = PorterStemmer()

stemmer.stem('cooking')

stemmer.stem('cookery')
```

is a technique to remove affxes from a word, ending up with the stem. For example, the stem of cooking is cook, and a good stemming algorithm knows that the ing suffx can be removed. Stemming is most commonly used by search engines for indexing words. Instead of storing all forms of a word, a search engine can store only the stems, greatly reducing the size of index while increasing retrieval accuracy