Introduction

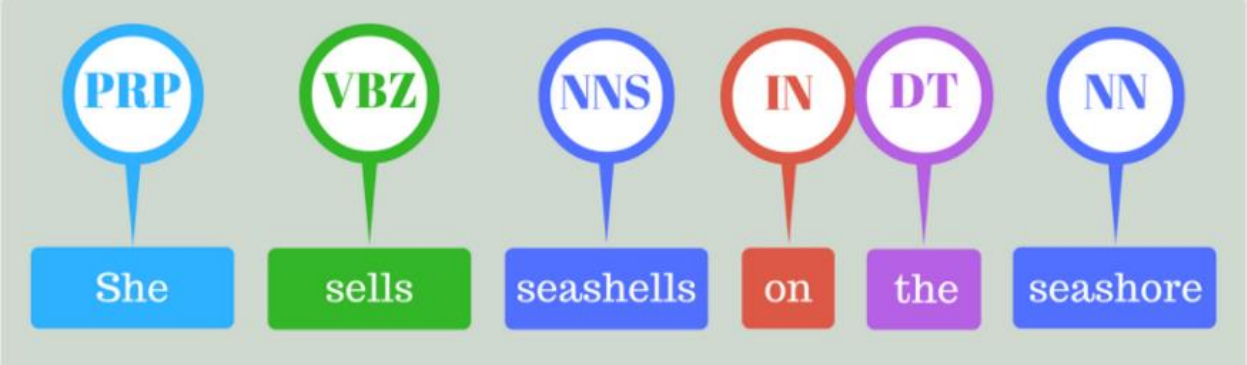
This document will explain you the Part of Speech (POS) tagging and chunking process in NLP using NLTK. , POS and Chunking helps us overcome this weakness:

(Bag-of-words fails to capture the structure of the sentences and sometimes give its appropriate meaning.)

What is Part of Speech?

The part of speech explains how a word is used in a sentence. There are eight main parts of speech – nouns, pronouns, adjectives, verbs, adverbs, prepositions, conjunctions and interjections.

Most POS are divided into sub-classes. POS Tagging simply means labeling words with their appropriate Part-Of-Speech.



POS tagging is a supervised learning solution that uses features like the previous word, next word, is first letter capitalized etc. NLTK has a function to get pos tags and it works after tokenization process.

What is networkx for POS ?

NetworkX is a Python package for creating, manipulating, and analyzing complex networks or graphs. One of the potential applications of NetworkX is in analyzing and visualizing the relationships between parts of speech in natural language processing (NLP)

Overall, NetworkX provides a flexible and powerful framework for working with parts of speech and analyzing their relationships, making it a valuable tool for natural language processing and computational linguistics research.

In this project we will represent POS by Networkx

DATA DESCRIPTION

Overview

This is an entity-level sentiment analysis dataset of twitter. Given a message and an entity, the task is to judge the sentiment of the message about the entity. There are four classes in this dataset: Positive, Negative ,Neutral and Irrelevant . We regard messages that are not relevant to the entity (i.e. Irrelevant) as Neutral

Usage

using twitter_training.csv as the training set
and twitter_testing.csv as the test set.

Content

DATA CONTAIN 5 COLUMNS AND 74682 ROW

- INDEX : INDEX OF A ROW
- TWEET ID : ID OF TWEET
- ENTITY : TYPES OF VIDEO GAMES BRANDS
- SENTIMENT : Positive, Negative ,Neutral and Irrelevant
- TWEET CONTENT : A TWEET ITSELF

TOOLS

- Python
- Nltk
- Matplotlip
- Numpy
- Networkx
- Tensorflow
- Holoviews

Baseline Experiments

THE GOAL → part of speech (POS) tagging that represented by network graph .

Steps of Experiments :

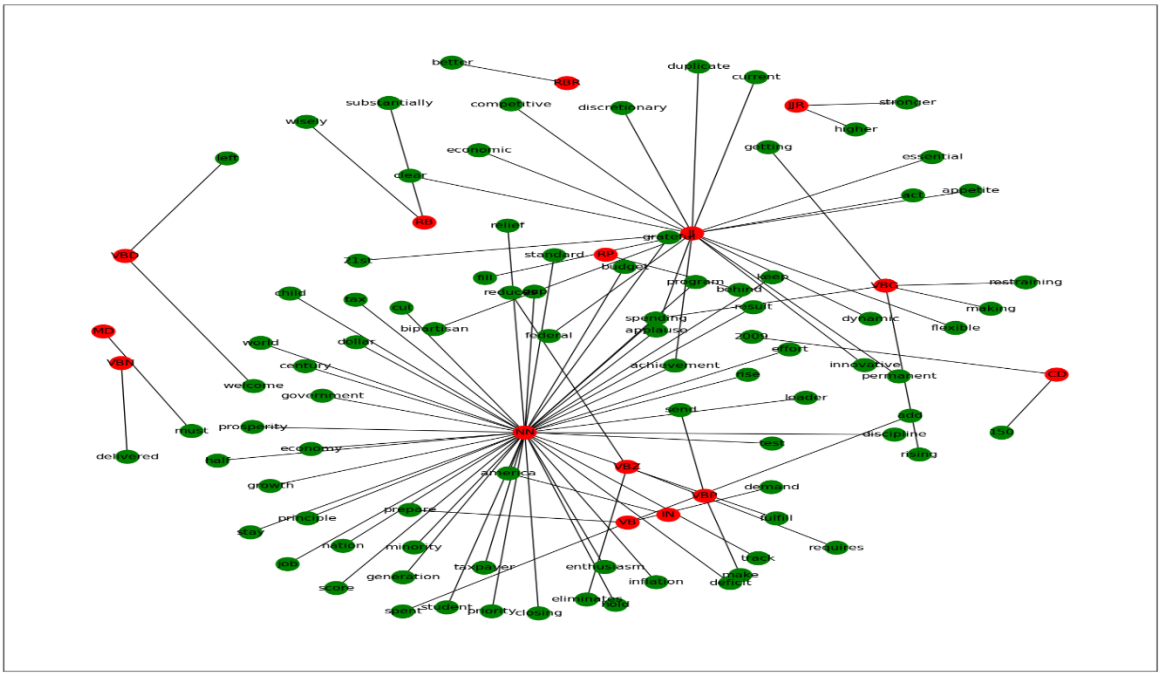
1) Load data (Data was splited to Train & Test File)
2) Text pre-processing

   Not all the information is useful in making predictions or doing classifications. Reducing the number of words will reduce the input dimension to your model. The way the language is written, it contains lot of information which is grammar specific. Thus when converting to numeric format, word specific characteristics like capitalisation, punctuations, suffixes/prefixes etc. are redundant. Cleaning the data in a way that similar words map to single word and removing the grammar relevant information from text can tremendously reduce the vocabulary. Which methods to apply and which ones to skip depends on the problem at hand.

   - Lower casing
   - Remove all the special characters - remove all single characters
   - Substituting multiple spaces with single space
   - Removing prefixed
   - remove stop words
   - word tokenize - word lemmatization
   - POS tag for each word
3) building graph using networkx
   - Add nodes for each word and POS tag
   - Set node colors based on POS tag
   - Draw the network graph
   - Show the network graph

RESULTS

pairings between words , tags and the network graph :



CONCLUSION

In conclusion, using NetworkX graphs for part of speech has several benefits in natural language processing and computational linguistics research. NetworkX provides a flexible and powerful framework for representing, analyzing, and visualizing the relationships between parts of speech, allowing researchers to gain insights into the syntactic and semantic structure of language.

References

- https://networkx.org/documentation/stable/reference/drawing.html
- https://www.geeksforgeeks.org/nlp-part-of-speech-default-tagging/