MACHINE
LEARNING

Natural Language Processing (NLP)
Machine Learning

Diptangshu Banik

Twitter @dipbanik

Overview

- Natural Language Processing (or NLP) is applying Machine Learning models to text and language.
- Sentiment Analysis – use text review to predict if the review is a good one or a bad one.
- predict some categories of the articles.
- predict the genre of the book

- Use NLP to build a machine translator or a speech recognition system
- Question Answering – Chatbots
- Document Sumarization
- Natural Language Understanding – LUIS, Google Dailogflow, Amazon Lex.

Python Libraries

- Natural Language Toolkit – NLTK
- SpaCy
- Stanford NLP
- OpenNLP

Algo

- Most of NLP algorithms are classification models
  - Logistic Regression
  - Naive Bayes
  - CART which is a model based on decision trees
  - Maximum Entropy again related to Decision Trees
  - Hidden Markov Models which are models based on Markov processes
  - C 5.0 – Decision Tree based model

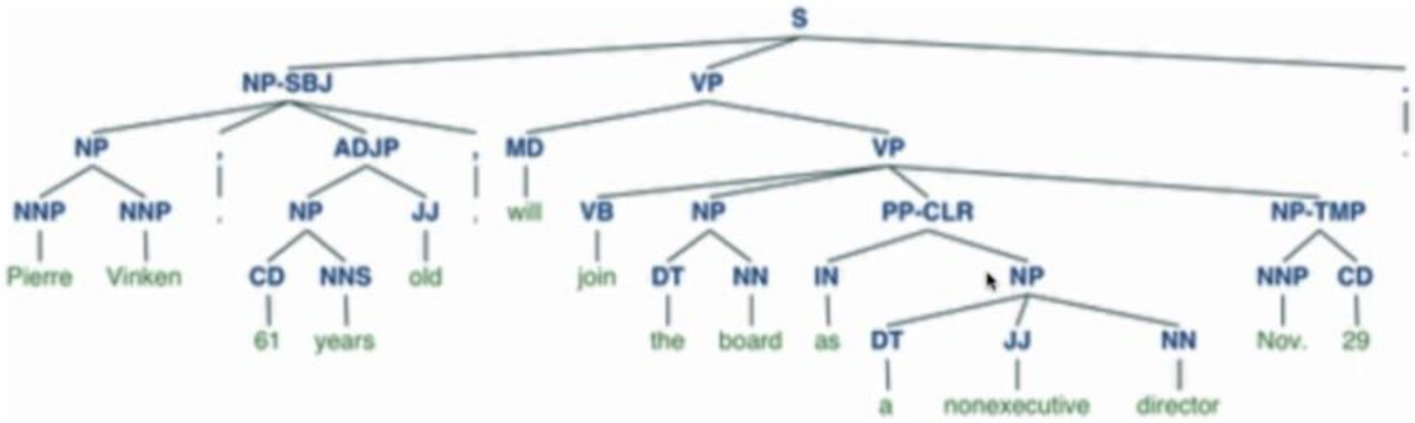
Algo

- A very well-known model in NLP is the Bag of Words model. It is a model used to preprocess the texts to classify before fitting the classification algorithms on the observations containing the texts.

Steps - Bag of Words

- Clean texts to prepare them for the Machine Learning models
- Create a Bag of Words model
- Apply Machine Learning models onto this Bag of Worlds model.

Parts of Speech Tagging



Parts of Speech Tagging for a sentence

Parts of Speech Tagging

- Wikipedia Desc –

Marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context — i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

Stop Words

- These are words which do not add any value to the text and are only used for stitching words together.

- Examples – for, the, a, an, of, it, . , ?, etc.

Stemming

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma.

Love = love

Loved = love

Loving = love

TF-IDF

Term frequency–inverse document frequency

- A numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.
- Used as a weighting factor in searches of information retrieval, text mining, and user modeling.
- The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general.
- Tf–idf is one of the most popular term-weighting schemes today; 83% of text-based recommender systems in digital libraries use tf–idf.