Python-Machine Learning

Natural language processing (NLP)

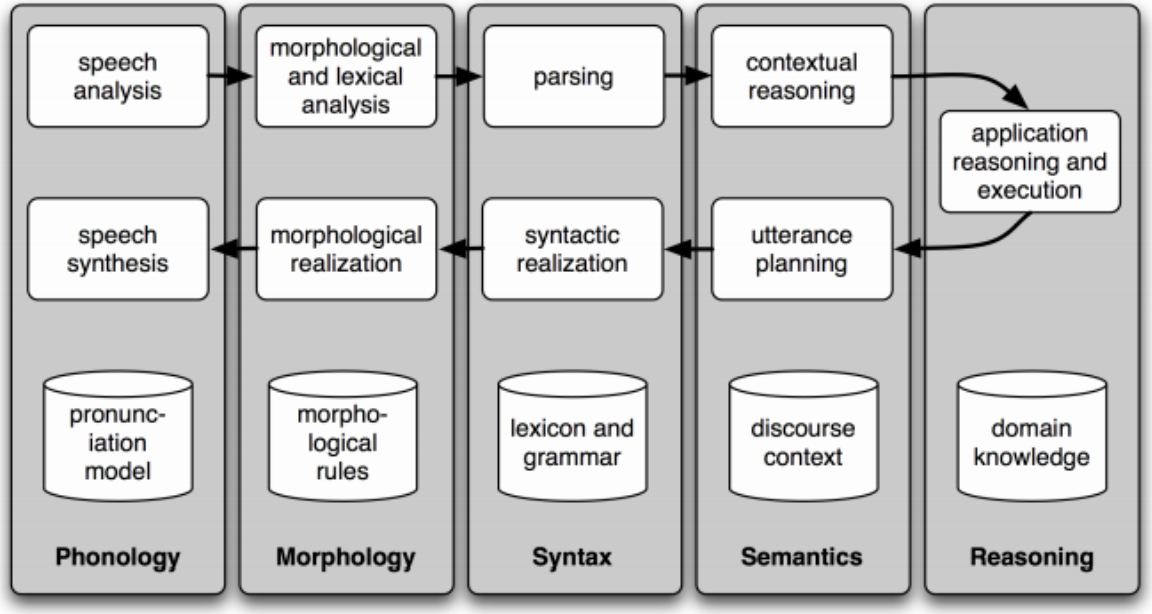
using NLTK package

Dr. Sarwan Singh

Natural Language
Analyses with NLTK

long-standing challenge within computer science has been to build intelligent machines

The chief measure of machine intelligence has been a linguistic one, namely the Turing Test: can a dialogue system, responding to a user's typed input with its own textual output, perform so naturally that users cannot distinguish it from a human interlocutor using the same interface? Today, there is substantial ongoing research and development in such areas as machine translation and spoken dialogue, and significant commercial systems are in widespread use

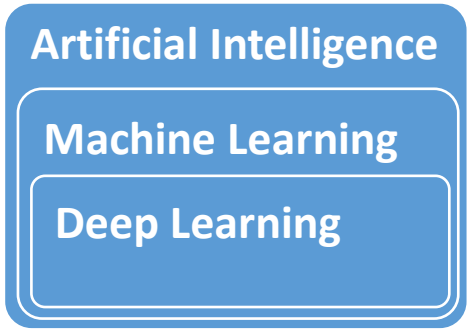


Simple Pipeline Architecture for a Spoken Dialogue System

Agenda

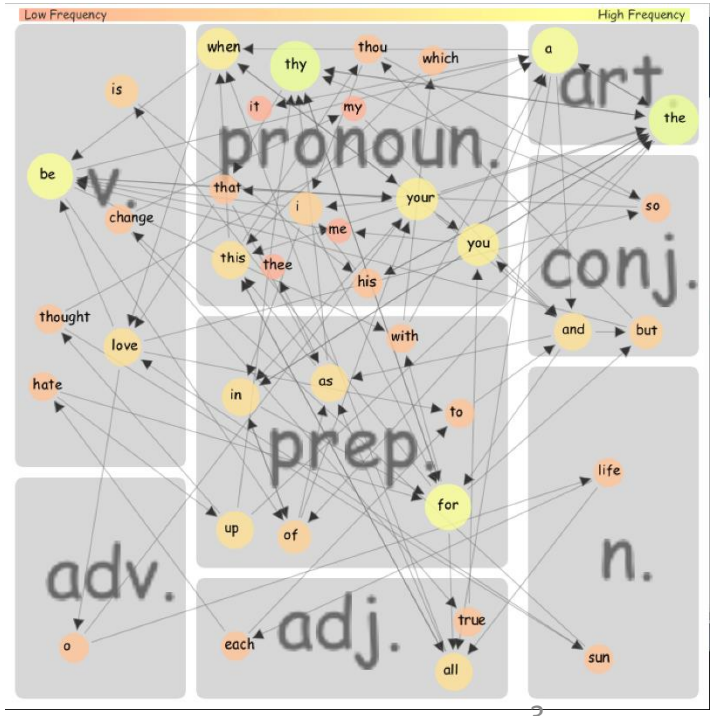
- Introduction - Sentiment Analysis

Natural Language Analyses with NLTK



Covered in part 1

- Introduction (NLTK Toolkit)
- History, benefits ,
- Libraries, Uses
- Tokenize Text Using NLTK
- Wordnet, Lemmatizing Words



Machine learning is a branch in computer science that studies the design of algorithms that can learn.

TEXT CLASSIFICATION FOR SENTIMENT ANALYSIS – **NAIVE BAYES CLASSIFIER**

Introduction - **Sentiment Analysis**

- It is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.
- It is becoming a popular area of research and social media analysis, especially around user reviews and tweets. It is a special case of text mining generally focused on identifying opinion polarity, and while it’s often not very accurate, it can still be useful.

Source: kaggle.com, google NLTK Cookbook

Principle of Naive Bayes Classifier

- A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

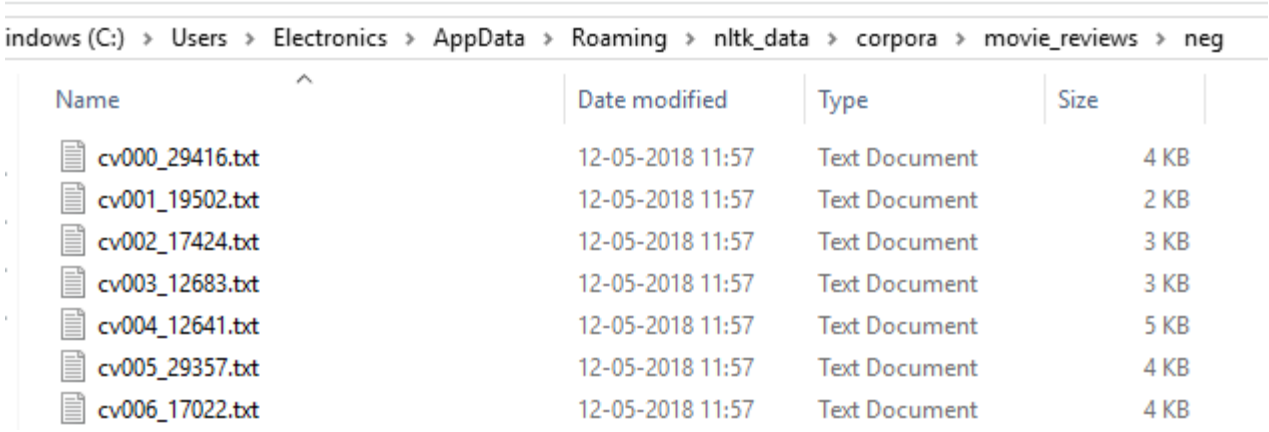
- Using Bayes theorem, we can find the probability of **A** happening, given that **B** has occurred. Here, **B** is the evidence and **A** is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

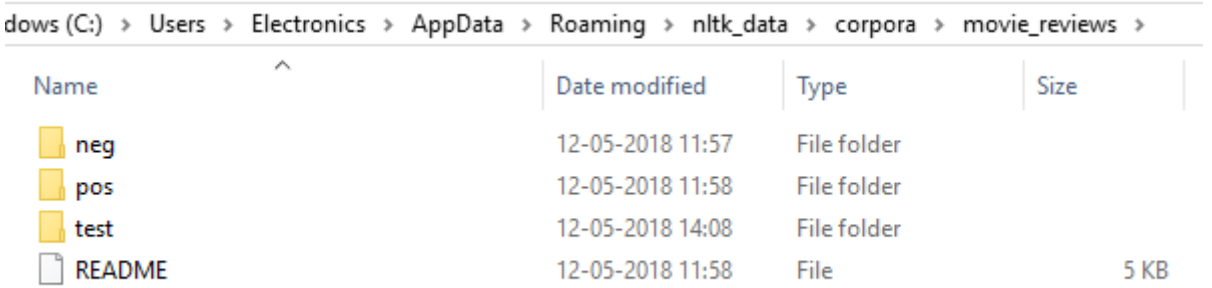
Conlusion

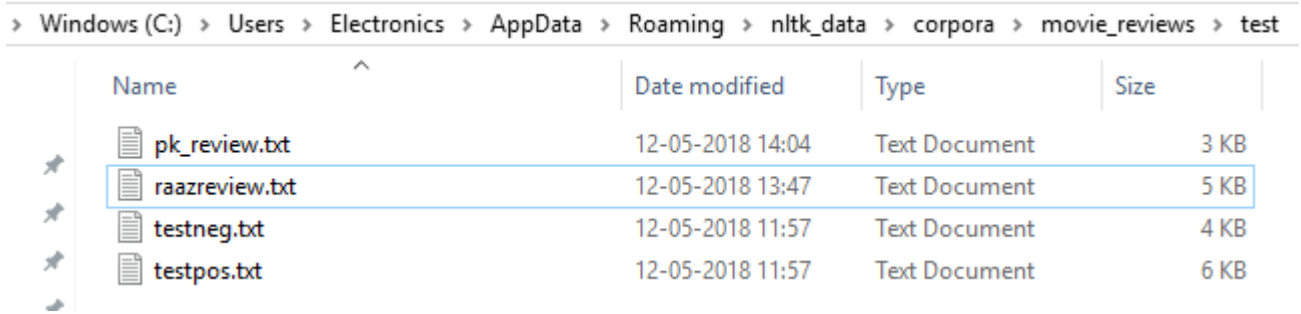
- Naive Bayes algorithms are mostly used in sentiment analysis, spam filtering, recommendation systems etc.
- They are fast and easy to implement but their biggest disadvantage is that the requirement of predictors to be independent.
- In most of the real life cases, the predictors are dependent, this hinders the performance of the classifier.

NaiveBayesClassifier

- NLTK comes with all the pieces you need to get started on sentiment analysis:
  - a movie reviews corpus with reviews categorized into pos and neg categories, and
  - a number of trainable classifiers.
- So, start with a simple NaiveBayesClassifier as a baseline, using boolean word feature extraction.


indows (C:) › Users › Electronics › AppData › Roaming › nltk_data › corpora › movie_reviews › neg
Name
Date modified
Type
Size
cv000_29416.txt 12-05-2018 11:57 Text Document 4 KB
cv001_19502.txt 12-05-2018 11:57 Text Document 2 KB
cv002_17424.txt 12-05-2018 11:57 Text Document 3 KB
cv003_12683.txt 12-05-2018 11:57 Text Document 3 KB
cv004_12641.txt 12-05-2018 11:57 Text Document 5 KB
cv005_29357.txt 12-05-2018 11:57 Text Document 4 KB
cv006_17022.txt 12-05-2018 11:57 Text Document 4 KB


dows (C:) › Users › Electronics › AppData › Roaming › nltk_data › corpora › movie_reviews ›
Name
Date modified
Type
Size
neg 12-05-2018 11:57 File folder
pos 12-05-2018 11:58 File folder
test 12-05-2018 14:08 File folder
README 12-05-2018 11:58 File 5 KB


Windows (C:) › Users › Electronics › AppData › Roaming › nltk_data › corpora › movie_reviews › test
Name
Date modified
Type
Size
pk_review.txt 12-05-2018 14:04 Text Document 3 KB
raazreview.txt 12-05-2018 13:47 Text Document 5 KB
testneg.txt 12-05-2018 11:57 Text Document 4 KB
testpos.txt 12-05-2018 11:57 Text Document 6 KB