Byte The Correct Apple

hackerRank challenge as solved by Krzysztof Strug

Problem description

Provided a text file, with a number of lines. Each line contains either a sentence or a paragraph or a text snippet which could either be related to Apple, the computer company, or the apple, the fruit. Your task is to perform disambiguation between these two groups and identify which one is being referred to. It is possible that the plural or the possessive form of Apple might exist in some of the tests (apples, Apple's).

Problem solution

As the suggested time limit for the task was too short to develop and implement original ideas, the solution was based on Natural Language Toolkit (nltk) - a library for Python that have been written to facilitate and teach processing of natural languages by computers.

One of the most recommended by authors of the toolkit algorithms for text classification is Naive Bayes that is based on Bayesian probabilistic theorem with (naive) assumption of independence of parameters.

All the relevant code is in apples.py.

The hardest part of using NLTK is figuring out the desired data format for classifier. In the case of NaiveBayes the data format for training seems to be: dictionary in form {'contains(WORD)': True}.

Results

Predictive accuracy of constructed model is not very high. HackerRank gives it 12 points out of 100 possible. Removing stopwords, analysing most frequently used words, adjusting parameters of the model could be done to improve the result.

Future work

There are also many other more sophisticated machine learning algorithms that could have proven more succesful. Some examples include: decision trees, random forests, support vector machines.