Categorizing News Articles with Machine Learning

jonathan.c.jauhari@gmail.com

June 2020

Project repository link

Aims

- To develop a news headline text classifier.
- To compare 3 machine learning approaches (Naive Bayes, SGD, neural network) in medium-sized text classification.
- To roughly estimate what kind of articles the New York Times has been publishing in 2020 (expecting a spike in health-related articles due to the pandemic).

Models used

Three models were compared before deciding on which one to use in the New York Times analysis (the detailed results of the comparison are listed in the README):

- Multinomial Naive Bayes classifier
- Stochastic Gradient Descent (SGD) classifier
- Neural network classifier with 2 hidden layers (16 nodes each)

The (Multinomial) Naive Bayes is well-known to be a great baseline model for text identification (its assumption of feature independence is more or less true for tokens in a document). The SGD classifier is based on optimising gradients, and is also a model that is recommended by the scikit-learn library for text classification.

Ultimately, however, the neural network approach proved to be the most accurate, at the cost of higher training and preprocessing time. The neural network model was therefore chosen for the New York Times headlines analysis.

Datasets used

- News Aggregator Dataset: Headlines and categories of 400k news stories from 2014, derived from the UCI Machine Learning Repository dataset.
- The New York Times 2020 headlines of monthly free to read articles that I gathered from: their site map.

The News Aggregator Dataset was chosen, among the many **labeled** news article datasets that are available, because it had by far the most entries when compared to other datasets. Though it only had a few categories, each category was broad enough to be sufficient for the analysis (a rough estimate).

It also did not require as much preprocessing as some other datasets, such as the Reuters news dataset, as it was already in a CSV format, with few missing values, and well-documented features.