News Category Prediction

Jon Dickerson

Stevens Institute of Technology

2016-12-06

Introduction

The dataset I chose to analyze was Kaggle's News Aggregator Dataset.[1] It contains "headlines and categories of 400k news stories from 2014". The goal of this project is to attempt to predict the category of a news story given its headline.

[1] https://www.kaggle.com/uciml/news-aggregator-dataset

Data Structure

The columns included in this dataset are:

- ID : the numeric ID of the article
- TITLE : the headline of the article
- URL : the URL of the article
- PUBLISHER : the publisher of the article
- CATEGORY : the category of the news item; one of:
  - b: business
  - t: science and technology
  - e: entertainment
  - m: health
- STORY : alphanumeric ID of the news story that the article discusses
- HOSTNAME : hostname where the article was posted
- TIMESTAMP : approximate timestamp of the article's publication

Code Plan

The code consists of a few small scripts for ease of use:

- *00-feature_extraction.py*: stem and vectorize the headlines
- *01-build_models.py*: train a few models to compare performance
- *02-evaluate_models.py*: evaluate the above models
- *03-tune_winner.py*: take the model with the best baseline performance and tune it using grid search
- *04-evaluate_winner.py*: rerun the evaluation on the tuned model

Results

The winning model was a linear support vector machine with gradient descent. The final confusion matrix is given below.

$$\begin{bmatrix} 26,832 & 486 & 231 & 1535 \\ 453 & 37,170 & 120 & 296 \\ 475 & 322 & 10,543 & 156 \\ 1430 & 489 & 126 & 24,941 \end{bmatrix}$$