Stacey Camp

Capstone Project #2

**Title:** Using Clustering, Natural Language Processing and Modeling Techniques to Classify Op-Ed Articles at the New York Times by their Writers

**Links:**

* Jupyter Notebook: https://github.com/spaceport729/thinkful/blob/master/NYtimes\_Editorials\_Capstone.ipynb
* Data:

https://github.com/spaceport729/thinkful/blob/master/nytimes\_oped\_articles.csv

**Project Summary**: The goal of this project was to use several supervised and unsupervised learning techniques to classify articles by writer. After collecting opinion articles from the New York Times, text was cleaned and then features generated using the natural language processing library spaCy. Data was split into training and test sets and then two approaches to clustering and three types of supervised models were applied. Logistic regression using features generated primarily through tf-idf (62% accuracy for the test group) was the most effective combination of the techniques evaluated. This model could likely be improved with more data and additional engineered features.

**Data Collection:** URLs to the most recent opinion pieces for each of the 14 staff Op-Ed writers were collected using the New York Times API. Thirty urls to articles for each writer were initially collected using the Requests Python library. Then, full text articles were scraped by making a second request to the NY Times API using Beautiful Soup to identify article content using HTML tags for each url. Articles with a type other than Op-Ed were removed from the dataset leaving 307 total articles used for feature generation and classification. The resulting text output was cleaned by removing non-ascii characters, brackets and parentheses.

**Feature Engineering**: Data was organized into several different DataFrames to facilitate feature engineering. Full text articles were tokenized using spaCy so that tf-idf methods could be used with the entire article defined as the document (as opposed to a paragraph or sentence). A second database with each article broken down by sentence is was generated for collecting information on word frequency by sentence.

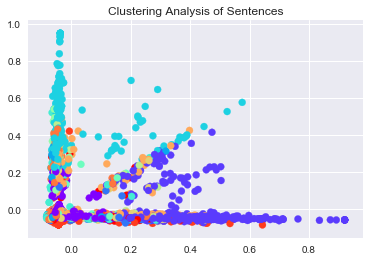
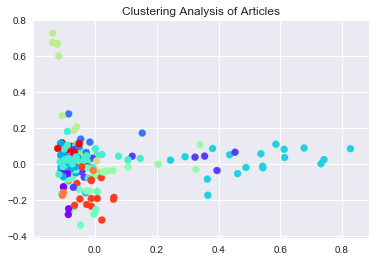
Two feature dataframes were created. The first, with one record per article had word features generated using tf-idf. The second had one sentence per record with word based features generated with the bag of words technique.

**Figure 1:** Data was collected by querying the New York Times API for each of the 14 Op-Ed writers. Data included in the output included subject keywords, full text, date and word count.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| byline | byline | date | full\_text | subjects | word\_count |
| 0 | ANDREW ROSENTHAL | 10/19/2017 | When most Americans think of domestic terroris... | ['Blacks', 'Police Brutality, Misconduct and S... | 753 |
| 1 | ANDREW ROSENTHAL | 10/4/2017 | It’s time to talk about taking away guns — not... | ['Gun Control', 'Firearms', 'Las Vegas, Nev, S... | 776 |
| 2 | ANDREW ROSENTHAL | 5/18/2017 | In normal times, the appointment of a former F... | ['Special Prosecutors (Independent Counsel)'] | 686 |
| 3 | DAVID LEONHARDT | 11/29/2017 | This article is part of the Opinion Today news... | ['United States Politics and Government', 'Tax... | 528 |
| 4 | DAVID LEONHARDT | 11/2/2017 | This article is part of the Opinion Today news... | ['Taxation', 'Baseball', 'Officiating (Sports)... | 725 |

**Clustering:** K-means, Mean-Shift and Spectral clustering approaches were tried on both the tf-idf and bow datasets clustering. For K-means and Spectral clustering, cluster size was set to the number of writers (14). Other cluster sizes were used to check for interesting patterns in the data, such as if certain writers tend to be grouped together. Example output is included below in Figure 2. None of the clustering techniques or cluster sizes created distinguishable groups either for sentences or articles. Mean-shift under multiple parameter conditions only returned a single cluster. Spectral clustering and K-means did not produce clusters distinct clusters at the cluster sizes tried and definitely did not cluster in a pattern that related to the author.

**Figure 2:** Clustering Analysis of full op-ed articles and sentences using K-means. Full text articles were clustered using features generated using tf-idf while sentences were clustered using features from the bag of words technique. Results shown in both charts are for 14 clusters to evaluate the ability of this technique to identify articles written by different authors. The table is the output for clustering by articles.



|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Cluster # | | | | | | | | | | | | | |
| **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** |
| ANDREW ROSENTHAL | 0 | 2 | 0 | 7 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| BRET STEPHENS | 0 | 1 | 1 | 1 | 0 | 10 | 1 | 0 | 0 | 1 | 2 | 4 | 0 | 1 |
| CHARLES M. BLOW | 3 | 1 | 0 | 8 | 0 | 0 | 7 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| DAVID BROOKS | 0 | 0 | 0 | 0 | 1 | 7 | 3 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| DAVID LEONHARDT | 0 | 1 | 0 | 3 | 8 | 5 | 1 | 4 | 0 | 0 | 0 | 1 | 5 | 0 |
| FRANK BRUNI | 2 | 1 | 0 | 2 | 0 | 5 | 2 | 1 | 0 | 0 | 0 | 3 | 3 | 0 |
| GAIL COLLINS | 0 | 5 | 0 | 0 | 0 | 3 | 0 | 1 | 1 | 1 | 0 | 0 | 3 | 0 |
| MAUREEN DOWD | 0 | 1 | 0 | 4 | 0 | 0 | 2 | 3 | 1 | 0 | 0 | 0 | 0 | 0 |
| MICHELLE GOLDBERG | 0 | 0 | 0 | 2 | 1 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 2 | 0 |
| NICHOLAS KRISTOF | 1 | 0 | 0 | 0 | 1 | 6 | 1 | 1 | 3 | 1 | 0 | 0 | 0 | 0 |
| PAUL KRUGMAN | 0 | 0 | 0 | 2 | 6 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| ROGER COHEN | 1 | 1 | 0 | 3 | 0 | 10 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| ROSS DOUTHAT | 4 | 2 | 0 | 0 | 1 | 8 | 1 | 2 | 1 | 0 | 0 | 1 | 1 | 0 |
| THOMAS L. FRIEDMAN | 0 | 1 | 5 | 0 | 0 | 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 6 |

**Modeling**: Naïve Bayes, Logistic Regression and Random Forest models were attempted for the tf-idf feature set and the BoW dataset feature set. Each model was run for both tf-idf and BoW. Best results for each dataset and model combination are included in Table 1. After identifying logistic regression using features generated with tf-idf as the strongest performer, accuracy was further improved by incorporating word count and unique word count as additional features.

**Table 1:** Accuracy results for different model’s abilities to predict writers using tf-idf and bag of words features.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Dataset | Training Set Accuracy | Test Set Accuracy |
| Naïve Bayes | Tf-idf | 0.97 | 0.38 |
|  | BoW | 0.46 | 0.32 |
| Random Forest | Tf-idf | 1.0 | 0.32 |
|  | BoW | 0.97 | 0.21 |
| Logistic Regression | Tf-idf | 0.94 | 0.57 |
|  | BoW | 0.60 | 0.44 |
|  | **Tf-Idf+Word\_counts** | **0.90** | **0.62** |

**Table 2:**  Crosstab evaluation of the values predicted by logistic regression compared to the target values for the test set. Although the test set was a stratified sample, it still appears that the model was better at correctly predicting writers with more articles in the dataset.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | AR | BS | CB | DB | DL | FB | GC | MD | MG | NK | PK | RC | RD | TF |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| AR | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| BS | 0 | 2 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 |
| CB | 1 | 0 | 6 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| DB | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DL | 0 | 1 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| FB | 1 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| GC | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MD | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| MG | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| NK | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| PK | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| RC | 1 | 2 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 5 | 0 | 0 |
| RD | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 5 | 0 |
| TF | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 5 |

**Evaluation of Clustering versus Modeling:** The clustering techniques used in this analysis did not effectively group articles or sentences by author. It’s possible that the texts were too similar to distinguish with this method. However, it is also possible that further analysis into the cluster output, trying additional clustering techniques or generating different types of features would have offered alternative insights into the articles. For instance, it’s possible that clusters could identify patterns in the articles in terms of subject matter or writing style.

All three supervised models had a problem with overfitting. However, using a logistic regression model with the tf-idf generated features (and a few additional custom features) was fairly effective at predicting the writer of articles. The model could likely be further improved with additional feature generation and many more articles used to train the model.