Progress Report

Wilson Fearn, Andrew Hale, Victor Lazaro, Spencer Seeger

March 13, 2018

- A description of the problem
  - The problem that we are solving is allowing machine learning algorithms to take news articles and identify the topics of them. This is applicable to sites that suggest articles for users to read. For example, Google will suggest articles for users based on articles they have read in the past. By using classification, Google would be able to suggest articles that are of the same topic as what a user has read before.
- What machine learning model we are initially trying to learn with
  - The machine learning algorithms we are planning on using include Naïve Bayes, Subject Vector Machines, and Grid Search. All of these learning algorithms can be applied to text classification. We plan on training on all three of them to see which one gives the best overall accuracy, and then optimizing the best machine learning algorithm.
- How and from where we are gathering data
  - We will be gathering data from many different news websites, including www.washingtonpost.com and www.cnn.com. To get this data, we will have to manually identify website articles and their classification. After this is done, a simple Python script will fetch the article and store it as "article text, classification" in a spreadsheet file. The next section includes an instance of our data.
- A description of our data set including:
  - Actual example instances, including a reasonable representation (continuous, nominal, etc.) and values for each feature

Article	Classification
Like the Huawei Mate 10 Pro that we reviewed earlier this week...	Technology
President Trump on Thursday doubled down on his idea...	Political

  - How many instances and features we plan to have in our final data set
    - We hope to have 400 instances to train and test on, and the features we will have will differ from what our original dataset contains. This is because we will need to transform the data into a different form before handing it to our chosen machine learning algorithm. The text from the articles will be passed to Sklearn which will use Bucket of Words to obtain word counts for each article. This data will then be used for training the model. Thus, as we gather more articles, the final set of features will change due to the inclusion of new words from different articles. Below is an example of what the the machine learning algorithm will actually use to train.

Upcoming	Displays	We	...	Like	Classification
0.0543789	0.0234958	0.05830	...	0.03093	Technology
0.0907910	0.0000	0.1257	...	0.07839	Political

- Brief discussion of plans and schedule to finish the project
  - March 13th, 2018 - Submitted Progress report
  - March 16th, 2018 - Finalized list of website articles and their classifications
  - March 19th, 2018 - Finished data processing (grabbing data from web pages)
  - March 23rd, 2018 - Chose machine learning algorithm to use from original list
  - March 30th, 2018 - Finished fine tuning of chosen machine learning algorithm
  - April 6th, 2018 - Completed final draft of project write up and presentation
  - April 9th, 2018 - Finished final Editing of the project write up and presentation
  - April 10th, 2018 - Submitted final project