Dan Burke

Dburke01@syr.edu

Homework 5 – Option 2

MB Naïve Bayes

05/22/2023

|  |  |
| --- | --- |
| Homework 5 – Option 2 | |
| **Introduction** | For some the latest “get rich quick” strategy is investing in cryptocurrencies. Many fortunes have been made and lost within this space, a space which often is ruled by what currency possesses the most headlines. This report attempts to address one application of sentiment analysis in the form of predicting an articles topic/classification – a specific cryptocurrency given its text contents; to better parse through what cryptocurrencies news articles gravitate towards. This classification of articles can then be leveraged to further identify follow on approaches for sentiment analysis.  This report intends to apply the Naïve Bayes approach to article data textual data sourced from NewsAPI.org. After data preparation this data possesses two columns, “content” – denoting article content in the form of a summary and “label” the label derived from the API query string (Bitcoin, Ethereum and Dogecoin).  Naïve Bayes is a probabilistic machine learning process utilizing the Bayes’ theorem. This process assumes that all features (columns) of the data set are conditionally independent. Within the scope of this report the features refer to induvial words; all individual words contained within the dataset (content column/feature). |
| **Analysis** | data preparation The crypto currency article has been sourced from NewsAPI.org. The data is fetched via an API (Application Programming Interface) call (http request), whose response is formatted list of news articles in the form of JSON (JavaScript Object Notation).  The API HTTP request, JSON format response required this effort to both create custom methods to retrieve and format the data. This was combined into one topic agnostic method, which accepts the starting date, sorting method, topic to query, API key and Boolean “save (or export if in virtualized environment) to file” parameters. This custom method retrieves, formats and leverages regular expressions to remove unwanted/irrelevant text from the raw data.  This custom data retrieval and data formatting method was the best approach as this effort required multiple topics to query from NewsAPI.org. Encapsulating this functionality into a custom method allows for reusability and reproducible results (in the form of formatting) when retrieving data from this source. This method also labels the data on a per topic basis (using the term to query parameter), as article data fetched from NewsAPI.org does not possess an explicitly defined label (nor should it as it is only a data source not a tool for conducting analysis). This method also accounts for label data type; properly assigning this column/feature as categorical as the Multinominal Naïve Bayes approach can only accept numerical datatypes.    Figure Data Retrieval and Preparation Custom Function    Figure Method Calls, Exporting to CSV File    Figure Data Preparation Method Resulting Data Frame  **Label Extraction and Fitting the Model**  MB Naïve Bayes models only allows for one “label” or “classifier”. This required the effort to combine 3 data frames containing NewsAPI.org article data on Bitcoin, Ethereum and Dogecoin. Each data frame resulted from an individual API call (custom method call) and was hence formatted identically. This allowed for a simple concatenation of data frames via the pandas function “pandas.concat()” (pandas below is imported as “pd”).    Figure Data Frame Concatenation  With these data frames (Bitcoin, Ethereum and Dogecoin) it was now necessary to count vectorize the aggregate textual data, in order to format it properly for generation of a MB Naïve Bayes model. A custom method was created to count vectorize and return a clean data frame for use within a Naïve Bayes model. This was done in order to both possess reusable code as well as creating an economical workflow.    Figure Count Vectorize Method    Figure Count Vectorize to Data Frame Method  To aid in ease of workflow and create reusable code which leverages the NewAPI.org API, this effort has created a “all in one” function which leverages the previous count vectorizer code as well as performing the test, train, split (80/20), fitting to a model, running prediction and generating a confusion matrix (as displayed below).    Figure All in One Function  . |
| **results** | technical results **Article Crypto Currency Topic Prediction**    **Confusion Matrix - Article Crypto Label Prediction**    **Word Cloud**    **Word Cloud Code**    Figure Word Cloud Code |
| **conclusions** | The Multinomial Naïve Bayes approached produced results of unacceptable accuracy ~58% with article cryptocurrency label prediction.  First and foremost, expansion of the dataset may increase prediction accuracy, however a custom stop words list must be investigated. Much of the crypto currency space uses new terms and abbreviations which are alien to many outside of this space. A custom stop words list and or weighting of terminology specific to individual cryptocurrencies may aid in article classification. In conclusion these results are not ready for implementation into a production environment.  . |