DOCUMENT CLASSIFICATION

36350 Final Project by
Johnny Bae
Benjamin McGrath
Wan Xin Teo

Project Description

Document classication You will get news stories from the New York Times with known subject classication. You will write code to extract the text, turn the word frequencies into features, and automatically sort the text into categories, using regression-like methods. We will provide texts for two groups, one learning to discriminate articles about art from articles about music, the other learning to discriminate news stories from editorials

57 art texts

45 music texts

Word frequencies

Sort text into categories

Methodology

1. Extract words from dataset
2. Find word frequencies unique to each category
3. Build regression model
4. Testing
5. Final Model

Extracting Text from HTML Code

- Using Regex
- Remove links
  - Not useful in determining categories
  - Difficult to differentiate from texts
- Kept the headlines & subheadings
  - Could have looked for "<block class=\"full_text\">” and extract that up till end of body class "</body.content>"
- Remove punctuation and digits
- Change all words to lower case

Finding Word Frequencies

- Top 20 words (using sort)
  - In each article in the same category
  - In all articles in each category
- Ignored words such as
  - the, a, for
- Looked at the frequencies of words such as
  - art* -> arts, artist, artistic **removed article
  - compos* -> compose, composer, composes, composition, compositions

Regression Model

- Chose to use Probit Model
- Included variables such as
  - Theatre
  - Broadway
  - Museum
  - Art*
  - Music*

Testing

- Leave one out cross validation
  - Remove one row (article) of the data set
  - Fit a model based on all but that one row
  - Based on the word frequencies of the row that is not in the sample, predict whether that article is about art or music
  - Repeat this process once for each row
  - Count what percentage of articles were correctly predicted when it was not included in the data used to create the model

Final Model

- Repeat the leave-one-out cross-validation methods several times with different models (removing and adding word frequencies) to maximize the accuracy of the model
- Final words that were included in the probit regression:
  - Art, Music, Exhibit, Sculpture, Paint, Street, Museum, Color, Photo, Gallery, Opera, Band, Show, Play, Compose, Perform.
- Final model
  - glm(formula = Art ~ Art.freq + Music.freq + Exhibit.freq + Sculpture.freq + Paint.freq + Street.freq + Museum.freq + Color.freq + Photo.freq + Gallery.freq + Opera.freq + Band.freq + Show.freq + Play.freq + Compose.freq + Perform.freq, family = binomial(link = "probit"))

Difficulties

- Extracting text from html code
- Finding word frequencies
- Using REGEX
  - art.{0,10} -> includes articles
- Building a regression model
  - More variables for higher Rsquare or less variables for stability
- Cannot test on out-of-sample texts because extraction code would be different