Labeling data reuse statements through Active Learning

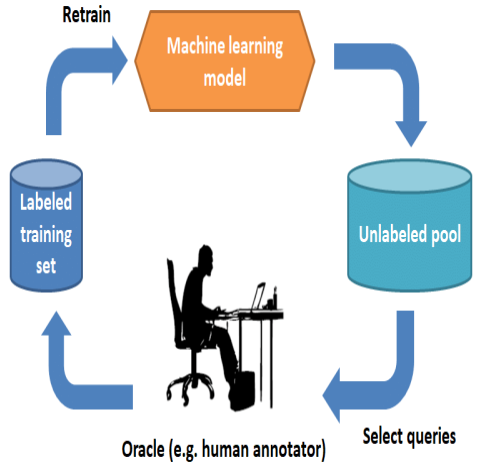
Gilbert Wong
Loyola University Maryland

May 5, 2020

Introduction

- Problem: Providing the client a capability to obtain labeled data through Active Learning.

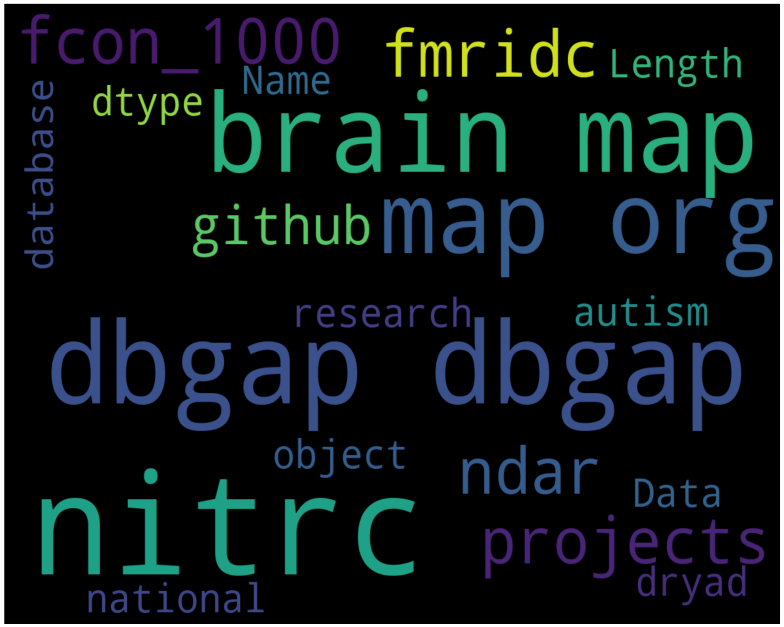
National Institute of Mental Health

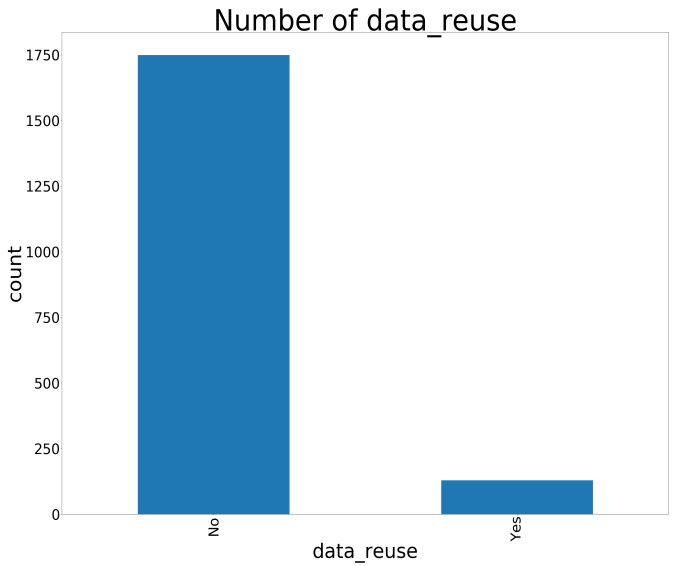


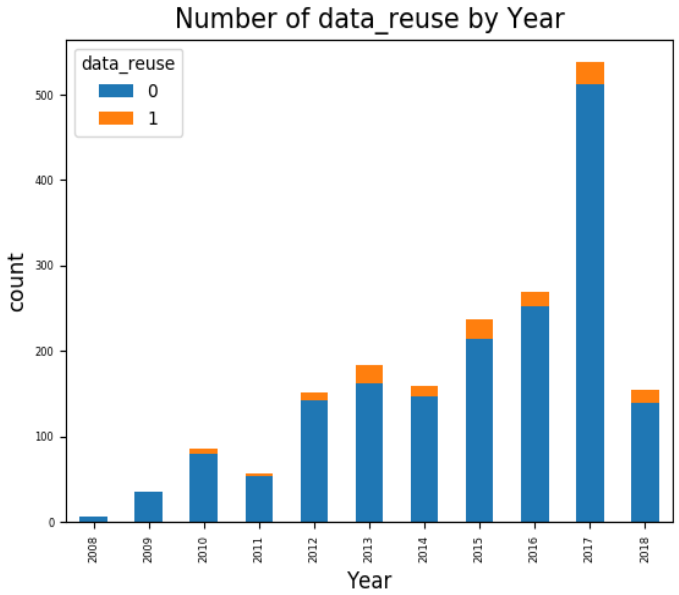
Dataset/Experiments

- Dataset
  - Initial training set consists of 1893 records, 15 records with missing 'text' field for a total of 1878 records (684 unique papers).
  - Obtained more data using a REST API (40,808 papers retrieved).
  - Filtered the 40,808 papers using regular expressions and resulted in 2674 records (1261 unique papers).
- Experiments
  - Vectorizer methods (CountVectorizer and TfidfVectorizer)
  - Normal machine learning workflow
  - Active learning (interactive labeling)

Exploratory data analysis







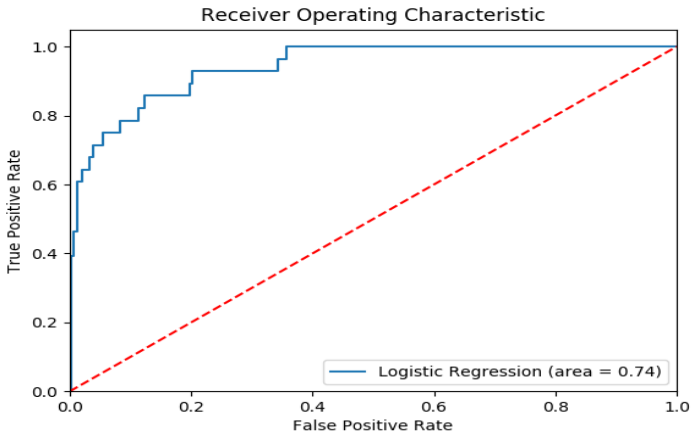
Normal machine learning workflow analysis

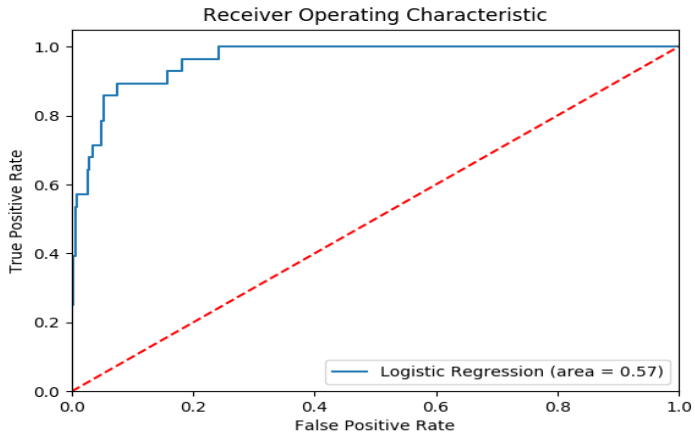
- Preprocessor and Tokenizer (sample tokens)
  - www.yeastgenome.org', 'www.wtccc.org.uk', 'www.sfari.org', 'www.scandb.org', 'www.r-project.org', 'www.qiagen.com', 'www.python.org', 'www.pubatlas.org'
- Four ML algorithms (Logistic Regression, Naive Bayes, Support Vector Machines, Random Forest)
- Classification reports (CountVectorizer/TfidfVectorizer)

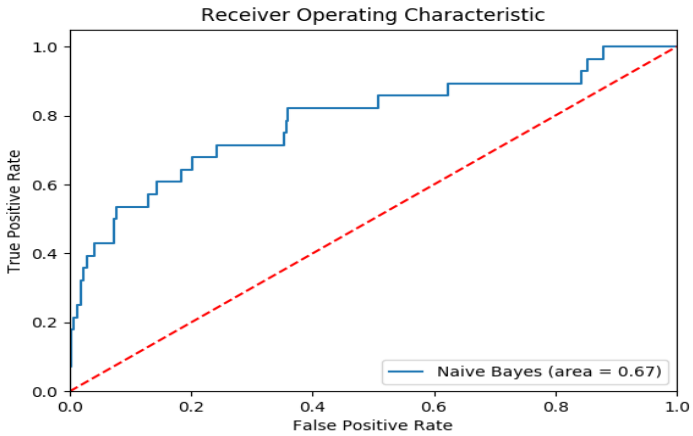
Metrics	Logistic Regression	Naive Bayes	SVM	Random Forest
Recall	0.50/0.14	0.36/0.00	0.57/0.54	0.29/0.18
Macro-avg	0.74/0.57	0.67/0.50	0.77/0.76	0.64/0.59

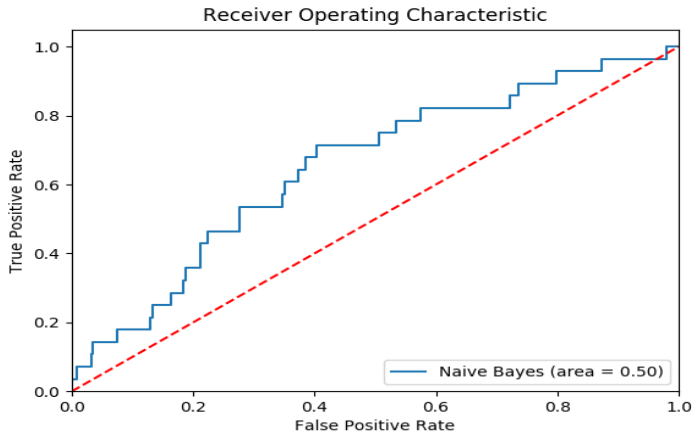
Normal machine learning workflow analysis cont.

- ROC graphs

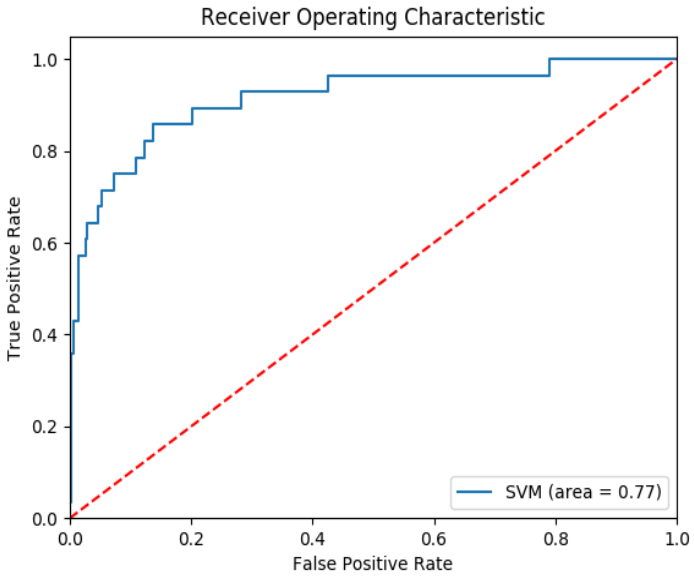


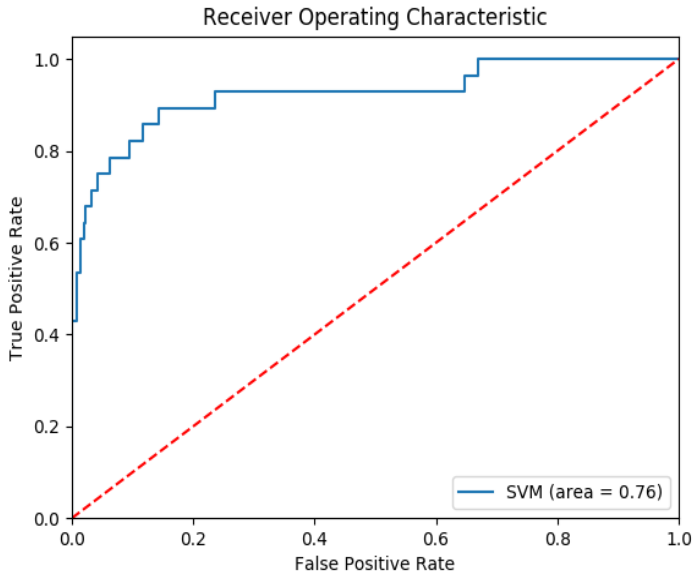


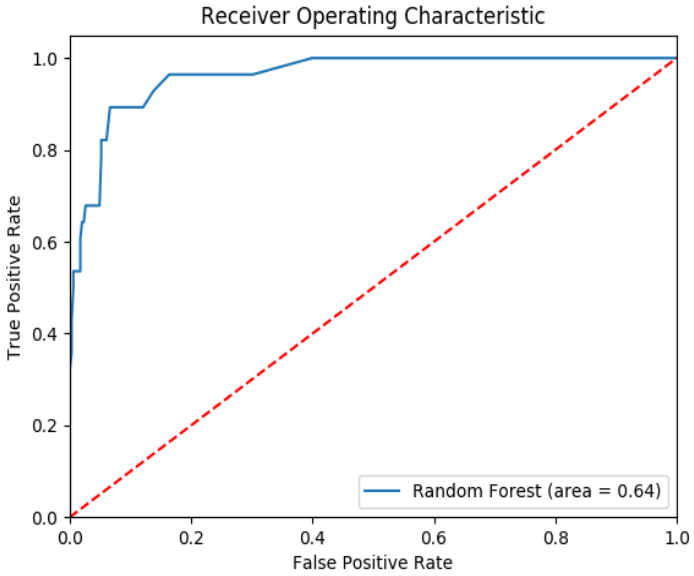


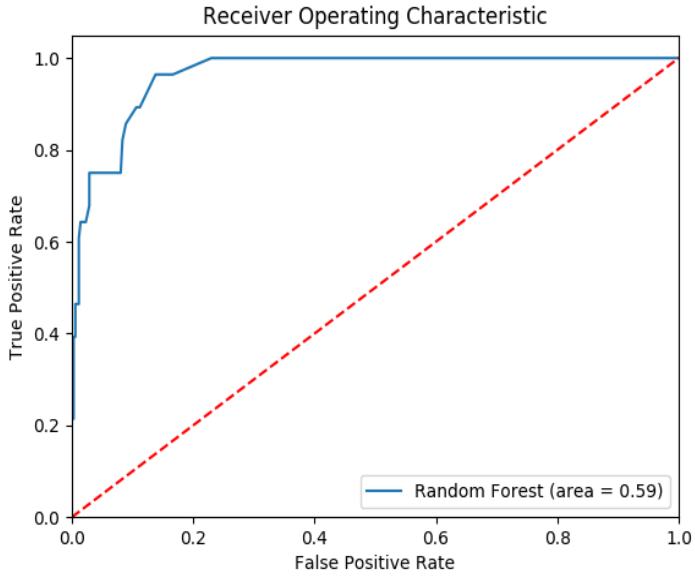


Normal machine learning workflow analysis cont.





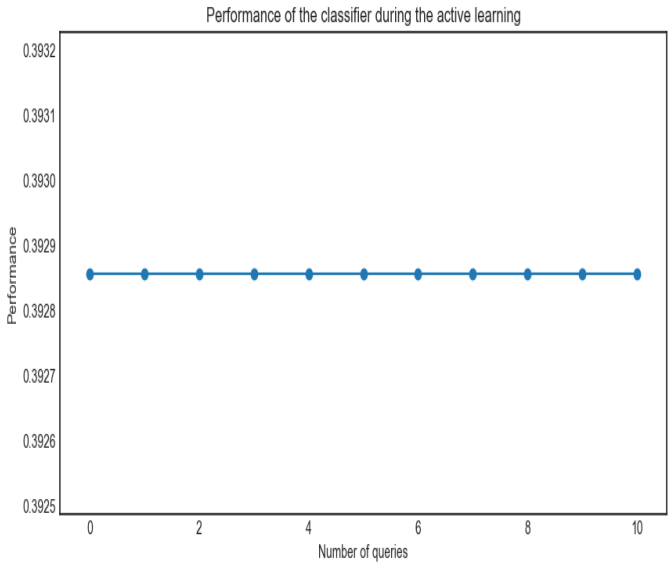


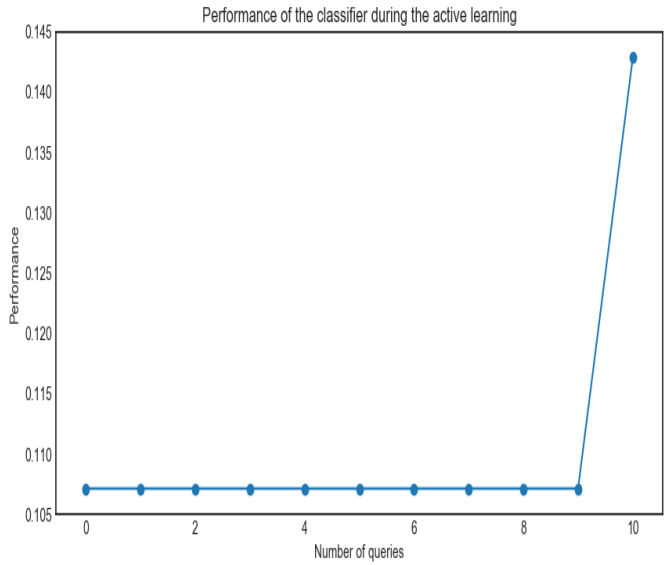


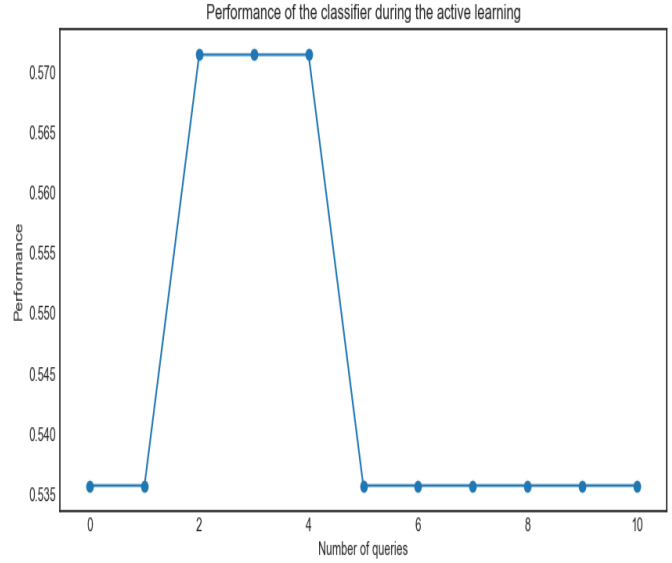
Active learning workflow analysis

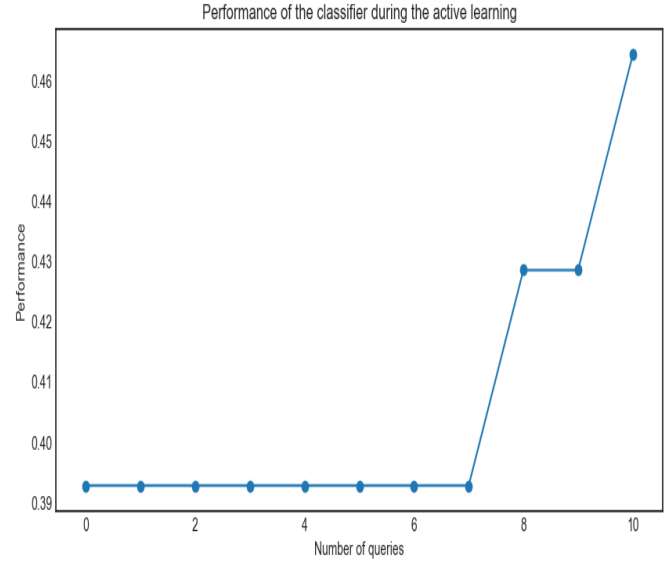
- Interactive labeling
- Allow users to choose the following:
  - Vectorizer method
  - Machine learning algorithm
  - Query strategy
  - Number of queries
- Example output of program
  - Diagrams in d and g are modified from the Allen Mouse Brain Atlas, Allen Institute for Brain Science; available from http://mouse.brain-map.org/.
    Is this a data reuse statement or not (1=yes, 0=no)?
    1
  - Data from mouse connectivity map of Allen Brain Atlas: id 263242463, http://connectivity.brain-map.org/).
    Is this a data reuse statement or not (1=yes, 0=no)?
    1

Active learning workflow analysis cont.

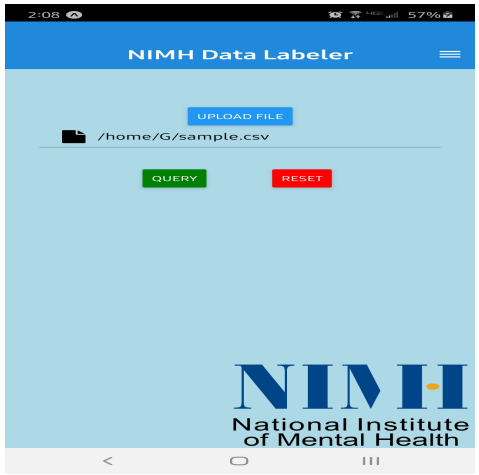


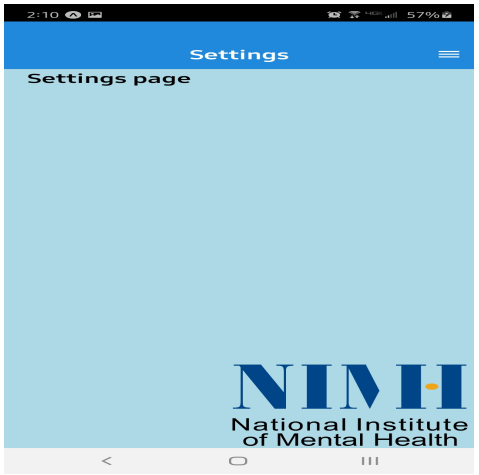


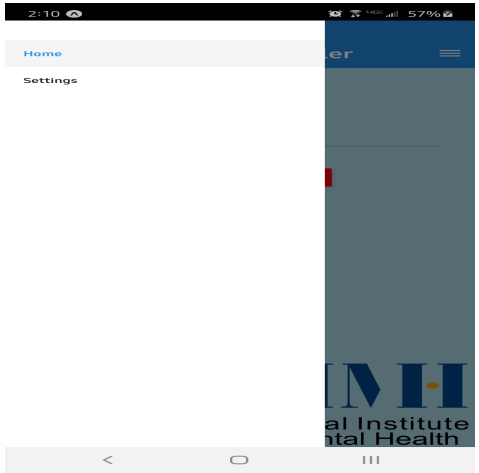




Mobile application deployment







Conclusion

Git repository

https://github.com/gwong11/data-science-project

- Working with unstructured data like text has its challenges.
- Choosing the right vectorizer and how you perform the preprocessing/tokenizing step affects the model performance.
- Both Logistic Regression and Support Vector Machines saw a better performance.
- Number of queries performed during active learning determines how the model is improving.

Questions?