Assignment 3: Sentiment Analysis on Amazon Reviews

Apala Guha

CMPT 733

Spring 2017

Readings The following readings are highly recommended before/while doing this assignment:

- Sentiment analysis survey:
  - *Opinion Mining and Sentiment Analysis*, Bo Pang and Lillian Lee, Foundations and trends in information retrieval 2008.
  - *Opinion Mining and Sentiment Analysis*, Bing Liu, Web Data Mining, 2011.
- Sentiment analysis tutorial at `https://www.kaggle.com/c/word2vec-nlp-tutorial`
- Spark TF-IDF: Spark documentation and Ryza et al. Page 105.

1 Introduction

In this assignment, we will use as our training dataset `reviews_Pet_supplies_5.json.gz` and as our test dataset `reviews_Pet_Supplies.json.gz`. Please see the file format at `http://jmcauley.ucsd.edu/data/amazon/`. We are interested in `reviewText` (the review) and `overall` (the rating) columns.

2 Preprocessing

Clean the reviews by converting them to lower case, splitting into tokens at whitespaces and characters that are not letters, and removing stop words.

3 TF-IDF

Compute the TF-IDF vectors for each review and build a linear regression model for the ratings using the training dataset. Use an appropriate hash table size for computing the TF vectors and normalize them. Also, since the vectors will be large, use an appropriate regularization parameter. Use 5-fold cross validation to build the model on the training dataset. Report the RMSE error on the training and test datasets.

4 Word2Vec

Compute Word2Vec vectors for each review. Use the default `PySpark ML` parameters. Build a linear regression model for the ratings and set appropriate regularization. Use 5-fold cross-validation to build the model on the training dataset. Report the RMSE error on the training and test datasets.

5 Word2Vec Clustering

Take the vocabulary of Word2Vec vectors and cluster them using kmeans. Use a cluster count that is appropriate for the given vocabulary size. Study some of the word clusters to understand how similar words are clustered. List two clusters (as set of words) that you found interesting. Use the clustering to map reviews into cluster frequency vectors. Normalize the vectors and build a linear regression model with the training set using 5-fold cross validation. Use appropriate regularization. Report the RMSE error on the training and test datasets.