ANLY 565 Final Project

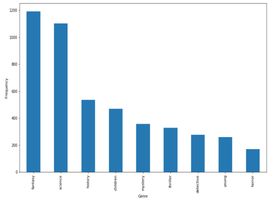
Various Dimension Reduction Techniques for Book Recommendation and Classification

Weile Chen, Yanou Yang, Siyao Peng, Xuejun Zhang

Introduction & Data Preprocessing

This project achieves recommendation system and classification for books on Wikipedia by applying different dimension reduction techniques, including LSA, LDA, PCA and UMAP.

The Number of Books By Genre



Data Preprocessing:
35991 books have wiki pages -> 20080 wiki pages include plots -> 11000 after cleaning:
Remove outliers, non-english words, lemmatization, tokenization;
Training set: 9212(80%)
Testing set: 2302 (20%)

Methodology

Principal Component Analysis (PCA)

Principal Component Analysis seeks a N-dimensional linear subspace that solve

$$min \sum_{i-1}^{n} ||X^{(i)} - Proj_v X^{(i)}||^2$$

PCA mainly has three steps:

1. Compute the covariance matrix of the data
2. Compute the eigenvalues and eigenvectors to select only the most important principle components and transform your data into uncorrelated linear combinations to reduce dimensionality

Latent Dirichlet Allocation (LDA)

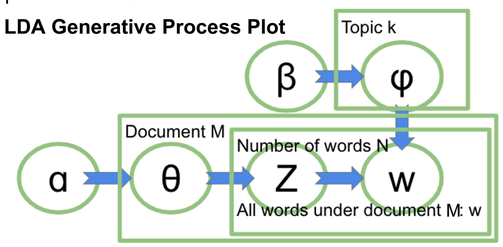
LDA assumes the generative process for a corpus with M documents each of length Mi (Document i has Mi words):

α : Parameter of the Dirichlet prior on the per-document topic distributions
β : Parameter of the Dirichlet prior on the per-topic word distribution
θi : Topic distribution for document i
φ_k: Word distribution for topic k

1. Choose θi ~ Dir(α) where i ϵ {1,2,3,...,M} and Dir(α) is a Dirichlet distribution
2. Choose φ_k ~ Dir(β) where i ϵ {1,2,3,...,K}
3. For each word positioned at i,j, where i ϵ {1,2,3,...,M}, j ϵ {1,2,3,...,Ni}
   a. Choose a topic Z_{ij} ~ Polynomial(θi)
   b. Choose a word W_{ij}~Polynomial(φ_{zij})

Briefly, the LDA that each document mix with various topics and every topic mix with various words

LDA Generative Process Plot



Latent Semantic Analysis(LSA)

First, X presents documents by TF-IDF:

$$X = f(t,d) log \frac{N}{|\{d \in D : t \in d\}|}$$

Next, LSA uses Singular Value Decomposition technique to reduce the dimensionality of input

$$X = U\Sigma V^T$$

Umap

UMAP can preserve the data's global structure very well.It uses graph layout algorithms to arrange data in low-dimensional space, making it structurally similar as in high dimension.

Results

Method	Coherence Score	Num Topics	TPR
LDA	0.42	37	0.06
LSA	0.46	6	0.06

Table 1: Books Recommendation System Results

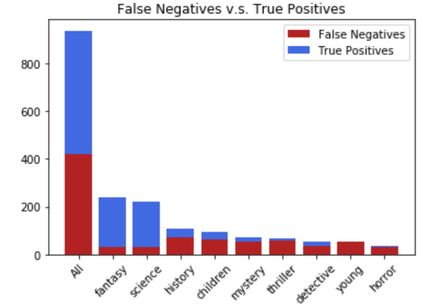
Method	Recall	Precision	Accuracy
PCA	0.10	0.06	0.25
LDA	0.36	0.50	0.52
LSA	0.38	0.68	0.55

Table 2: Books Multiclass Classification Results

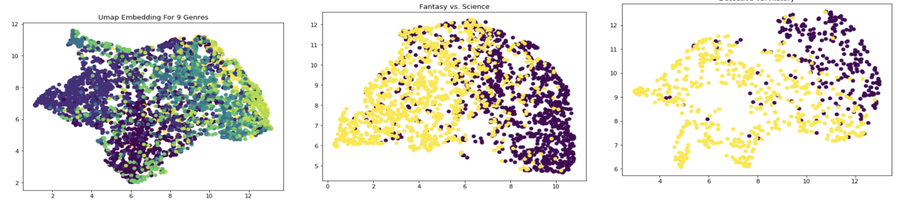
Book Recommendation System Demo

PCA Multiclass Classification

LSA Multiclass Classification



UMAP Embedding After LSA Topic Modeling



For the book recommendation part, topic modeling techniques including LDA and LSA are performed to about 11,000 books. Results are shown in Table 1. Coherence measures the score of a single topic by measuring the degree of semantic similarity between high scoring words in the topic, a higher coherence score is better. We measure the true positive rate by splitting each book in the testing set into two halves. The first half is used as input, the second half is incorporated into the training set as candidate pool(~11000 book plots), select 10 out of the pool, if the second half of the input book is matched with the 1 out of 10 book plots, we consider it as true positive.

For the multiclass classification part, 4000 books are classified based on dimension reduction techniques including PCA, LDA, LSA. Plots for different books vary a lot, we only consider plot size from 50 to 2000. There are totally 9 genres of books. The results are shown in Table 2.

Conclusion

In the book recommendation part, the result is good despite that the true positive rate is very low. The reason is the measurement is very strict as stated in the result section. Moreover, the possibilities of 10 out of 11,000 that the first half plots token match with the second half plot token is only about 0.09%, and the model team made is 6%.

In the classification part, although both LDA and LSA have better results than PCA, they still do not reach the benchmark. The main reason is that the size of sample is not large enough and each sample in the training set does not contain enough tokens, so both models could not estimate the true parameters at high accuracy.

PCA also performs badly in NLP classification problem, because some distinguishing words always have low contribution to explain variance,while PCA select those who contribute the most. Besides, the input matrix of PCA are sparse matrix, which might cause some inaccuracy when calculating inverse matrix.

Next Step

- Data Set: Incorporate more information like author, country, language. People tend to read books from the same author and books of their mother language. Besides, we should improve the data quality by removing documents with fewer words as well as collecting more documents.
- Methodology: Use word2vec for word embedding to better understand the plots of books.
- Solve data imbalance problem before classification.
- Add user data to do collaborative filtering.

Reference

[1] Building a LDA-based Book Recommender System. (2020). Retrieved 1 May 2020, from https://humboldt-wi.github.io/blog/research/information_systems_1819/is_lda_final/
[2] Wall, R., 1971. *Introduction To Information Retrieval [Sammensat Materiale].*
[3] Landauer, T., 2014. *Handbook Of Latent Semantic Analysis.* New York: Routledge.
[4]Ghosh, J., Delampady, M., & Samanta, T. (2006). *An introduction to Bayesian analysis.* Springer Verlag New York.

https://github.com/SteveChenWL/Various-Dimension-Reduction-Techniques-for-Book-Recommendation-and-Classification