**ANLY580 Project #2 Proposal**

**Topic: A LDA-Based Book Recommender Engine**

*AsianFusion66*

Members:

Weile Chen, Shaoyu Feng, Yunjia Zeng, Mengtong Zhang

* **Motivation**

A recommender system is a computer program that helps a user discover products and content by predicting the user’s rating of each item and showing them the items that they would rate highly. Recommendation systems are everywhere. Recommender system plays an important role helping users find the product and content they want. In our project, we want to build a book recommender system based on Latent Dirichlet Allocation(LDA) model, which is widely used for topic modeling and opinion analysis.

* **Relevant Literature**

Blei, Ng and Jordan proposed Latent Dirichlet Allocation (LDA) model in 2003[1]. LDA is being applied in various Natural Language Processing tasks such as for topic modeling [2], and for learning word classes[3]. Also, there has already been some research in using LDA to power recommender system. [4]

* **Data**

For this project, we will use Wikipedia book data. Example data format: [*https://github.com/WillKoehrsen/wikipedia-data-science/blob/master/data/found\_books\_filtered.ndjson*](https://github.com/WillKoehrsen/wikipedia-data-science/blob/master/data/found_books_filtered.ndjson)

* **Plan**

In this project, we will first use LDA model for topic modeling and find the best number of topic with the word representations of each topic. Then we will build a content-based model recommender system based on document similarity. At this point the algorithm is fully content based, but if later on user data is available, we may also consider building high-level business implementable hybrid recommender system.

* **Expected Results**

Our content-based recommender system aims to help users find books which they may be interested in based on the books the users read or liked before. Given a text document, our model will generate a topic classification result, as well as the top recommended books for the document.

[1] *(Blei, D.M., Ng, A.Y., Jordan, M.I. Latent Dirichlet allocation. Journal of Machine Learning Research 3 (2003a) 993–1022.)*

[2] *Jojo, S.m., Wong, M., Dras, M.J.: Topic modeling for native language iden- tification. In: Proceedings of the Australasian Language Technology Association Workshop, Canberra, Australia (2011) 115–124*

[3] *Chrupala,G.: Efficient induction of probabilistic word classes with LDA. In: Proceedings of the Fifth International Joint Conference on Natural Language Processing (IJCNLP 2011). (2011) 363–372*

*[4] Ralf K.,* *Peter F.: Latent dirichlet allocation for tag recommendation (2009)*