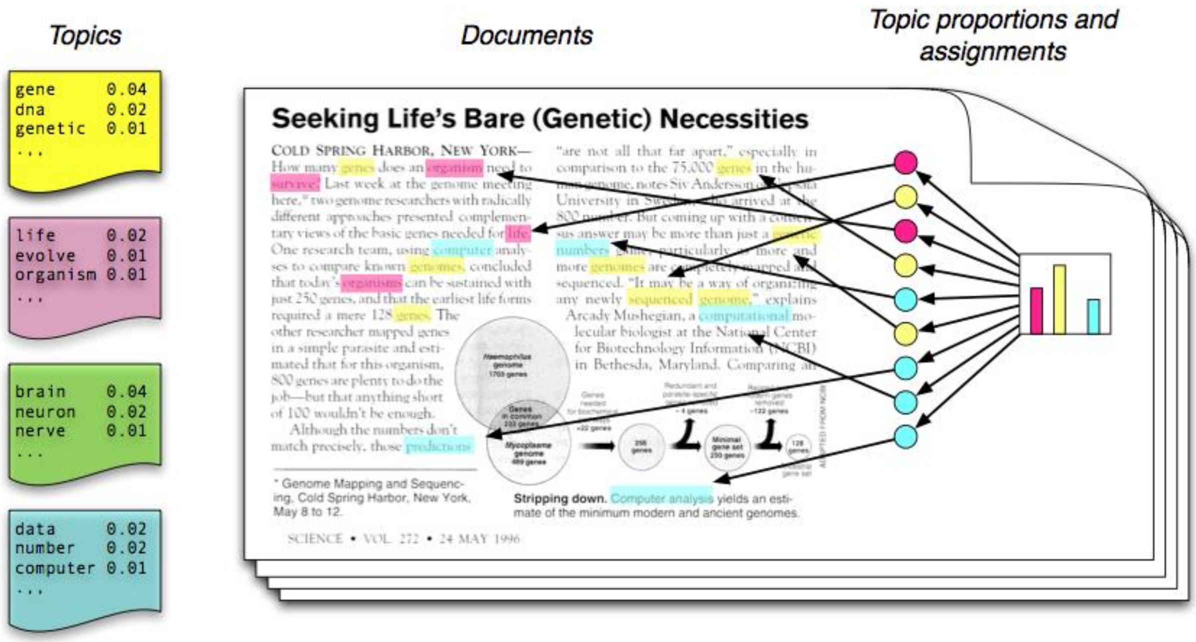
MOOC Data Analytics:
Probabilistic Topic Modeling of Discussion Forum Data

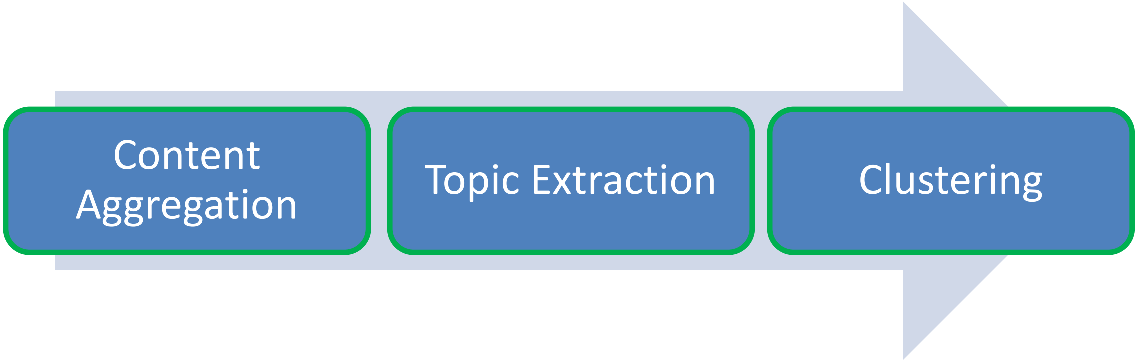
Lei Sun
Supervised by Prof. Dit-Yan Yeung

Introduction

The MOOC data, including students' course record and behaviors on forum may provide an approach to improve online learning experiences. This project aims to utilize the topic distribution of the aggregated forum posts from users to explore cluster information and even predict user’s curriculum performance.



Methodology


Content Aggregation

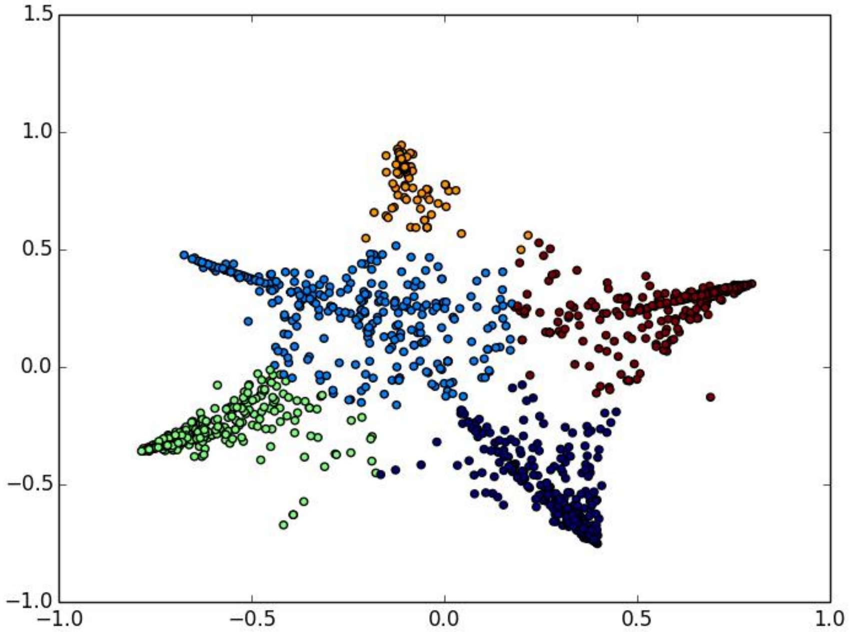
Topic Extraction

Clustering

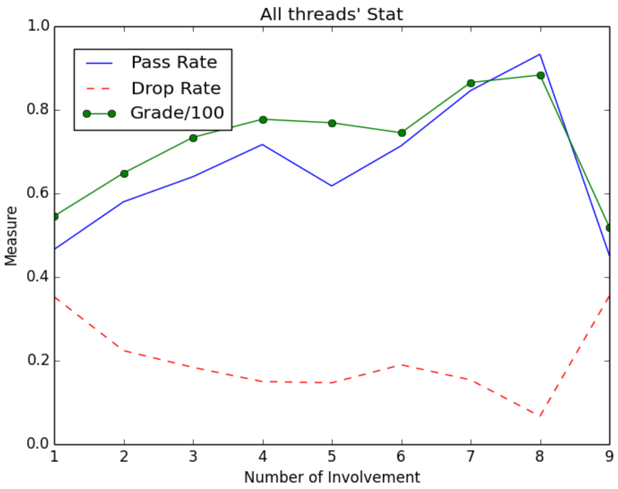
The methods to aggregate different posts into bigger documents is one of the key factors that influence the clustering result. Also the different aggregation methods imply different directions to utilize the cluster information for further prediction. Mainly two approaches are explored in this study:
(1). User-based aggregation with one thread/ week.
(2). Thread-based aggregation.

The LDA probabilistic topic modeling is used for topic extraction in this study.

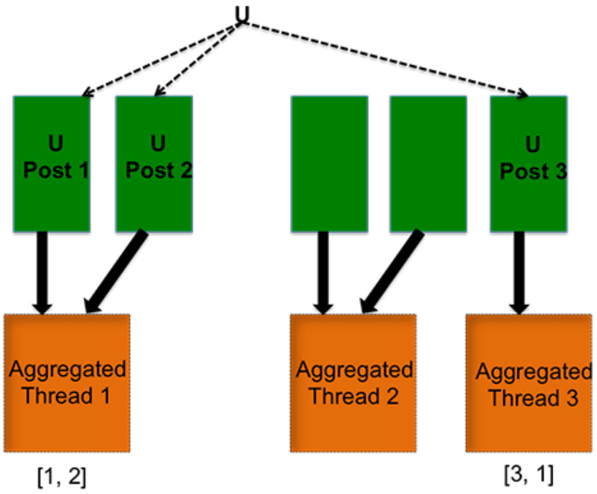
Feature Exploration



The thread-based aggregation made all the posts inside one discussion forum thread as a single super-thread, so that the total number of threads can be reduced by regarding multiple similar threads in terms of their topic distribution as a single one.

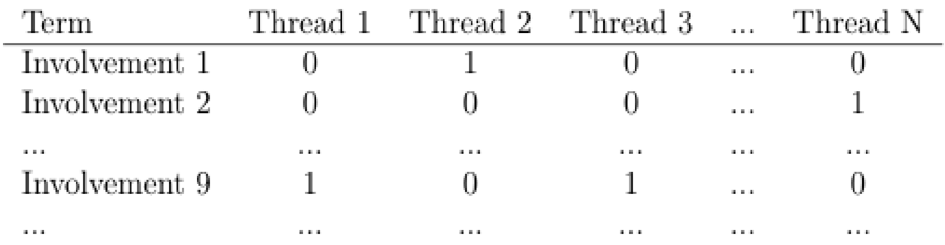


The statistical information inside one super-thread is explored for their potential power to demonstrate any tendency on users' performance who talked in those threads. It is shown that more times that a user talked (higher involvement), the higher performance will she achieved in some clusters.

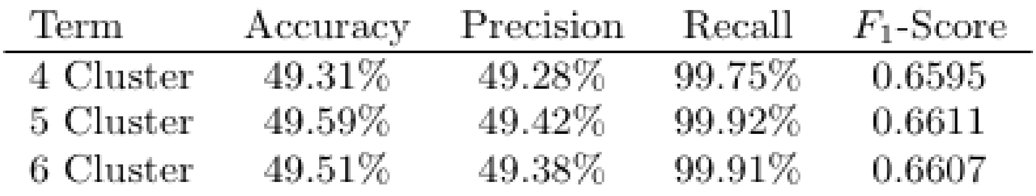


Prediction

The prediction model is like a simple 0/1 tow-dimensional table that predict whether a user pass or fail in the course, trained by Decision-tree like model.

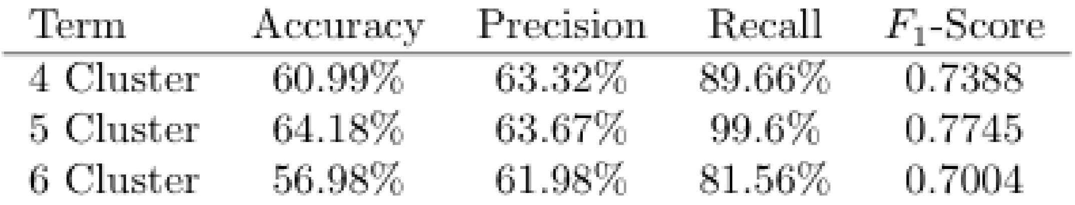


Term	Thread 1	Thread 2	Thread 3	...	Thread N
Involvement 1	0	1	0	...	0
Involvement 2	0	0	0	...	1
...	...	...	...	...	...
Involvement 9	1	0	1	...	0
...	...	...	...	...	...



Term	Accuracy	Precision	Recall	F_1-Score
4 Cluster	49.31%	49.28%	99.75%	0.6595
5 Cluster	49.59%	49.42%	99.92%	0.6611
6 Cluster	49.51%	49.38%	99.91%	0.6607

By removing some “confident users”, which are confidently predicted (Receive high score in SVM model) by other non-content related features from data set, the performance can be improved.



Term	Accuracy	Precision	Recall	F_1-Score
4 Cluster	60.99%	63.32%	89.66%	0.7388
5 Cluster	64.18%	63.67%	99.6%	0.7745
6 Cluster	56.98%	61.98%	81.56%	0.7004