# Team

Team name: textminer

Team Members

- Angus Jyu (angusfj2)

- Manu Vinod Shesha (manuv3)

# Topic

The topic is “Intellijent Topic Modeling and Index building of Course on EducationalWeb[1]”. This loosely aligns to Theme 2: Intelligent Learning Platform, with sub-area of “ConceptView”.

# Problem

Courses like “CS-410 Text Mining and Text Analysis” are presented on MOOCs (like Coursera/EducationWeb) as a series of lectures, each lecture roughly discussing one main idea and some related concepts. However, the title of a lecture video may reveal underneath topics/concepts, to a small degree. As a user, it may be difficult to find out where was a topic or concept discussed in all of course lectures. The other challenge is that many times individual lectures contain more than one topic, and all the topics may not be apparent by just looking at the lecture title.

# Proposal

We believe Topic Modeling may be one effective way to solve this challenge. Intuitively, a list of topics extracted from a particular lecture transcript, should provide a snapshot of what the lecture covers. At the same time, such a model, once built, for a Course, can act as building block for building other solutions which can ease user’s learning experience. As a possible application of such a model, we want to demonstrate a topic “index” for the course.

So, at high level, we want to:

- discover primary topic(s) as well as secondary topics in a given lecture. Here, we want to do comparative analysis of algorithms: PLSA and LDA, to measure the relative effectiveness of these algorithms in a paradigm of scientific/technical corpus (like course transcript). This will also give us opportunity to compare effectiveness and ease of use of tools like Lemur, Gensim etc.

- explore ways to do Automated Labeling, to create phrases (like 2-gram phrases), which define the “concepts” in more meaningful ways, than just “bag of word” representation of topics in traditional sense of Topic Modeling. For example: individual topics like “Dirichlet” and “distribution” can be more meaningful for a user if we can do a bit of semantic analysis and discover labels like “Dirichlet distribution”. The idea is to use techniques which can improve parameters: **Topic Coverage** (which means that discovered topics fully represent the document) and **Topic Differentiation** (which means that the topics are specific enough to differentiate between two documents). Here we want to explore the techniques like using NLP “chunkers”, and “Context Model”. The inspiration is the research paper by Professor C Zhai: **Automatic Labeling of Multinomial Topic Models.**

*(Please note that this is an analysis/discovery goal for this project, and may not necessarily apply to our final solution. At minimum, we will discuss what techniques were applied and how it impacted the topic models)*

- extend EducationalWeb UI to provide the index-like representation of the topic models, which can provide a mapping of topics to relevant lectures. This will greatly enhance the learning experience as it will be useful for the learner to quickly navigate to specific segments of lecture based on "topic". This idea goes beyond the basic search provided by EducationalWeb, where a user can search for a topic and this search task is treated as a simple "bag of word" search to give a whole list of results which may and may not be relevant for the user. Also, a traditional search on such learning platforms are not comparable to full-fledged browser, and here, the assumption is that the user has some idea on what he is searching for. Instead, we want to provide a push-kind of model, where we generate an index of relevant topics covered in a course (spread across a series of lectures).

*(As an option, we will try to index specific video segments, based on topics. However, this is an optional goal for this project, depending on current capability of EducationalWeb)*

# **Datasets, Algorithms and Tools**

## Datasets

* For our primary goal of topic mining for lectures:

- CS410 course transcipts available on EducationWeb/Coursera as our Collection Corpus. This means that “N” documents refer to number course lectures we have.

- Human generated topics (by TAs and former students) for the lectures here [add reference link], to evaluate our models.

- Built-in Background language models, in tools like Gensim, Lemur etc.

* For secondary goal of Automatic labeling:

- Abstracts of scholarly articles related to Text Analysis and Mining, from the ACM digitallibrary[add reference link] and Google Scholar [add reference link]

## Algorithms

We plan to use PLSA (Probabilistic Latent Symantic Analysic) or LDA (Latent Diriclet Allocation) to generate the topic models.

For automated labeling, we would use the techniques like Chunking, Ngram testing, Semantic Relevance scoring and Context model, discussed in the paper [add reference link].

## Tools

- Gensim

- EducationalWeb

## Language

Python

References:

[1] http://timan102.cs.illinois.edu/explanation/

[2] https://docs.google.com/spreadsheets/d/1V\_PvgwqUifq6QlmXl-0EDQutaCVErYfSDl09HJwE6W0

[3] <https://radimrehurek.com/gensim/>

[4] <http://www-personal.umich.edu/~qmei/pub/kdd07-label.pdf>

[5] <http://www.acm.org/dl>

[6] <https://scholar.google.com/>