Team: Early Birds

What are the names and NetIDs of all your team members?

First	Last	NetID	
Anthony	Ghabour	ghabour2	Team Captain
Maciej	Wieczorek	maciejw2	
Quan	Nguyen	quanhn2	
Rick	Suggs	rsuggs2	
Tina	Tang	ytang30	

What topic have you chosen?

Theme: Intelligent Learning Platform

The goal of this project is to implement enhancements to an existing intelligent learning platform (specifically, *SmartMOOCs*) and facilitate a more intuitive and efficient learning experience. In particular, this project aims to:

- incorporate keyword and query search functionality **(*primary objective*)**, and
- identify better ways to segment lectures based on topic transitions **(*stretch goal)***.

Why is it a problem?

Online learning platforms and MOOCs such as Coursera have democratized education and enabled millions of students to access knowledge at a previously unthinkable scale. These platforms host thousands of hours of extremely rich content that not only expose their audiences to valuable concepts and ideas, but also serve as crucial reference materials after the content has been initially viewed (for example, while studying for an exam or upon attempting to implement a learned concept in the real world).

However, in their current state, most MOOCs are not fully mature platforms and do not provide users with effective tools to quickly find and access specific content. In general, revisiting a target topic requires considerable effort on the part of the user and often involves clicking through multiple (potentially) relevant lectures and exhaustively scrubbing through videos to access the concept of interest.

SmartMOOCs is an experimental MOOC platform at UIUC aimed at developing technology to enhance the student experience with MOOCs, and can serve as a demonstration testbed for advanced features. At present, the lectures on SmartMOOCs are uniformly partitioned into one-minute segments acting as proto-topics which does not accurately represent topics. A more

effective approach could be to segment lectures based on the slides in a slideshow - one slide per topic. This approach benefits from using the same paradigm of information segmentation used in creating the slideshows, where each slide typically contains one key idea. A topic found based on the transcript or slide text would thus have a likely logical beginning and end, which creates new opportunities for presenting results to a user, for example as a playlist.

How does it relate to the theme and to the class?

The primary theme of this class (especially with respect to the Text Retrieval portion of the course) involves processing user queries and specifying document relevance. In the context of this project, documents will correspond to transcripts or slide text from the lectures such that applying many of the tools and techniques learned in the course will be central to effectively executing this objective. *See next section for details.* If we are able to accomplish the ***stretch goal***, these techniques will be applied not only across course lectures, but also *within* lectures to identify timestamps specifying precisely when/where queried concepts are covered in a lecture.

Datasets, Algorithms and Techniques

This data for this project will consist entirely of lecture videos, slides and english transcripts (possibly including timestamps) downloaded from Coursera, as of the writing of this proposal.

We plan to implement the following algorithms and techniques:

- Data preparation and processing
    - Eliminating stop words
    - Tokenization
- Search Engine (“Pull Mode”) System Implementation
    - Document representation and indexing
    - Ranking functions (BM25/Okapi)
        - Term frequency weighting
        - Inverse document frequency weighting
        - Document length normalization
- Text Retrieval System Evaluation
    - Basic measures (accuracy, precision, recall, F-score)
    - Ranked list measures (precision@n, MAP & gMAP)

Stretch goal implementation would also involve:

- Offline processing of lecture videos for topic detection on slide transitions

How will you demonstrate that your approach will work as expected?

We will demonstrate our approach worked by comparing retrieved results against target results for a range of predetermined queries as judged by human users. The test corpus will be the contents of CS410 so that the human verifiers will have appropriate domain knowledge to make judgements.

Which programming language do you plan to use?

- Back-end: Python (metapy) for text retrieval and video processing
- Front-end: Javascript(React), HTML, CSS

Projected Workload

Given the ambitious scope, we are confident that total hours required to complete this project will exceed 20*N = 100 hours, (N=5). We optimistically anticipate the core tasks involved will be completed within the following time commitments:

Task	**Est. Hours**
Data Collection and Cleaning	10 - 15
Database Development	10 - 15
Search Engine Development	10 - 15
Hosting Infrastructure	10 - 15
Front-end Development	10 - 15
Architecture/API	10 - 15
Video processing	10 - 15
Integration and Testing	10 - 15
Documentation	10 - 15
Demo and Presentation	10 - 15
TOTAL	**100 - 150**