Authors:

Magd Bayoumi (Piazza) **<mb2363@cornell.edu>**
Jenna Kressin (Piazza) **<jek343@cornell.edu>**
Souleiman Benhida (Piazza) **<sb2342@cornell.edu>**
Sheetal Athrey (Piazza) **<spa42@cornell.edu>**
Oscar So (Piazza) **<ons4@cornell.edu>**

GitHub Link: https://github.com/oscarso2000/Piazza
Piazza Link: https://piazza.com/class/k5h3t0fy1gm5vv?cid=418 (@418)
URL: https://www.mycourseindex.com/

High level goal/Use case:

This tool aims to help students relevant course materials to their coursework questions.
Use case: User writes query i.e. “Fast cosine similarity on Kardashians transcripts”
=> Results: Piazza @26 @112 (for example)
=> Link to worksheet from website on this
Inputs are free text and outputs are links/blobs of text

Data Sources:

The primary data source we would be using are existing Piazza questions and answers as a corpus to the query form, this will help us with our aim to improve the way Piazza search queries work. Other data sources that we plan to use are course notes, powerpoint presentations, assignments as well as if available the course video lectures (and transcripts), and the course textbook (if free). Another source we plan to possibly use is the User’s own notes that they would upload. These additional sources will be what we would use to build the tool that would use keyword queries and give relevant excerpts from piazza answers as well as the course site.

Input:

The user would input a single free form keyword query that could be a concept, course material question, or technique. In addition, the user would specify the course that we should search within.

Output:

A user would receive an output in the form of a list of one or multiple relevant Piazza questions and answers, or an excerpt or page from the respective, inputted course site.

Social Component:

We are using a question-answer forum in Piazza as our human generated data. This offers a rich environment with questions and answers that we plan to use in our information retrieval component as a way to answer their information needs. Piazza also offers information about whether a post has been answered and is considered a “good” post. To further encourage social engagement, we can prompt users to ask their question on Piazza if no good matches are found in order to enhance the corpus. Additionally, we can gather pseudo relevance feedback by noting which Piazza posts and resources are clicked on in order to improve results.

Information Retrieval Component:
Hopefully, we will have access to a bot inside a Piazza course which will allow us to web scrape and gain information on any posts. Thus, we will then be able to preprocess the data and compare user searches with TF-IDF, Cosine Similarity, etc. Thus, this will allow us to get more relevant and similar piazza posts to answer any queries that the user has.

Machine Learning Component:
Another critical feature for our users is finding related concepts in their queries. We plan to train a BERT model to tokenize keywords or concepts and map them to a higher order concept (similar to an NER model) on a per class basis.