Text Miners Team Project

1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.

Team member	**NetId**
Ashish Pradhan [Captain]	apradh6@illinois.edu
Kirti Magadum	magadum2@illinois.edu
Bhuvaneswari Periasamy	bp14@illinois.edu

2. What system have you chosen? Which subtopic(s) under the system?

   We have chosen to enhance the Educational Web System and intend to add new features, focusing on the following subtopics:

- **Scale up the current system**

  Add more slides and courses from multiple sources e.g. Coursera, UIUC courses, etc. and run the existing algorithms on them.

  We intend to create an automatic crawler which could classify a webpage containing slides correctly and subsequently download them.

- **Allow downloading slides in bulk:**

  Downloading the entire collection of slides for a particular course or interest.

  Our goal is to primarily focus on creating a successful crawler and then aim for creating the ability to allow slides in bulk after that.

3. Briefly describe the datasets, algorithms or techniques you plan to use

Dataset	Create a dataset of all the UIUC pages which has slides available for download and also some random pages which are "negative" examples.
Algorithms	Dirichlet prior, EM Algorithm, All algorithms currently in use. For classification, we intend to use all the standard available ones such as SVM, Clustering, Logistic Regression to achieve maximum accuracy.

Techniques	Automatic web crawler. All techniques currently in use.

4. If you are adding a function, how will you demonstrate that it works as expected? If you are improving a function, how will you show your implementation actually works better?
    - **Scale up the current system**

      We are adding an automatic web crawler. We will crawl data related to selected courses from the mentioned dataset. Users should be able to select newly added courses and our updated code should be able to display relevant data to users.
      Also upon selection of a newly added course, options related to course should get updated.
        - Courses [Newly added courses]
        - Recently Visited slides
        - Lectures
        - Search result.

- **Allow downloading slides in bulk**

  We will enhance the existing functionality from downloading a single slide into downloading multiple slides that fits under the same course/slide heading.

5. How will your code communicate with or utilize the system? It is also fine to build your own systems, just please state your plan clearly
   We will aim to automate the crawling to feed slides directly into the system so that users will not have to create a zip file and then manually upload it into the system.
   For the second part, we will aim to provide the user two options: to download the slide for a particular lecture or the whole course (as a zip file).

6. Which programming language do you plan to use?
   We will be writing our code predominantly in Python but might implement Javascript, CSS and HTML if needed.

7. Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

- **Scale up the current system**

Task	**Hours (hours * N team member)**
Detailed understanding code for existing system	6 (2 * 3)
User Interface changes	3 (1 * 3)
Automatic Web crawling (Including creation of training dataset and a classification model)	33 (11* 3)
Deployment and testing	12 (4 * 3)
Bug Fix	3 (1 * 3)
Performance Test	3 (1 * 3)
Total	60 Hours

8. **Allow downloading slides in bulk**

   NOTE: Any remaining time from the above section will be utilized to complete this task.