Stage 1

Dan Howe and Keith Funkhouser
CS838: Data Science
February 8, 2017

- Experimental questions
  - Can we infer insights about courses in UW-Madison's CS department over the last two decades?
    - Has there been grade inflation?
    - Are there professors who should be taken to maximize GPA?
    - Are there professors who should be taken to maximize happiness with the course?
  - Can we build an intuitive interface for browsing historical grade and student review data for courses in UW-Madison's CS department?
- Data sources
  - Structured:
    - **Grade distributions**: grade distributions for all UW-Madison courses are publicly available (https://registrar.wisc.edu/course_grade_distributions.htm).
    - **(Pre-2013) course teaching history**: the CS department used to maintain a tool allowing users to search previous student ratings of courses in the department (http://www.cs.wisc.edu/evaluations/).
    - **Course list**: the CS department maintains a course list (https://www.cs.wisc.edu/courses/list).
    - **(Post-2013) course teaching history**: the above list contains links to individual course pages, each of which contains a list of when the course was offered and by whom (e.g. http://www.cs.wisc.edu/courses/367).
  - Text documents:
    - **Rate My Professors:** many sites are freely available which aggregate student reviews of particular courses and professors. One of the most popular of these is Rate My Professors (https://www.ratemyprofessors.com/).
- Methodology
  - **Grade distributions**: We downloaded the PDF file for Spring and Fall semesters from the 1996-1997 school year all the way up to the present. Structured information was extracted from these files using Tabula, a free tool for extracting tables from PDF's.
  - **Rate My Professors:** We used an open source NodeJS package (`rmp-api`) to scrape all reviews for all CS professors. In total, we scraped 576 reviews, each of which contains unstructured text in the form of course comments, at least a few sentences each.
  - **(Pre-2013) course teaching history**: We downloaded the full table, consisting of approximately 1,000 (Professor, Course) tuples.
  - **Course list**: We downloaded the full table, which includes over 100 courses.

    - **(Post-2013) course teaching history**: We will extract this data for the rest of the historical (Professor, Course) tuple data.
- Plan for information extraction phase
    - We plan to extract the reviewer's sentiment to determine if the reviewer was pleased with their experience in the course. Notice this is different than the discrete rating he or she leaves as to the quality of the professor.
- Tools used
    - Tabula/Tabula-java
        - http://tabula.technology/ and https://github.com/tabulapdf/tabula-java
        - Allows table extraction from PDF
    - Rate My Professors NodeJS API: https://www.npmjs.com/package/rmp-api
        - Allows easy querying (scraping) of Rate My Professors site