Semester.ly: Project Proposal **Spring 2017**

Query Search Engine for Semester.ly

Lecturer: *Scribes: Alex Ahn*

2.1 Courses Search Engine

This document is a proposal and a description of a function implementation for Semester.ly regarding with Courses Search Engine.

The goal of this document is to explain the following:

1. The objective/motivation of the Courses Search Engine based on user-typed queries.
2. Description of a methodology for document retrieval in a relevance ranking sense
3. Explanation of an algorithm and implementation (code)

2.2 Objective

Objective of the project is simple: to provide a functionality that serves users to find the most relevant courses based on a query typed in a search bar (using keyboard). I would like to expand the range of user-typed-query to fetch beyond the course titles, including course description, department, area, level and time using appropriate information retrieval methods.

Current system of Semester.ly has a search engine that enlists courses that match with words typed in the search bar. As far as I understand, the system currently returns the list of courses which the words in query that only match with the course titles. I believe that if we can extend the range of query from course titles to more generic and specific queries, the user experience may excel at more dynamic and useful course selection process.

For example, a student may wish to learn a library called "OpenCV". Current system does not return a course that explores "OpenCV". The goal of the Courses Search Engine is to enable users to acquire more interactive query-based course searching system.

2.3 Methodology: Document Information Retrieval

In order to implement Courses Search Engine (CSE), we make use of current document retrieval method based on document-vector modeling.

In short, document-vector modeling is a way to encode each document (pieces of information for each course) from corpus (all course information) to perform tasks such as clustering, relevance ranking, and classification. Our aim is to make use of relevance ranking.

Relevance ranking is a way to generate documents in a sorted order of relevance or similarity scores. By computing similarity scores in various ways (i.e. cosine similarity, jaccard similarity, overlap similarity, dice similarity), one can retrieve most relevant courses based on user input query (or even another course).

For example, an interactive query search may look as below (examplified using terminal).

```
Type in your query:computer vision using opencv and python advanced

   ************************************************************
        Documents Most Similar To Interactive Query number 0
   ************************************************************
   Similarity   Doc#  Author      Title
   ==========   ===  ========     ============================
   0.44269158   AS.050.814.  Research Seminar in Computer Visio
   0.36339759   EN.600.657.  Advanced Topics for Computer Graph
   0.31984176   EN.600.461.  Computer Vision.  3.00 Credits.
   0.28247201   EN.600.683.  Vision as Bayesian Inference.  3.0
   0.26160350   EN.600.662.  Topics in Illumination and Reflect
   0.22084968   EN.600.624.  Advanced Topics in Data-Intensive
   0.21314169   EN.600.642.  Advanced Topics in Cryptography.
   0.20477709   EN.600.661.  Computer Vision.  3.00 Credits.
   0.20339910   EN.600.643.  Advanced Topics in Computer Securi
   0.17538866   EN.600.591.  Computer Science Workshop I.  1.00
   0.16957908   EN.600.485.  Probabilistic Models of the Visual
   0.16464851   EN.600.471.  Theory of Computation.  3.00 Credi
   0.15812178   EN.600.646.  Computer Integrated Surgery II.  3
   0.15396174   EN.600.775.  Selected Topics in Machine Learnin
   0.15211572   EN.600.323.  Data-Intensive Computing.  3.00 Cr
   0.13901609   EN.650.624.  Advanced Network Security.  3.00 C
   0.13494486   EN.600.668.  Advanced Topics in Software Securi
   0.13310435   EN.500.745.  Seminar in Computational Sensing a
   0.13287487   EN.600.745.  Seminar in Computational Sensing a
   0.13089716   EN.600.104.  Computer Ethics.  2.00 Credits.
   0.12988993   EN.600.357.  Computer Graphics.  3.00 Credits.
```

2.4 Algorithm and Implementation

For document-vector based information retrieval modeling, we make use of several techniques briefly explained below to excel precision and overcome limitations.

1. Stemming
   : a way to generalize words by stemming (a technique that is widely used for generalization. I.e. PorterStemming)

2. Stopwords
   : a method to ignore non-meaningful words such as 'is, was, the, a, etc"

3. TF-IDF (term frequency-inverse document frequency)
   : a method to generalize term-frequency based on normalizing factor that bases on number of doucments that contains a term

4. Similarity Scores for query/documents
   : implement mutiple similarity score calculations and choose the best one (cosine, dice, jaccard, overlap, etc)

5. N-gram
   : a way to incorporate word-sequences as features (i.e. 'computer vision' is a one bi-gram used as a feature)

6. Synonyms class
   : a way to regard synonyms as equivalent-words (i.e. parallel == concurrent)

7. Term-weighting
   : a way to weight terms differently by the type of source (i.e. words from 'Title' gets 4 weight units, 'Area' for 2, 'Description' for 1 and so on)

2.5 Examples from current implementation on CS courses website

```
Type in your query:mobile application development android frontend user interface and experience
   ************************************************************
        Documents Most Similar To Interactive Query number 0
   ************************************************************
   Similarity   Doc#  Author      Title
   ==========   ===  ========     ============================
   0.57795837   EN.600.250.  User Interfaces and Mobile Applica
   0.15130705   EN.600.629.  Wireless Networks.  3.00 Credits.
   0.08139565   EN.600.105.  M & Ms: Freshman Experience.  1.00
   0.07942902   EN.600.355.  Video Game Design Project.  3.00 C
   0.07720303   EN.600.638.  Computational Genomics: Data Analy
   0.07672403   EN.600.485.  Probabilistic Models of the Visual
   0.07298485   EN.600.357.  Computer Graphics.  3.00 Credits.
   0.06766051   EN.650.624.  Advanced Network Security.  3.00 C
   0.06739036   EN.600.438.  Computational Genomics: Data Analy
   0.06470664   EN.600.108.  Introduction to Programming Lab.

Type in your query:entrepreneurship design new project for semester making app
   ************************************************************
        Documents Most Similar To Interactive Query number 0
   ************************************************************
   Similarity   Doc#  Author      Title
   ==========   ===  ========     ============================
   0.22219152   EN.600.355.  Video Game Design Project.  3.00 C
   0.16711996   EN.600.255.  Introduction to Video Game Design.
   0.16307776   EN.600.256.  Introduction to Video Game Design
   0.14955493   EN.600.411.  Computer Science Innovation & Entr
   0.11149195   EN.600.446.  Computer Integrated Surgery II.  3
   0.11105571   EN.600.519.  Senior Honors Thesis.  3.00 Credit
   0.10349961   EN.600.443.  Security & Privacy in Computing.
   0.09859848   EN.600.321.  Object Oriented Software Engineeri
   0.08760342   EN.600.591.  Computer Science Workshop I.  1.00
```

```
   0.08577464   EN.600.615.  Big Data, Small Languages, Scalabl
   0.08450382   EN.580.694.  Statistical Connectomics.  3.00 Cr


Type in your query:genomic data health biology and computer vision machine learning data mining
   ************************************************************
        Documents Most Similar To Interactive Query number 0
   ************************************************************
   Similarity   Doc#  Author      Title
   ==========   ===  ========     ============================
   0.46410761   EN.600.441.  Machine Learning for Genomic Data
   0.45226459   EN.600.641.  Advanced Topics in Genomic Data An
   0.37586795   EN.600.775.  Selected Topics in Machine Learnin
   0.36913014   EN.600.438.  Computational Genomics: Data Analy
   0.32798125   EN.600.479.  Representation Learning.  3.00 Cre
   0.31121121   AS.171.205.  Introduction to Practical Data Sci
   0.29647428   EN.600.475.  Machine Learning.  3.00 Credits.
   0.28143110   EN.600.624.  Advanced Topics in Data-Intensive
   0.26534834   EN.600.340.  Introduction to Genomic Research.
   0.25738778   EN.600.675.  Statistical Machine Learning.  3.0
   0.25376389   EN.600.692.  Unsupervised Learning: From Big Da
```

2.6 Final words

I really believe that Semester.ly has created a positive impact on student bodies for facilitating a course scheduling process that can become overwhelming. I am an active user of Semester.ly myself and I remember being very impressed with all of its functionalities and your dedication as developers. From the course that I am currently taking, I realized that this functionality can definitely be something helpful for users of Semester.ly, a function that can save a lot of time from digging in for potential courses. I also regard this opportunity as a chance for me to contribute back to the community which I am very much grateful for, and share the excitement for learning something useful as a function of software. If you guys are interested in having me continue to work on this project suited for Semester.ly, let me know!

Thank you for listening.

Best, Alex