IS559A: Network Analysis
Research Project Proposal

Team:2
Team Members:
Pratik Shrivastava: pratiks2@illinois.edu
Vaishnavi Padala: vpadala2@illinois.edu

Research questions:

- Are the course descriptions reflective of the similarity between various courses at iSchool, and the extent of similarity in a way that students even without prior experience can gain a mental picture of the courses' landscape?
- In the present text descriptions, will students be able to identify clusters of similar or complementary courses?

In this project, we are trying to create a semantic network of courses offered by the iSchool using the course descriptions of each course. Using courses as nodes and the similarity between them as edges, we can identify possible ways in which they are similar or augmenting to each other.

Such an analysis will provide evidence and insights into the gaps in helping students understand the course content and expectations. This may not be visible to an experienced person or the team responsible to draft course descriptions as they may not be familiar with all of the courses and the students' information seeking behavior. This can help in redesigning the course catalog accordingly. A better understanding of how courses relate to each other can help in identifying connections and transferability across interdisciplinary domains and aid such cross pollination at iSchool.

The aim is to quantitatively express the intuitive relationships and connections noticed among course descriptions. Such techniques and measures are a part of:

- An intersection of text mining, data modeling(ontologies) and network analysis, i.e. Semantic networks and Relational content analysis.
- "A **semantic network**, or frame network, is a network that represents semantic relations between concepts. This is often used as a form of knowledge representation. It is a directed or undirected graph consisting of vertices, which represent concepts, and edges, which represent semantic relations between concepts."
- Thematic analysis- Human coding of content.

These will help in identifying the themes that are conflicting, cooperating or augmenting each other among the courses. For example, there are some common themes between data mining and text mining, like clustering. However, they differ in the techniques, terminology, and tools used. Having such understanding of metadata and semantics of the course content can help students better identify opportunities in the job market that align with coursework as the job descriptions are written using terms that can fall in any of the categories of tools, skills, purpose, techniques etc. Hence depicting the ways that courses are similar or different can help in transferring skills and pitching across disciplines or coursework.

Such a concept is described in a similar topic in a paper by professor Jana Deisner- "Network Text Analysis (NTA) is one method for encoding the relationships between words in a text and constructing a network of the linked words (Popping, 2000). The technique is based on the assumption that language and knowledge can be modeled as networks of words and the relations between them (Sowa, 1984). Several NTA methods exist (for an overview see Popping, 2000; Popping & Roberts, 1997, for discussion of empiric studies see Monge & Contractor, 2003), such as Centering Resonance Analysis (Corman et al., 2002), Functional Depiction (Popping & Roberts, 1997), Knowledge Graphing (Bakker, 1987; James, 1992; Popping, 2003), Map Analysis (Carley, 1988; Carley & Palmquist, 1992), Network Evaluation (Kleinnijenhuis, Ridder & Rietberg, 1996), and Word Network Analysis (Danowksi, 1982). Since the terror attacks on September 11, 2001 in the USA, research also focuses on visualizing covert networks extracted from texts (Krebs, 2001; Batagelj, Mrvar & Zaveršnik, 2002; Johnson & Krempel, 2004)".

The following are the concepts used for analyzing the relationships:

- Words co-occurrence
- Nodes similarity /alignment score

IS559A: Network Analysis
Research Project Proposal

- Centrality, Betweenness and Eigen-vector measures
- Topology of the network

There has been significant research conducted in the area of citation analysis, media content analysis and digital humanities at the intersection of networks and text mining which are related to the project. While there is a high incidence of social network and text analysis applications, there is relatively less applications and research about structure of data, particularly ontologies and semantics. Certain works also dealt with analysis of abstracts/ research papers to predict the concepts and themes of the papers and compare them with others to build recommendation engines or research forums like the academia.edu

The course name, id and description are the key fields for finding the similarities between the different courses using the course description. The Moodle course page provides us the reliable resource and accurate resource for collecting the data. It has the course description with the required fields such as course name, course Id which are the keys for generating our network graph. The course explorer site also provides the same data but differs in the format. Some of the courses are bundled and listed under a single course for example "IS 490", whereas on Moodle these are listed as different courses. The data will be collected using the libraries of python into a csv file or can be requested from the iSchool department in csv format for completion of the project.

The dataset contains the text data and will be in the below format for analysis.

Course ID	Course Name	Course Description
SP18IS390RGI	Race, Gender & Info Technology	Open to sophomores, juniors and seniors. Description: In this course we will critically examine the ways in which information and communication technologies (ICTs) are shaped by -- and help to shape -- social relations of race and gender; and we will extend our review to other categories of identity and exclusion as well, such as age, ability, geography and ethnicity. We will also explore the various benefits and burdens of the information society and how these are socially distributed, and conduct case-studies of policies, practices, and programs designed to enhance opportunities and/or mitigate disadvantages through the creative or disruptive use of ICTs.

For the different courses, the course description will be used for creating the corpus of words for our collection. The corpus will be created post data preprocessing (stop words and tokenizing process) steps. The tf-idf calculation will be performed for the words present in the corpus, which will be used for getting the similarity scores. The Euclidean or Cosine similarity measures will be used for getting the similarity scores between the course description (which will be replaced by course name) of various courses. A matrix of course name and similarity measure will be created which will be used for generating the network graphs. The nodes in the network will be the courses and the edges will be similarity measures. If we use the Cosine similarity, we can create a sigmoid function in which the course name having similarity score of 1 will be connected whereas the courses with score 0 will not.

The nodes in our data will be the course names, and the edges will be the similarity measure. From the website, we found 146 distinct courses which gives us the number of the nodes. The no of edges in the default case would be ${}^{146}C_2 = 146 * 145 = 21{,}170$. But this can be reduced if we use a transformation function on our similarity measures.

The network described here can be expanded to a heterogeneous network that includes Professors' research keywords and also job titles and descriptions

Citations:

Tom Brughmans; Networks of networks: a citation network analysis of the adoption, use, and adaptation of formal network techniques in archaeology, *Literary and Linguistic Computing*, Volume 28, Issue 4, 1 December 2013, Pages 538–562, https://doi.org/10.1093/llc/fqt048

Hoser B., Hotho A., Jäschke R., Schmitz C., Stumme G. (2006) Semantic Network Analysis of Ontologies. In: Sure Y., Domingue J. (eds) The Semantic Web: Research and Applications. ESWC 2006. Lecture Notes in Computer Science, vol 4011. Springer, Berlin, Heidelberg

IS559A: Network Analysis
Research Project Proposal

Online links-

https://www.researchgate.net/publication/254825638_Semantic_Network_Analysis_Techniques_for_Extracting_Representing_and_Querying_Media_Content
https://authors.library.caltech.edu/35731/9/nrn3354-s8.pdf
http://ieeexplore.ieee.org/abstract/document/6921602/?anchor=keywords
https://www.researchgate.net/publication/254825638_Semantic_Network_Analysis_Techniques_for_Extracting_Representing_and_Querying_Media_Content
https://academic.oup.com/dsh/article/28/4/538/1079224
http://www.casos.cs.cmu.edu/publications/protected/2000-2004/2003-2004/diesner_2004_usingnetwork.pdf

Dataset Links:

Moodle: https://courses.ischool.illinois.edu/course/index.php?categoryid=50

Course Explorer:

https://courses.illinois.edu/search?year=2018&term=spring&keyword=&keywordType=qs&instructor=&collegeCode=LP&subjectCode=&creditHour=&degreeAtt=&courseLevel=&genedCode1=&genedCode2=&genedCode3=&genedType=all&partOfTerm=&_online=on&_open=on&_evenings=on