DATA 620 Web Analytics

Derek G Nokes

August 27, 2017

Course Summary:

Organizations, both commercial and community, can benefit from deep analysis of their website interactions and mobile data. Social networks have also become a source of information for companies; search engines are an important referral mechanism. Popular social networks and other online communities provide rich sources of user information and (inter-) actions through their application programming interfaces. This data can help to identify a number of individual user preferences and behaviors, as well as fundamental relationships within the community. Search engines use algorithms to rank sites. Students will learn how to analyze social network data for types of networks, the fundamental calculations used in social networks (e.g., centrality, cohesion, affiliations, and clustering coefficient) as well as network structures and roles. Beyond social network data, students will learn about important concepts of analyzing website traffic such as click streams, referrals, keywords, page views, and drop rates. The course will touch on the fundamentals of search algorithms and search engine optimization. To provide a basic context for understanding these online user and community behaviors, students will learn about relevant social science theories such as homophily, social capital, trust, and motivations as well as business and social use contexts. In addition, this course will address ethical and privacy issues as they relate to information on the Internet and social responsibility.

Course Learning Outcomes:

At the end of this course, students will be able to:

–Analyze text data, including natural language processing and text representation, word association, topic mining, opinion mining and sentiment analysis, and text-based prediction.

–Perform network analysis, including creating graphs, calculating statistics on nodes, and graph visualization.

–Work with various social network APIs, including Twitter, Facebook, and Linked In.

Students will be required to:

–Apply what they learn about network analysis and text mining in a series of increasing complex projects and associated presentations.

How is this course relevant for data analytics professionals?

Text mining is about working with unstructured data. Network analysis focuses more on relationships than entities. These are two of the fastest growing sub-fields of data science, and are increasingly important for success in the workplace.

Assignments and Grading:

Assignments (8 x 25): 20%

Projects (4 x 100): 40%

Final Project (1 x 200): 20%

Final Project Presentation (1 x 50): 5%

Discussion Participation (15 x 10): 15%

TOTAL: 100%

Week 1 - Set up Development Environment

Week 2 - Network Analysis and Text Mining Use Cases

Natural Language Processing with Python, Chapters 1 and 2

Social Network Analysis for Startups, Chapter 1

Setting Up GitHub Repository

GraphLab Create

two key data structures: SFrames and SGraph

Week 3 - Network Analysis: Graph Theory, Definitions

Social Network Analysis for Startups, Chapter 2

Set Up Neo4j

download neo4j

pip install neo4j-driver

Practice Using an Adjacency Matrix

Calculating Graph Diameter

Assignment

1) Load a graph database of your choosing from a text file or other source. If you take a large network dataset from the web (such as fromhttps://snap.stanford.edu/data/), please feel free at this point to load just a small subset of the nodes and edges.
2) Create basic analysis on the graph, including the graph's diameter, and at least one other metric of your choosing. You may either code the functions by hand (to build your intuition and insight), or use functions in an existing package.
3) Use a visualization tool of your choice (Neo4j, Gephi, etc.) to display information.
4) Please record a short video (~ 5 minutes), and submit a link to the video in advance of our meet-up.

Week 4 - Network Analysis: Centrality Measures

Social Network Analysis for Startups, Chapter 3

Assignment [due date: end of day on Monday September 26th]

Centrality measures can be used to predict (positive or negative) outcomes for a node.

Your task in this week's assignment is to identify an interesting set of network data that is available on the web (either through web scraping or web APIs) that could be used for analyzing and comparing centrality measures across nodes. As an additional constraint, there should be at least one categorical variable available for each node (such as "Male" or "Female"; "Republican", "Democrat," or "Undecided", etc.)

In addition to identifying your data source, you should create a high level plan that describes how you would load the data for analysis, and describe a hypothetical outcome that could be predicted from comparing degree centrality across categorical groups.

For this week's assignment, you are not required to actually load or analyze the data. Please see also Project 1 below.

You may work in a small group on the assignment. You should post your document to GitHub

Project 1

October 3rd October 11th

Identify and load a network dataset that has some categorical information available for each node.

For each of the nodes in the dataset, calculate degree centrality and eigenvector centrality.

Compare your centrality measures across your categorical groups.

Project should be delivered in an IPython Notebook, and posted in GitHub

Week 5 - Network Analysis: Clustering

Social Network Analysis for Startups, Chapter 4: "Cliques, Clusters, and Components."

Project 1 is due end of day October 3rd

Week 6 - Network Analysis: 2-Mode Networks

Social Network Analysis for Startups, Chapter 5: "2-Mode Networks" and Chapter 6: "Going Viral! Information Diffusion."

Week 7 - Text Mining: Natural Language Processing

Natural Language Processing with Python, Chapters 3 and 4

Week 8 - Text Mining: Word Association

Natural Language Processing with Python, Chapters 5 and 6

Week 9 - Text Mining: Topic Mining 1

Natural Language Processing with Python, Chapters 7 and 8

Week 10 - Text Mining: Topic Mining 2

Natural Language Processing with Python, Chapter 9

Week 11 - Network Analysis: Sentiment Analysis

Natural Language Processing with Python, Chapters 10 and 11

Week 12 - Text Mining: Text-Based Prediction

Natural Language Processing with Python, Chapter 6

Week 13 - Nothing

Week 14 - Network Analysis and Text Mining: Longitudinal Analysis

Network Science, Albert-László Barabási, "Chapter 10: Spreading Phenomena"