Course number and name: **CS 02625 - Data Quality and Web/Text Mining**

Credits and contact hours: 3 credits / 3 contact hours

Instructor's or course coordinator's name: Anthony Breitzman

Instructional materials: None required. If you feel you want a book, here are a couple of good ones. Larose – Discovering Knowledge in Data; Han, Kamber, Pei – Data Mining Concepts and Techniques

Specific course information

Catalog description: This course studies data quality problems and solutions in the context of text and web mining, which is the exploration of vast amounts of digitized text for use in knowledge discovery or more particularly drug discovery in the biomedical field.

About this course: This hands-on course will teach students how to:

- extract digitized text from publicly available databases, web sites, reddit feeds, and other sources; parse and standardize the text, and load it into a large database to be mined.
- understand the problems associated with acquiring data from multiple disparate sources, and learn methods to clean, standardize, and normalize text so that it can be exploited by text mining tools.
- know key algorithms used in natural language processing, information retrieval, database building, and general data mining, and learn how to exploit those algorithms for text mining.
- know Python and know how to exploit powerful open source tools including NLTK (Natural Language Tool Kit) and SQLite. (SQLite is an open source database system accessible via most high level languages such as Python, C/C++, Java, etc. as well as commercial database front ends such as Oracle, Microsoft Access or SQL server.)

Prerequisites: Acceptance into MS-DS or MS-CS programs. (Knowledge of Linear Algebra and Python will be helpful)

Specific Topics

- Introduction to Text Mining
- R, R-studio, Jupyter, Python and sci-kit learn
- Obtaining Data; Building Web Crawlers (etiquette, and spider traps);

- The Google Search API and Reddit API
- Word Clouds
- Data Quality, Data Cleansing, HTML stripping, XML Parsing, BeautifulSoup
- Identifying Sentence Boundaries, Stemming, Part of Speech Tagging
- Building Dictionaries and Thesauri
- NLTK-Natural Language Toolkit
- Sentiment Analysis
- Word embedding; Word2Vec
- Part of speech taggers, regular expressions, chinking, chunking, synonym substitution
- Most important sentence identification, automatic abstracting, automatic keyword finding
- Automatic summarizing
- Vector Space Models; TF/IDF; Similarity Matrices; Clustering
- K-means and hierarchical Clustering; Search Engines; Retrieval; Precision/Recall; F1-Scores
- Topic Modeling/Text Classifiers, Bayes Classifiers
- Named Entity Recognition