Rationale

With the explosion of data coming in a form of unstructured text (e.g., news articles, web pages, Twitter messages, Amazon product reviews, blog posts, etc.) text mining has become an important area of computational research. Text is currently analyzed for learning opinions and shifts in opinions, trends, ideas, connections among people, etc.

Description

The course encompasses ideas from both computer science and computational linguistics. The course covers the study of different representations of textual data as well as algorithms used to glean new/interesting information from text data. This course provides a detailed explanation of how to use linguistically–motivated features for a variety of data mining tasks applied to text, and how to use statistical patterns within text applications.

This course satisfies the "Advanced Natural Language Processing" requirement of the CUNY Graduate Center Computational Linguistics MA/PhD Certificate Program. Linguistics students must have successfully completed Methods in Computational Linguistics I and II. The students are supposed to have programming experience with Python. To successfully pass the class, the students will complete several programming assignments and a class research project.

Topic list (topics may include but are not limited to)

1. Word statistics: regular expressions, word frequencies, bag-of-words, unigrams, n-grams
2. Text representation: tf, tf*idf, stop words, word vectors, application to Information Retrieval
3. Corpus analysis
4. POS tagging: sequential labeling, HMM, classification
5. Document categorization: Bayesian classification, SVM (kernels)
6. Text categorization / classification text mining applications: authorship identification, text polarity, etc.
7. Language modelling, topic modelling, recommender systems
8. Information extraction: IE as text segmentation, named entity recognition
9. Semantics, word sense disambiguation
10. Sentiment analysis and psycholinguistics
11. Text analysis and crowdsourcing
12. Clustering and outlier detection

Learning goals

Upon successful completion of this course, a student can expect to be able:

- Implement a number of text representation techniques
- Perform document and sentiment classification and implement simple classification algorithms
- Correctly evaluate natural language processing applications
- Write systems that integrate disparate information
- Understand the basics of crowdsourcing
- Use current text mining software with practical, real-world data sets in a way that aids decision making.

Assessment:

The class does not have either midterm or final exam. The assessment will be based on the completed homework assignments, final project which students present in class, class participation, research papers discussion.