NATURAL LANGUAGE PROCESSING
STA 9792
Course Description, Requirements & Syllabus

COURSE DESCRIPTION

This course provides a survey of the challenges, concepts and methodologies employed in Natural Language Processing (NLP). The subject brings together the modeling of the underlying structure of human language with the flexibility and power of neural networks and other algorithmic approaches. The course covers modeling the parts of speech, disambiguation, text similarity, maximum entropy methods, neural networks, and computational semantics.

COURSE LEARNING OBJECTIVES

- Demonstrate a basic understanding of the tools of semantics and how those are used to model the structure of language;
- Demonstrate the ability to understand and employ the algorithms currently being used to automate the handling of natural language;
- Show familiarity with the mathematical structures required to capture the subtlety of natural language;
- Prepare an individualized project demonstrating the ability to apply the concepts and techniques of NLP too an original topic;
- Adopt codes of ethical use of statistical methods in the presentation and analysis of natural language processing, including forthright appraisal of what can and cannot be achieved.

MS STATISTICS LEARNING GOALS

General Statistical Competence

Students will be able to apply appropriate probability models and statistical techniques when analyzing problems from business and other fields.

Statistical Practice

Students will become familiar with the standard tools of statistical practice for multiple regression, along with the tools of a subset of specialized statistical areas such as multivariate analysis, applied sampling, time series analysis, experimental design, data mining, categorical analysis, and/or stochastic processes.

Technology Competency

Students will learn to use one or more of the benchmark statistical software platforms, such SAS or R.

COURSE MATERIALS

Required texts:

Jurafsky, Daniel and James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Second Edition. Prentice Hall, 2008

Draft: https://nlp.stanford.edu/~manning/xyzzy/JurafskyMartinEd2book.pdf

Reference texts:

Manning, Christopher D and Hinrich Schütze. Foundations of Statistical Natural Language Processing. MIT Press, 1999

Bird, Stephen, Natural Language Processing with Python. O'Reilly, 2009

Koehn, Philipp, Statistical Machine Translation. Cambridge, 2010

Bengio, Yoshua, *Learning Deep Architectures for AI*. Technical Report, 2009

Jelinek, Frederick, Statistical Methods for Speech Recognition. MIT Press, 1998

Allen, James, Natural Language Understanding. Benjamin/Cummings, 2ed, 1995

Software: R, Python

Course Requirements:

Weekly assignments and an individualized final project.

Evaluation Criteria:

Assginments: 50%; Project: 50%

See http://www.baruch.cuny.edu/advisement/grades-and-gpa.html for the way grades will be assigned in general at Baruch.

Details of Course Requirements:

Students will do assignments involving problems relevant to the course materials. All analyses will be done using the stated software, but will presented in Word documents . All assignments and their solutions

can be found on the course Blackboard website. Students are expected to complete the assignments and the final project, demonstrating as they go that they have achieved a good understanding of the material.

Academic Integrity:

Academic dishonesty is unacceptable and will not be tolerated. Cheating, forgery, plagiarism and collusion in dishonest acts undermine the college's educational mission and the students' personal and intellectual growth. Baruch students are expected to bear individual responsibility for their work and to uphold the ideal of academic integrity. Any student who attempts to compromise or devalue the academic process will be sanctioned. Please see the Baruch College Website for Further Information: http://www.baruch.cuny.edu/academic/academic_honesty.html

Counseling and Student Health:

Students may occasionally have personal issues that arise in the course of pursuing higher education that may interfere with academic performance. If you are facing problems affecting your coursework you are encouraged to seek confidential assistance at the Baruch College Counseling Center 646–312-2158 or contact the Office of Graduate Programs 646-312-1300.

Students with Disabilities:

To qualify for special accommodation you must first register with the Baruch College Disability Services office 646-312-4590.

TOPIC 1: Intro and Regular Expressions
(1.0 week)

Regular expressions (ch 2.1)
Finite state automata (ch 2.2)

TOPIC 2: N-grams
(1.0 week)

Word counting (ch 4.1-4.2)
Training and performance (ch 4.3-4.4)
Smoothing and estimation (ch 4.5-4.8)

TOPIC 3: POS Tagging, Sequence Classifiers
(1.5 weeks)

POS tagging (ch 5.1-5.7)
Hidden Markov models (ch 6.1-6.5)
Maximum entropy (ch 6.6-6.9)

TOPIC 4: Context-free Grammars, Parsing
(1.5 week)

Formal grammars (ch 12.1-12.7)
Syntactic parsing (ch 13.1-13.4)
Probabilistic CFGs (ch 14.1-14.7)

TOPIC 5: Lexical Semantics
(1 week)

Word senses (ch 19.1-19.3)
Event participants (ch 19.4-19.5)

TOPIC 6: **Question Answering, Summarization**
(1 week)

Information retrieval and QA (ch 23.1-23.2)
Summarization (ch 23.3-23.8)

MS Learning Goals	**Significant Part of Course**	**Moderate Part of Course**	**Minimal Part of Course**	**Not Part of Course**
Oral Communication	☐	☒	☐	☐
Written Communication	☒	☐	☐	☐
Technology Literacy	☒	☐	☐	☐
Ethical Awareness	☒	☐	☐	☐
Global Awareness	☐	☐	☒	☐
Quantitative Analysis	☒	☐	☐	☐
Teamwork and Leadership	☐	☐	☒	☐
Knowledge Integration	☐	☒	☐	☐

Assignments	***Course Learning Goals***	***MS Statistics Learning Goals***
Homework Weekly problems using course materials and software to solve big data problems.	➢ Mastery of fundamental l statistical concepts and computational methods ➢ Use appropriate software ➢ Develop critical thinking skills about applications to real-world problems	Statistical Competence Statistical Practice Technology Competence
Lectures: In class time will focus on student-initiated discussion points from the reading assignment, be that a text or lecture notes prepared by the teacher. Lecture time will also be spent developing additional themes that parallel and enhance understanding of the written material.	➢ Critical thinking about business and social problems in quantitative terms ➢ Adopting codes of ethical use of statistical methods in the presentation and analysis of data	Statistical Competence Statistical Practice Technology Competence
Project: Students will develop their own NLP project in which methods developed in class will be employed on a practical level. Each student works up an individual topic; demonstrate skill in applying the full range of methods developed in the course to that topic and data; prepares the needed programs; and uses error measurements to assess the performance quality; and is alert to the challenges of broader issues such as bias-variance tradeoffs and the curse of dimensionality.	➢ Applying appropriate statistical tools, techniques, and procedures for regression analysis ➢ Mastery of fundamental descriptive and inferential statistical concepts and procedures proficiently ➢ Use appropriate software proficiently ➢ Achieving excellence in statistical graphics.	Statistical Competence Statistical Practice Technology Competence