





Computer Science (COMP) 659

Statistical Language Processing for Text Analytics (Revision 1)

Status: Replaced with new revision, see the [course listing](#)  for the current revision 

Delivery mode: [Grouped study](#) 


Credits: 3

Area of study: Information Systems

Prerequisites: **COMP 501** (or an equivalent high-level programming language course) and the essentials of undergraduate-level probability and/or statistics, or course coordinator approval.

Precluded: None

Faculty: [Faculty of Science and Technology](#) 

Notes: This is a graduate level course and students need to apply and be approved to one of the graduate programs or as a non-program **School of Computing and Information Systems**  graduate student in order to take this course. Minimum admission requirements must be met. Undergraduate students who do not

meet admission requirements will not normally be permitted to take this course.

Instructor:

Dr. Dunwei (Grant) Wen [↗](#)

Overview

There has been an increasing demand for better retrieval, processing, and analysis of textual information in modern society in recent years due to the availability of a huge and ever growing amount of textual data from both inside organizations and the Internet. Well known examples include web search engines (e.g., Google), document and content management systems, email filtering, social media sentiment analysis, automated question answering (e.g., IBM Watson® on *Jeopardy!*), natural language interfaces in games and mobile devices, and big data text analytics for business/competitive intelligence. Natural language processing (NLP), also known as computational linguistics, which aims to process and understand natural languages and text, is the driving force that makes these tasks and systems possible.

To better meet the needs of the big data era, this course focuses on the principles and technologies of statistical machine-learning-based NLP and their application in text analytics, including retrieval, extraction, recognition, and analysis of information from large textual collections.

Outline

This course covers the core topics in statistical NLP and several applications of text analytics as follows:

- Unit 1: Linguistics and Statistics Essentials
- Unit 2: Python for Text Processing
- Unit 3: Language Models for Information Retrieval
- Unit 4: Hidden Markov Models for POS Tagging



- Unit 5: Probabilistic Grammar and Parsing
- Unit 6: Statistical Machine Learning
- Unit 7: Text Classification and Clustering
- Unit 8: Semantic Structures and Parsing
- Unit 9: Named Entity and Relation Extraction
- Unit 10: Web Search and Question Answering
- Unit 11: Topic Modeling, Opinion Mining, and Sentiment Analysis

Learning outcomes

Upon successful completion of this course, you should be able to

- explain fundamental concepts, principles, models and algorithms of natural language processing (NLP), including language models, PoS tagging, syntactic and semantic parsing, named entity and relation extraction, question answering, opinion mining and sentiment analysis.
- discuss the state-of-the-art statistical and machine learning algorithms and techniques and their connection with the implementation of statistical NLP tasks and text analytics.
- apply machine learning and NLP algorithms to real natural language data for language and text processing.
- analyze large text collections by selecting and applying suitable statistical NLP approaches.
- evaluate and improve the performance of a selected statistical learning machine for a specific NLP task.
- design system structures and integrate open source components for statistical NLP and text analytics applications.
- review research articles from well-known NLP, machine learning, and AI journals and conference proceedings regarding NLP and text analytics.
- carry out a research project and write a research proposal, report and paper.

Evaluation

To **receive credit**  for COMP 659, you must achieve a cumulative course grade of **B- (70 percent)**  or better, and must achieve an average grade of at least 60% on the assignments. Your cumulative course grade will be based on the following assessment.

The weighting of the composite grade is as follows:

Activity	Weight
Assignment 1: Concepts, Design, Demonstration, Reading	15%
Assignment 2: Demonstration, Reading, Analysis	15%
Assignment 3: Research Project	40%
Assignment 4: Research Paper	20%
Assignment 5: Discussion Forum Participation	10%
Total	100%

Materials

Digital course materials

Links to the following course materials will be made available in the course:

S. Bird, E. Klein, and E. Loper. 2009. *Natural Language Processing with Python —Analyzing Text with the Natural Language Toolkit*. O'Reilly Media. Available at <http://www.nltk.org/book/>

T., Hastie, R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2ed. Springer. Available at <https://web.stanford.edu/~hastie/ElemStatLearn/>

C.D. Manning, P. Raghavan, and H. Shutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press. Available at <http://nlp.stanford.edu/IR-book/>

- Articles from journals such as *Computational Linguistics*, *Natural Language Engineering*, and *Machine Learning Journal* and conference proceedings (e.g., ACL, ICML, NAACL, EMNLP, HLT, AACL, IJCAI), all available through AU Library Services databases.
- Selected materials from online resources including *Wikipedia* and NLP Tutorials.
- Selected contents from a list of relevant books (all available online):

Reference Materials

D.Barber. 2012. *Bayesian Reasoning and Machine learning*. Cambridge University Press.

D. Jurafsky and J. H. Martin. 2008. *Speech and Language Processing—An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2ed. Pearson Prentice Hall, Upper Saddle River, NJ.

C. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Other Materials

The remaining learning materials are distributed in electronic format on the course site.

Programming Languages and Software Tools



The Python programming language and Python-based open source machine learning and NLP software tools are mainly used in this course. However, students may select either Java or C++ as an alternative language and use its relevant open source machine learning and NLP tools for their assignments and research projects.

Special Note

Students registered in this course will **NOT** be allowed to apply for a course extension due to the nature of the course activities.

Important links

- › [Future Course Offerings](#) 

- > [Important Dates and Deadlines](#) 
- > [MSc IS Contact Information](#) 

Athabasca University reserves the right to amend course outlines occasionally and without notice. Courses offered by other delivery methods may vary from their individualized study counterparts.

Updated January 27, 2025
