Ling 583 Statistical Methods in Text Analysis

Schedule #22040
Spring 2018
Th 4:00am–6:40pm
Room SHW-243 (Computational Linguistics Lab)

According to industry statistics, up to 80% of the information owned by large organizations is in the form of 'unstructured' text documents. During this course, students will collaboratively build applications that extract usable information from a large collection of texts. We will cover techniques for collecting, organizing and annotating textual databases. The course activities will familiarize students with the use of a range of standard statistical methods for analysis of large texts, including Markov models, Bayesian classifiers, maximum entropy models, support vector machines, and neural nets, as applied to tasks such as topic modeling, relation detection, and sentiment analysis. After completion of this course, students will be able to:

- Apply techniques for collecting and preparing text data for computational analysis.
- Describe a variety of text analysis steps, understand their interdependencies, and identify algorithms for each step.
- Choose and apply appropriate text classification and clustering algorithms
- Choose and apply appropriate semantic analysis techniques
- Integrate knowledge and apply skills acquired over the sequence of text analysis courses to solve real world problems.

Prerequisites

LING 571 or LING 572; and STAT 550 or STAT 551A or permission of instructor

Instructor

Rob Malouf
Office: SHW 244
Office hours: Tu 8:00–9:00, Th 1:00–2:00, or by appt
Email/GTalk: `rmalouf@mail.sdsu.edu`
Phone: (619) 594-7111

Requirements

The final grade will be based on four projects over the course of the semester. Taken together, these projects will constitute all the components of a complete end-to-end text analytics application. Students may work in small groups, but each student will be required to submit their own write-up. Graduate students will also be required to write a paper describing the design of a (hypothetical) text mining application. In addition, we will be doing hands on lab exercises in class each week. Submitted lab write ups will make up 10% of the grade for both grads and undergrads.

	Undergrads	Grads
Labs	10%	10%
Project 1	15%	15%
Project 2	15%	15%
Project 3	30%	20%
Project 4	30%	20%
Paper	—	20%

No form of academic dishonesty, including cheating or plagiarism, will be tolerated in the class. Following Executive Order 1006, all instances of academic dishonesty will be reported to the Center for Student Rights and Responsibilities for investigation. For more information about the judicial process, see `http://csrr.sdsu.edu`. For more information about what plagiarism is and how to avoid it, see `http://its.sdsu.edu/tech/plagiarism.html`.

If you are a student with a disability and believe you will need accommodations for this class, it is your responsibility to contact Student Disability Services at (619) 594-6473. To avoid any delay in the receipt of your accommodations, you should contact Student Disability Services as soon as possible. Please note that accommodations are not retroactive, and that accommodations based upon disability cannot be provided until you have presented your instructor with an accommodation letter from Student Disability Services. Your cooperation is appreciated.

Readings

The required textbook for this course is:

- Andreas C. Müller and Sarah Guido. 2017. *Introduction to Machine Learning with Python: A Guide for Data Scientists.* O'Reilly.

It should be available in the bookstore, on Amazon, etc., or in electronic form via the library.

Additional readings will be made available on the class website.

Proposed schedule

- Week 1 **Introduction**
  Background · Uses for text analytics · Using the Computational Linguistics Lab

- Week 2 **Text repositories**
  Linguistic corpora · Pubmed · Open data collections

- Week 3 **Web scraping**
  Spiders · HTML and XPATH · Data cleanup

  Project 1: Gathering text due

- Week 4 **Annotation**
  Planning annotation schemes · Manual annotation · Inter-annotator agreement

- Week 5–6 **Sequence models**
  Sequence annotation · Language models · Hidden Markov Models · Conditional random fields

- Week 7–8 **Classifiers**
  Text classifiers · Naive Bayes · Maximum Entropy models · Support Vector Machines · Sentiment analysis

  Project 2: Annotation due

- Week 9–10 **Topic models**
  Vector spaces · Latent Semantic Analysis · Latent Dirichlet Allocation · Embeddings

- Week 11 **Clustering**
  K-means · Hierarchical clustering · Visualization

  Project 3: Extracting topic clusters

- Week 12 **Dependency parsing**
  Dependency vs. constituents · Projective and non-projective algorithms

- Week 13–14 **Relation detection**
  Entities and relations · Taggers and classifiers for relations · 'Deep' learning and neural nets

- Week 15 **Project presentations**

  Final projects and papers due