CS4984: Computational Linguistics
Virginia Tech, Fall 2014

Description:
With support from a grant from the National Science Foundation, this course will give students the opportunity to engage in active learning about how to work with large collections of text, one aspect of 'big data'. An 11-node Hadoop cluster, along with other tailored computing resources, will aid handling of over 500 million tweets and over 10 terabytes of webpages. Using methods employed in search engines, including linguistic analysis and natural language processing, as well as statistical techniques, students will engage in problem based learning with the semester long challenge of analyzing content collections automatically, extracting key information, and generate easily readable summaries of important events in English. Just-in-time learning will allow development of an understanding of concepts, techniques, and toolkits so students will master the key methods related to computational linguistics.

Instructor:
Professor Edward A. Fox, fox@vt.edu, http://fox.cs.vt.edu, 231-5113

Prerequisites: senior standing in CS, or instructor permission
Text: http://www.nltk.org/book/

Topics:
- Lexical, syntactic, semantic, discourse, and statistical analysis of texts
- Automatic text generation
- Natural Language Toolkit
- Tweet and webpage analysis
- Indexing (stopwords, stemming/lemmatization, morphology, phrases)
- Named entity recognition and extraction
- Ontology building and utilization
- Cluster-based processing with Hadoop, Solr, and other tools

Evaluation:
- 50% team term project (sum of: 3% proposal, 10% midterm presentation, 15% final presentation, 22% project report – released in VTechWorks; with adjustment based on team peer assessment)
- 15% midterm exam
- 35% final exam

Computational Linguistics (CL) Course Plan

Different Aspects of the Common Project

- All students will work with some portion of the 11TB of webpages and the 500M tweets collected with IDEAL.
- Students will work in groups of 5, preferably each group having people covering a mix of skills, e.g., Python experience, exposure to linguistics, …
- Each group will pick a particular class of events, e.g., hurricane, earthquake, political election, …
- Each group will identify relevant parts of the available content, and develop a way to generate summaries for instances of their chosen class of events.

Connection with Ensemble

- From this course will come a collection in Ensemble (computingportal.org).
- This collection will be usable by others who want to teach CL.
- Instructors should be able to easily tailor a new course from the collection of educational resources.

Tools

- Students will learn how to use each of the key commonly employed CL tools.
- They will learn them when they are needed.
- Learning about a tool will be aided by a module, like those used in the DL curriculum project. It will refer to YouTube videos/lectures, tutorials, papers, primers, etc.
- Tools also will include those used for webpage and tweet processing.
- Tools also will include those used in our Hadoop cluster.

Prototypes, Iterative Refinement

- Students will devise a rapid prototype with naïve assumptions in month 1.
- Students will implement a series of ever better versions during the course.
- Each version will be more complex and yield higher quality results.
- Thus, they will rapidly achieve full success, but will see how to improve.

Programming

- Student will use NLTK and program in Python.
- Students will learn high-level languages used with the various tools.