CS-4984 SS: Big Data Text Summarization

Prerequisites: senior standing, CS3114 or CS3654 or CS4624 or permission of instructor

CS-5984 SS: Big Data Text Summarization

Prerequisites: graduate standing, permission of instructor

Lecture, 3 credits. Pedagogical approach: team project / problem based learning. Capstone.

Summary: Natural language processing methods applied to big data text collections selected from billions of tweets and terabytes of webpages will produce useful summaries.

Instructor: Professor Edward A. Fox, fox @ vt.edu, http://fox.cs.vt.edu, 540-231-5113

Topics:

- Cluster-based processing with Hadoop, Solr, and other tools
- Tweet and webpage analysis
- Text pre-processing: stop word removal, stemming/lemmatization
- Text analysis packages: part of speech, named entities, topic analysis
- Natural language processing toolkit
- Text classification, document clustering
- Approaches to summarization, including extractive, abstractive, deep learning

Evaluation:

- 70% team term project (sum of: 35% modules - focused on iterative refinement of term project solutions, 10% final presentation, 25% project report - released in VTechWorks; with adjustment based on team peer assessment)
- 10% midterm exam
- 20% final exam

Different Aspects of the Common Project:

- All students will work with some portion of the webpages and tweets collected in connection with a series of NSF-funded projects running since 2007.
- Students will work in groups of 4-5, preferably each group having people covering a mix of skills, e.g., use of a Hadoop cluster, Python experience, exposure to linguistics.
- Each group will work with a particular class of events / trends, undertake pre-processing and use of text analysis packages, and produce various types of summaries.

Prototypes, Iterative Refinement :

- Students will devise a rapid prototype with naive assumptions, early in the course.
- 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 success, but will see how to improve in stages, achieving useful intermediate goals along the way.

Programming:

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

References:

- Textbook: Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly, 2009. ISBN: 0596516495. Free version at http://www.nltk.org/book . See also http://shop.oreilly.com/product/9780596516499.do and http://www.nltk.org/
- Free book: Jimmy Lin and Chris Dyer. Data-Intensive Text Processing with MapReduce. Morgan & Claypool. 2010. 177 pages. ISBN: 9781608453429. DOI: 10.2200/S00274ED1V01Y201006HLT007. http://dx.doi.org/10.2200/S00274ED1V01Y201006HLT007. Note that all of the M&C books can be freely downloaded from http://www.lib.vt.edu.
- Other references, as appropriate will be used, each discussed in the related curricular modules.