Summarization of Documents and Interaction

11-899
Fall 2014, Tu/Th 1:30-2:50, WEH 5304

Instructor: Dr. Carolyn P. Rosé, cprose@cs.cmu.edu
Office Hours: GHC 5415, By appointment

Units: 12 (PhD/Master’s)

Books:
Most readings will be provided in PDF form on Blackboard. The following is an exception, although you may be able to find it in PDF form on the web:

Noah Iliinsky and Julie Steele (2011). Designing Data Visualizations, O’Reilly.

We will read some excerpts from the following book, which will be provided. The book as a whole is recommended supplementary reading. The book is available on Amazon.com.


Alternatively, this older book is available in e-book form through the CMU library system:


Prerequisites: Assignments will involve programming in R and Python.

Course Description:
As more of modern life is conducted online, the wealth of data that exists expands at a dizzying rate. Furthermore, changes in publication practices at all levels have dramatically increased the magnitude of available recorded knowledge. Soon after the turn of the century, awareness grew that information overload was a problem impacting the world at a grand scale. In response, the fields of machine learning and language technologies have developed technical solutions for getting a handle on data and transforming data into knowledge. The field of summarization has blossomed in its various foci from document summarization to summarization of behavior traces, including summarization of social interaction. Application areas in education, health, finance, and disaster relief have taken these technologies and applied them with the goal of positive human impact. The threefold focus of this course is to increase familiarity with the relevant technology in the areas of text mining, summarization and visualization, to raise awareness of theoretical and methodological insights from the field of Human-Computer Interaction that raise research questions, set desiderata for positive impact, and
guide our evaluation of technical solutions, and to offer a survey of applications in education, health, and business.

A term project will involve selecting a topic in summarization, doing a literature review, and either doing a small scale user study or a prototype implementation using open source tools. Tried and true technologies, beginning with an opensource toolkit we will provide, offer the possibility of building end-to-end working prototypes within the scope of a semester course project.

**Grading Criteria**
- Class participation (10%)
- 3 Homework Assignments (10% each)
- Class presentation (10%)
- Term project (40%)
- Take Home Final exam (10%)

**Week 1: Introduction to Summarization**
- **Tue: Course Intro**
- **Thur: Simple Summarization Overview**

  Nenkova, A. & McKeown, K. (2011). *Automatic Summarization*, NOW (Chapters 1 and 2)
  Assignment 1 assigned (due two weeks later)

**Week 2 Design and Evaluation Process: Summarization and Visualization**
- **Tue: Evaluation of Summarization**


  **Thur: Visualization overview**

  Iliinsky, N. & Steele, J. (2011). *Designing Data Visualizations*, O’Reilly (Chapters 1 and 2)

**Week 3 Visualization Continued**
- **Tue: Design Process**

  Iliinsky, N. & Steele, J. (2011). *Designing Data Visualizations*, O’Reilly (Chapters 3 and 4)

  **Thur: Design Evaluation**

  Iliinsky, N. & Steele, J. (2011). *Designing Data Visualizations*, O’Reilly (Chapters 5 and 6)
Assignment 2 assigned (due two weeks later)

Week 4 Genre Specific Summarization

**Tue: Contrasting Genres: A View from the Summarization Literature**


**Thur: Contrasting Genres: A View from the Linguistics Literature**


Deadline to pick dataset for Term Project
Deadline to sign up for a critique

Week 5/6 Advanced Techniques

**Tue Week 5 Leveraging World Knowledge**

Louis, A. (2014). A Bayesian Method to Incorporate Background Knowledge during Automatic Text Summarization, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*

Ng, J., Chen, Y., Kan, M., Li, Z. (2014). Exploiting Timelines to Enhance Multi-document Summarization, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*

**Thur Week 5 Sentence Compression**


**Tue Week 6 Zooming in on Text Visualization**


**Thur Week 6 Leveraging Insights from Cognitive Science**


Week 7 Summarizing News

**Tue Event Extraction**


**Thur Interactive News Visualization**


Cui, W., Qu, H., Zhou, H., Zhang, W., & Skiena, S. (2012). Watch the Story Unfold with TextWheel: Visualization of Large-Scale News Streams, *ACM Transactions on Intelligent Systems and Technology*, 3(2)

Week 8 Summarizing Personal Histories and Orientations from blogs and discussion forums


Week 9 Summarizing Work Activities

**Tue Meeting Summarization**


On Thur this week there will be a Guest Lecture: Summarization to support work in Wikipedia

Week 10 Summarizing Scientific Literature (Single Documents)

Cleveland, D. & Cleveland, A. (2013). *Introduction to Indexing and Abstracting*, Libraries Unlimited (Chapters 9 and 18)


Week 11 Summarizing Scientific Literature (Document Collections)


Ding, Y., Zhang, G., Chambers, T., Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis, *Journal of the Association for Information Sciences and Technology* 65(9), pp 1820-1833
Week 12 Summarization in the Health Domain


Group Project Proposal Due on Thur
Assignment 3 Assigned (due two weeks later)

Week 13 Summarizing Social Engagement in Online Learning Environments


Week 14 Summarizing the Crowd

Tue: Twitter Summarization


Thur: Thanksgiving Holiday, no class
Week 15

**Tue:** Work day to prepare for in class presentation.

No lecture. Scheduled meetings with project teams to get feedback in preparation for in class presentation.

**Thur:** Presentations
- In class presentations of final projects.
- Final exam handed out, due on Dec 5

Final Project Report due on Dec 15, 2014