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.

Ani Nenkova and Kathleen McKeown (2011). *Automatic Summarization*, Foundations and Trends in Information Retrieval, NOW.

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

Inderjeet Mani (2001). *Automatic Summarization*, John Benjamins Publishing

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

Nenkova, A. & McKeown, K. (2011). *Automatic Summarization*, NOW (Chapter 6)

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

Liu, S., Zhou, M., Pan, S., Qian, W., Cai, W., & Lian, X. (2012). TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis, *ACM Transactions on Intelligent Systems and Technology*, 3(2)

Nenkova, A. & McKeown, K. (2011). *Automatic Summarization*, NOW (Chapter 5)

Thur: Contrasting Genres: A View from the Lingusitics Literature

Cleveland, D. & Cleveland, A. (2013). *Introduction to Indexing and Abstracting*, Libraries Unlimited (Chapter 15)

Biber, D. & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge University Press (excerpts)

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

Kikuchi, Y., Hirao, T., Takamura, H., Okumura, M., & Nagata, M. (2014). Single Document Summarization based on Nested Tree Structure, in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics.*

Berg-Kirkpatrick, T., Gillick, D., and Klein, D. (2011). Jointly learning to extract and compress. In *Proceedings of ACL-HLT*, pp 481–490

Qian, X. & Liu, Y. (2014). Polynomial Time Joint Structural Inference for Sentence Compression, in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics.*

Tue Week 6 Zooming in on Text Visualization

Puretskiy, A., Shutt, G., & Berry, M. (2010). Survey of text visualization techniques, in Berry, M. & Kogan, J. (Eds). *Text Mining: Applications and Theory*, John Wiley & Sons

Chuang, J., Ramage, D., Manning, C., Heer, J. (2012). Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis, *Proceedings of ACM SIG-CHI.*

Thur Week 6 Leveraging Insights from Cognitive Science

Patterson, R., Blaha, L., Grinstein, G., Ligget, K., Kaveney, D., Sheldon, K., Havig, P., Moore, J. (2014). A human cognition framework for information visualization, *Computers and Graphics* 42, pp 42-58.

Week 7 Summarizing News

Tue Event Extraction

Pighin, D., Cornolti, M., Alfonseca, E., Filippova, K. (2014). Modeling Events through Memory-based, Open-IE Patterns for Abstractive Summarization, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*

Thur Interactive News Visualization

Candan, K., di Caro, L., & Sapino, M. (2012). PhC: Multiresolution Visualization and Exploration of Text Corpora with Parallel Hierarchical Coordinates, *ACM Transactions on Intelligent Systems and Technology*, 3(2)

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

Li, J. & Cardie, C. (2014). Timeline generation: tracking individuals on twitter, in *Proceedings of the 23rd international conference on the World Wide Web*.

Ranade, S., Gupta, J., Varma, V., Mamidi, R. (2013). Online debate summarization using topic directed sentiment analysis, *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining.*

Lloret, E., Balahur, A., Gomez, J., Montoyo, A., Palomar, M. (2012). Towards a unified framework for opinion retrieval, mining and summarization, Journal of Intelligent Information Systems, 39(3), pp 711-747.

Week 9 Summarizing Work Activities

Tue Meeting Summarization

Wang, L. & Cardie C. (2013). Domain-Independent Abstract Generation for Focused Meeting Summarization, in *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics*

Mehdad, Y., Carenini, G., Tompa, F., Y Ng, R. (2013). Abstractive Meeting Summarization with Entailment and Fusion, in *Proceedings of the 14th European Workshop on Natural Language Generation.*

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)

Angrosh, M., Cranfield, S., Stanger, N. (2013). Context identification of sentences in research articles: Towards developing intelligent tools for the research community, *Natural Language Engineering* 19(4), pp481-515

Qazvinian, V., Radev, D., Mohammad, S., Dorr, B., Zajic, D., Wihdby, M., & Moon, T. (2013). Generating Extractive Summaries of Scientific Paradigms, *Journal of Artificial Intelligence Research*, Volume 46, pages 165-201

Week 11 Summarizing Scientific Literature (Document Collections)

Nanba, H., Kando, N., & Okumura, M. (2000). Classification of research papers using citation links and citation types: Towards automatic review article generation. *Proceedings of the 11th ASIS SIG/CR Classification Research Workshop*, 117-134.

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

Dunne, C., Shneiderman, B., Gove, R., Klavans, J., Dorr, B. (2012). Rapid Understanding of Scientific Paper Collections: Integrating Statistics, Text Analytics, and Visualization, *Journal of the American Society for Information Science and Technology*, 63(12), pp 2351–2369.

Week 12 Summarization in the Health Domain

Baumel, T., Cohen, R., Elhadad, M. (2014). Query-Chain Focused Summarization, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.*

Bian, J., Topaloglu, U., Yu, F. (2012). Towards large-scale twitter mining for drug-related adverse events, *Proceedings of the 2012 international workshop on Smart health and wellbeing*

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

Week 13 Summarizing Social Engagement in Online Learning Environments

Stephens-Martinez, K., Hearst, M., Fox, A. (2014). Monitoring MOOCs: Which Information Sources Do Instructors Value? *In Proceedings of the first ACM Conference on Learning at Scale*

Huang, J., Dasgupta, A., Ghosh, A., Manning, J., Sanders, M. (2014). Superposter behavior in MOOC forums, *Proceedings of the First ACM Conference on Learning at Scale*

Chaturvedi, S., Goldwasser, D., Daumee III, H. (2014). Predicting Instructor's Intervention in MOOC Forums, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.*

Week 14 Summarizing the Crowd

Tue: Twitter Summarization

Lee, R., Wakamiya, S., Sumiya, K. (2012). Urban area characterization based on crowd behavioral lifelogs over Twitter, *Personal Ubiquitous Computing* 17(4), pp 605-620.

Cheong, M. & Lee, V. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter, *Information Systems Frontiers* 13, pp 45-59.

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