Computational Text Analysis

Federico Nanni
federico@informatik.uni-mannheim.de
SFB 884 - University of Mannheim

Overview

The course aims to offer a broad overview of natural language processing approaches and tools, together with their applications in the social sciences. It will specifically focus on learn how to properly use and evaluate them. No previous knowledge on programming or natural language processing are required, just be curious.

Takeaways. At the end of the course the students will be able to:

a) critically analyse a computational social science paper in all its aspects

b) re-implement the approaches presented in this research-area

c) adopt and adapt NLP approaches for their own research

Evaluation. Written exam (6 ECTS) + code (4 ECTS)

Materials. All slides will be available here. The code used in class is shared here.

[All slides and code from previous years are also available]

1st Week. Overview of the course, intro to Python.

Before coming to the first class try to install Jupyter notebook (http://jupyter.org/install.html). I highly recommend installing Anaconda (which contains Jupyter, among many other things that we will need). If you have any problem, just drop me an email.

Reading list:

- Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." Political analysis 21.3, 2013.
- O'Connor, Brendan, David Bamman, and Noah A. Smith. "Computational text analysis for social science: Model assumptions and complexity.", Public Health 41.42, 2011.

2. Intro to Computational Text Analysis, intro to Python.

The first two weeks are mainly focused on setting up a common ground on topics such as natural language processing, on learning Python syntax and collecting a corpus that we will use in the following classes.

Reading list:

- Barberá, Pablo. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." Political Analysis 23.1, 2014.
- Narayanan, Arvind, and Vitaly Shmatikov. "Robust de-anonymization of large sparse datasets." Security and Privacy, 2008.
- Deluca, L. and Pallitto, R. "The Contemporary Presidency: Digital Resources to Support Quantitative Scholarship in Presidential Studies. Presidential Studies Quarterly", 2018.

3. Text processing (tokenization, lemmatization, POS-Tagging, NER)

Week 3 and 4 are focused on learning the main components of a NLP pipeline, such as a tokenizer, part-of-speech tagger, lemmatizer and named entity recognizer.

Reading list:

- Manning, Christopher D., and Hinrich Schütze. Foundation of Statistical Natural Language Processing, 1999- for these two classes I suggest you to skim through chapters 3, 4 and 10.
- Cross, James P., and Henrik Hermansson. "Legislative amendments and informal politics in the European Union: A text reuse approach." European Union Politics 18.4, 2017.

4. Text processing (tokenization, lemmatization, POS-Tagging, NER)

Week 3 and 4 are focused on learning the main components of a NLP pipeline, such as a tokenizer, part-of-speech tagger, lemmatizer and named entity recognizer.

Reading list:

- Schrodt, Philip A., and David Van Brackle. "Automated coding of political event data." Handbook of computational approaches to counterterrorism. Springer, New York, NY, 2013.
- O'Connor, Brendan, Brandon M. Stewart, and Noah A. Smith. "Learning to extract international relations from political context." Proc. of ACL, 2013.

5. Text processing (Word Embeddings and Entities)

Week 5 and 6 are focused on extracting semantic properties from text, from distributional representations of word to disambiguated entities.

Reading list:

- Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors." Proc. of ACL, 2014.
- Shen, Wei, Jianyong Wang, and Jiawei Han. "Entity linking with a knowledge base: Issues, techniques, and solutions." IEEE Transactions on Knowledge and Data Engineering 27.2, 2015.

6. Text processing (Word Embeddings and Entities)

Week 5 and 6 are focused on extracting semantic properties from text, from distributional representations of word to disambiguated entities.

Reading list:

- Kraft, P., Jain, H., & Rush, A. M. "An Embedding Model for Predicting Roll-Call Votes", Proc. of EMNLP, 2016.
- Glavaš, Goran, Federico Nanni, and Simone Paolo Ponzetto. "Unsupervised Cross-Lingual Scaling of Political Texts.", Proc. of EACL, 2017
- Menini, Stefano, et al. "Topic-based agreement and disagreement in US electoral manifestos.", Proc of EMNLP, 2017.

7. Text Classification and Sentiment Analysis

Week 7 and 8 are focused on supervised machine learning, with the specific applications of text classification algorithms for sentiment analysis and topic detection.

Reading list:

- Allahyari, Mehdi, et al. "A brief survey of text mining: Classification, clustering and extraction techniques." arXiv, 2017.
- Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams Engineering Journal, 2014.

8. Text Classification and Sentiment Analysis

Week 7 and 8 are focused on supervised machine learning, with the specific applications of text classification algorithms for sentiment analysis and topic detection.

Reading list (sentiment analysis):

- Young, Lori, and Stuart Soroka. "Affective news: The automated coding of sentiment in political texts." *Political Communication* 29, no. 2, 2012.
- Murthy, Dhiraj. "Twitter and elections: are tweets, predictive, reactive, or a form of buzz?." Information, Communication & Society 18, no. 7, 2015.
- Soroka, Stuart, Lori Young, and Meital Balmas. "Bad news or mad news? Sentiment scoring of negativity, fear, and anger in news content." The ANNALS of the American Academy of Political and Social Science 659, no. 1, 2015.

Reading list (text classification - topic detection):

- Hillard, Dustin, Stephen Purpura, and John Wilkerson. "Computer-assisted topic classification for mixed-methods social science research." Journal of Information Technology & Politics 4, no. 4, 2008.
- Hopkins, Daniel J., and Gary King. "A method of automated nonparametric content analysis for social science." American Journal of Political Science 54, no. 1, 2010.
- Conover, Michael D., Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. "Predicting the political alignment of twitter users." InProc. of PASSAT, 2011.
- Zirn, Cäcilia, Goran Glavaš, Federico Nanni, Jason Eichorts, and Heiner Stuckenschmidt. "Classifying topics and detecting topic shifts in political manifestos.", 2016.

9. Clustering and Topic Models

Week 9 and 10 are focused on unsupervised machine learning, with the specific applications of clustering algorithms and latent dirichlet allocations for topic detection.

Reading list:

- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3, 2003.
- Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. "Reading tea leaves: How humans interpret topic models." In Advances in neural information processing systems, 2009.
- Brett, Megan R. "Topic modeling: a basic introduction." *Journal of digital humanities* 2.1, 2012.
- Graham, Shawn, Scott Weingart, and Ian Milligan. "Getting started with topic modeling and MALLET". The Editorial Board of the Programming Historian, 2012.

10. Clustering and Topic Models

Week 9 and 10 are focused on unsupervised machine learning, with the specific applications of clustering algorithms and latent dirichlet allocations for topic detection.

Reading list:

- Grimmer, Justin. "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases." Political Analysis 18, no. 1, 2010.
- Yano, Tae, William W. Cohen, and Noah A. Smith. "Predicting response to political blog posts with topic models." Proc. of NAACL, 2009.
- Roberts, Margaret E., et al. "Structural Topic Models for Open-Ended Survey Responses." American Journal of Political Science 58.4, 2014.
- Greene, Derek, and James P. Cross. "Exploring the political agenda of the European parliament using a dynamic topic modeling approach.", 2017.
- Menini, Stefano, et al. "Topic-based agreement and disagreement in US electoral manifestos." Proc. of EMNLP, 2017.

11. Scaling

Week 11 is focused on algorithms for determining the positions of actors regarding a specific policy in a space.

Reading list:

- Lowe, Will, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. "Scaling policy preferences from coded political texts." Legislative studies quarterly 36, 2011.
- Slapin, Jonathan B., and Sven-Oliver Proksch. "A scaling model for estimating time-series party positions from texts." *American Journal of Political Science* 52, 2008.
- Lowe, Will. "Understanding wordscores." Political Analysis 16, no. 4, 2008.
- Glavaš, Goran, Federico Nanni, and Simone Paolo Ponzetto. "Unsupervised cross-lingual scaling of political texts." Proc. of EACL, 2017.

12. Information Retrieval and Collection Building

Week 12 and 13 are focused on information retrieval and generating topic-specific collections from large corpora.

Reading list:

- Schütze, Hinrich, Christopher D. Manning, and Prabhakar Raghavan. Introduction to information retrieval. Vol. 39. Cambridge University Press, 2008. You can skim through chapters 6, 9, 11.
- Ponte, Jay M., and W. Bruce Croft. "A language modeling approach to information retrieval." Proc. of SIGIR, 1998.
- Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. "The PageRank citation ranking: Bringing order to the web". Stanford InfoLab, 1999.
- Liu, Tie-Yan. "Learning to rank for information retrieval." Foundations and Trends in Information Retrieval 3, no. 3, 2009.

13. Information Retrieval and Collection Building

Week 12 and 13 are focused on information retrieval and generating topic-specific collections from large corpora.

Reading list:

- Lepore, Jill. "The cobweb: Can the Internet be Archived?." *The New Yorker,* 2015.
- Gade, Emily, John Wilkerson, and Anne Washington. "The. GOV Internet Archive: A Big Data Resource for Political Science", 2016.
- Ben-David, Anat. "What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain." New Media & Society 18, no. 7, 2016.

14. Final Recap and Overview of the Exam

Last week will wrap up the course and describe the structure of the exam.