Text Categorization and Classification

Prof. Nicole Rae Baerg,
University of Mannheim,
nicole.baerg@uni-mannheim.de
Classroom: A5, 6 C317 & C109 for computer lab
Class Hours: 9:30 - 12:00, 14:30 - 17:00

July 7, 2016 - July 9, 2016

Course Objectives

One important problem with big data is finding meaningful associations in texts all the while automating as much of the process as we can. Text classification consists of assigning textual documents to one or more categories, based on the content of the document. In this module, we will examine three phases of text (or document) classification: Text annotation - how to annotate important features of text and how to use this information for document classification; Training -supervised, semi-supervised, and non-supervised approaches to text and document classification; Prediction (or classification) - classifying your texts and validating performance and accuracy.

We will take a very hands-on approach, starting with assessing theories in terms of theoretical implications and matching those with textual features. We will learn how to annotate texts, use annotated texts in supervised and semi-supervised approaches, and classify (and cluster) political texts. Finally we will also spend some time learning how to communicate and visualize our results.

Students are responsible for coming up with a simple research design problem in teams of 2 to 3 people that we will discuss in the afternoon of the last class. The research design should highlight the use of textual tools for categorization and or classification on your documents of interest as well as ideas about validation and visualization.

Background Readings

To make the best use of our brief time, we will assume all participants are familiar with certain basics of statistics. Students should read in advance the material covered in each unit of the course schedule. Time during the course can thus be spent re-reading important materials, and implementing some of the techniques developed in the class.

For both background and covered readings, I have tried to provide you with "cutting edge" examples in the field. Don't be alarmed if you do not "get" everything. We will draw from these readings throughout the sessions and think about them in the context of what we are doing. In the course schedule below, please note the particularly salient readings for the sessions and come to class having read those papers.

The main text that we will use is:

James, Gareth, et al. *An Introduction to Statistical Learning.* Vol. 112. New York: Springer, 2013.

In addition to this text, I also suggest that you read the following:

- 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): 267-297.
- Wallach, Hanna. Big Data, Machine Learning, and the Social Sciences: Fairness, Accountability, and Transparency NIPS Workshop on Fairness, Accountability, and Transparency in Machine Learning. 2014.

Course Schedule

Introduction to machine learning and document classification. Discussion of computational approaches for coming up with features. Introduction to easy annotation software for user-made annotations and labeling. Introduction to using crowd-sourcing for annotation and validation.

Day 1: Introduction, Examples, and Text Annotation

- Introduction to Supervised and Unsupervised Learning
- Supervised Learning and annotated texts
- Getting features using computational and hand coding approaches
- Introduction to Annotating a Corpus using MAE

Readings

James, Gareth, et al. *An Introduction to Statistical Learning.* Vol. 112. New York: Springer, 2013. **Read Chapter 2**

Jurka, Timothy P., et al. "RTextTools: A supervised learning package for text classification." *The R Journal* 5.1 (2013): 6-12.

Stubbs, Amber. "MAE and MAI: lightweight annotation and adjudication tools." Proceedings of the 5th Linguistic Annotation Workshop. *Association for Computational Linguistics,* 2011.

Card, Dallas, et al. "The Media Frames Corpus: Annotations of Frames Across Issues" *Association for Computational Linguistics,* 2015.

Day 2: Supervised Learning and ML Algorithms

Focus on Supervised learning and different algorithms for supervised learning. Model performance. Model interpretation. Model validation.

- Introduction to Classification

- K Nearest Neighbors
- Regression and Logistic Regression
- Support Vector Machines
- Decision Trees, Random Forests
- Naive Bayes
- Performance Evaluation

Readings

James, Gareth, et al. *An Introduction to Statistical Learning.* Vol. 112. New York: Springer, 2013. **Read Chapters 3,4,8,9**

Bonica, Adam. "Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning." Working Paper SSRN 2732913 (2015).

Evans, Michael, et al. "Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research." *Journal of Empirical Legal Studies* 4.4 (2007): 1007-1039.

Muchlinski, David, et al. "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data." *Political Analysis* 24.1 (2016): 87-103.

Day 3: Unsupervised Learning and Team Projects

Focus on unsupervised learning and different algorithms for unsupervised learning. Overview of some of the tools being used including scaling, scaling and voting, topic models, and scaling and topic models.

- Clustering K means
- Hierarchical Clustering
- Performance Evaluation
- Scaling and Topic Models
- Class projects

Readings

James, Gareth, et al. *An Introduction to Statistical Learning.* Vol. 112. New York: Springer, 2013. **Read Chapter 10**

Proksch, Sven-Oliver, and Jonathan B. Slapin. "Position taking in European Parliament speeches." *British Journal of Political Science* 40.03 (2010): 587-611.

Lauderdale, Benjamin E., and Tom S. Clark. "Scaling politically meaningful dimensions using texts and votes." *American Journal of Political Science* 58.3 (2014): 754-771.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." the *Journal of Machine Learning Research* 3 (2003): 993-1022.

Baerg, Nicole Rae, and Will Lowe. "A Textual Taylor Rule: Estimating Central Bank Preferences Combining Topic and Scaling Methods." Working Paper (2016).