Text Mining the Great Unread 2017

Curriculum

Title
Text Mining the Great Unread – An introduction to data-intensive methods and digital tools for analysis of texts in the humanities and social sciences

Description of qualifications
Key learning outcome

1. Demonstrate an ability to delineate and critically evaluate research problems related to text analysis in terms of text mining solutions. This involves accessing previous solution to similar problems.
2. Competences in design and implementation of knowledge discovery pipelines that solve research problems related to text analysis. This includes a basic understanding of the various pipeline elements and their dependencies (i.e., data selection, preprocessing and transformation followed by pattern discovery, data mining and evaluation) as well as implementation in open source software.
3. Have an understanding of how to communicate projects and findings in accordance with academic and industrial standards.

Contents

Texts have always been essential to research and education in the humanities and social sciences. Close reading and detailed interpretation have traditionally constituted the standard approach to texts, that is, we combine qualitative methods and theoretically motivated arguments to a small textual corpus with the purpose of understanding the meaning of that corpus. However, the rapid expansion of digital fulltext databases, increasingly faster computers, and advances in language technology are starting to impact the standard approach by offering a new digital and dataintensive paradigm in the study of text. Humanities and social science researchers are beginning to ask new types of questions and propose novel solutions to old problems by using faster and more efficient methods to collect, analyze, and visualize texts.
Many students (as well as researchers) experience a lack of digital competences when faced with text mining, that is, the application of tools and methods to analyze large sets of digitized texts. This is unfortunate because text mining 1) enables students to extract high quality information and acquire new knowledge in a fast and efficient manner; and 2) enhances the qualifications of students for a data-driven job market that is relying on the very same tools and methods. Finally, many tools and methods in text mining are in need of a thorough revision by academics who understand the importance of text meaning and context. Academia and industry alike are therefore in great need of students with text mining skills.
"Text Mining the Great Unread" is an introductory level course to text mining tools and methods in the humanities and social sciences, which will supply participants with sufficient knowledge and experience to develop and implement their own text mining projects. The core of the course is a series of handson workshops supplemented by lectures and tutorials by international researchers and industry

experts. Through the course, participants will become familiar with text mining methods and software for analyzing and visualizing texts. Participants will learn how to write their own text mining application in R and Python. Through the workshops, participants will also be presented with a range of paradigmatic studies and go through explain research design, best practice, and reporting standards. It is possible to work with one's own corpus, but historical and contemporary corpora (both works of fiction, historical documents and websites) are also available in class. Participants are not expected to have prior experience with text mining (i.e., programming, statistics, or visualization).

Program overview

Week 1	Time	Episode	Reading
July 24	09:00-12:00	Text analytics	Fayyad et al. 1996
	13:00-15:00	Computer literacy w. Unix	
July 25	09:00-12:00	Programming w. Python#1	ABSP, chp 7-14
	13:00-15:00	Programming w. Python#2	ABSP, chp 7-14
	15:00-17:00	Project session#1	
July 26	09:00-12:00	Processing raw text	NLPP, chp. 3
	13:00-15:00	Data cleaning	NLPP, chp. 3
	15:00-17:00	in groups	
July 27	09:00-12:00	Word frequencies	Slingerland et al. 2015, Reagan et al. 2015
	13:00-15:00	Ngrams and associations	Church et al. 1990, Pechenick et al. 2015
	15:00-17:00	in groups	
July 28	09:00-12:00	Project session #2	
	13:00-15:00	Visualization	
	15:00-17:00	Social event	

Week 2	Time	Episode	Reading
July 31	09:00-12:00	Machine learning	Brücher el al. 2002, Underwood 2016
	13:00-15:00	Document classification	NLPP, chp. 6
	15:00-17:00	in groups	
August 1	09:00-12:00	Document clustering	Jockers et al. 2010
	13:00-15:00	Latent variable models	Blei 2012, Tangherlini et al 2013
	15:00-17:00	in groups	
August 2	09:00-12:00	Entity extraction	Derczynski et al. 2015
	13:00-15:00	Statistics#1	ISR, chp 4/ISPY, chp. 4
	15:00-17:00	Business analytics	
August 3	09:00-12:00	Statistics#2	ISR, chp 6/ISPY, chp. 11
	13:00-15:00	Word embedding	Mikolov et al. 2013
	15:00-17:00	in group	
August 4	09:00-12:00	Project session #3	
	13:00-15:00	Project session #4	
	15:00-17:00	Social event	

Textbooks

[ABSP: Introduction to Python]
- Sweigart, A. (2015). Automate the Boring Stuff with Python: Practical Programming for Total Beginners (1 edition). San Francisco: No Starch Press.

[NLPP: Introduction to natural language processing]
- Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python (1 edition). Beijing; Cambridge Mass.: O'Reilly Media.

[ISR: Introduction to statistics & R]*
- Dalgaard, P. (2008). Introductory Statistics with R (2nd edition). New York: Springer.

[ISPY: Introduction to statistics & Python]*
- Haslwanter, T. (2016). An Introduction to Statistics with Python. New York: Springer.

* One book on statistics is sufficient, but depending on your field and personal preferences you might want to use *R* instead of Python for data analysis. If on doubt choose ISPY unless you want to broaden your programming knowledge to multiple languages.

Articles

- Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84.
- Brücher, H., Knolmayer, G., & Mittermayer, M.-A. (2002). Document classification methods for organizing explicit knowledge. Institut fur Wirtschaftsinformatik der Universität Bern.
- Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1), 22–29.
- Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., & Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. Information Processing & Management, 51(2), 32–49.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37.
- Jockers, M. L., & Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. Literary and Linguistic Computing, fqq001.
- Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg; Dean, Jeffrey (2013). Distributed Representations of Words and Phrases and their Compositionality. arXiv:1310.4546.
- Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books corpus: strong limits to inferences of socio-cultural and linguistic evolution. PloS One, 10(10), e0137041.
- Reagan, A., Tivnan, B., Williams, J. R., Danforth, C. M., & Dodds, P. S. (2015). Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs. arXiv Preprint arXiv:1512.00531.
- Slingerland, E., & Chudek, M. (2011). The Prevalence of Mind-Body Dualism in Early China. Cognitive Science, 35(5), 997–1007.
- Tangherlini, T. R., & Leonard, P. (2013). Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research. Poetics, 41(6), 725–749.
- Underwood, T. (2016). The Life Cycles of Genres. [Machine learning, Classification] Retrieved from https://www.ideals.illinois.edu/handle/2142/90161.

Preparation Details

- Answer pre-workshop questionnaire (you receive invitation by email)
- Read Part 1 in Automate the Boring Stuff with Python
- Read Chapter 1 in Natural Language Processing with Python
- Install Anaconda distribution of Python: `https://www.continuum.io/downloads`. If in doubt choose the Python 2.7 version.
- If you want to do statistics in *R* (or are simply interested) install *R*: `https://www.r-project.org/`
- Install Git: `https://git-scm.com/book/en/v2/Getting-Started-Installing-Git`. If your operating system is Windows use: `https://git-for-windows.github.io/`
- Make free GitHub account at `https://github.com/`

Should you have any specific requests, please contact: kln@cas.au.dk.

Description of Episodes

Almost all episodes are interactive workshops consisting of short introductory presentations followed by tutorials and exercises. You will all develop projects in groups that you work on during project sessions and in group episodes. Supervision is mandatory and available during in group episodes.

Text analytics

Text mining ($\sim$ text analytics) is a heterogeneous research field that focuses on extraction of meaningful patterns from unstructured and text-heavy data. The meaningful patterns are typically extracted by applying statistical learning (i.e., machine learning) to target data sets from large non-relational databases. This introductory lecture introduces data-intensive knowledge discovery in fulltext databases for humanities and social sciences and we focus on mapping the research possibilites available in the various domains of text analytics.

Computer literacy with Unix

In order to understand what computers can contribute, we practice how to think algorithmically with Unix shell scripting. A shell script is a computer program the is designed to run from a Unix command-line interface. Beyond training computational literacy, scripting in Unix-like systems turns is very handy for managing files.

Programming with Python#1

Introduction to Python basics in Jupyter Notebook. This episode introduce file manipulation, data types, flow control with booleans and conditional statements, and import of modules.

Programming with Python#2

Part 2 of the introduction to python basics which now transitions to the Spyder IDE. This episode introduces functions, classes, errors and debugging. Will will pay particular focus on defensive programming and how to write robust programs that run from the shell.

Processing raw text

How do we extract the relevant data from existing fulltext databases, urls, or even badly scanned PDF files? In this session we will develop principles for building a data set, extract text from websites and documents, and do simple preprocessing steps such as tokenization and stopword removal.

Data preparation

This episode introduces regular expressions for data clarning and information extraction, text normalization with stemming and lemmatization, and bag-of-words models and vector space representations for documents (term-document matricies).

Word frequencies

Having prepared our data set in the previous episodes, we now turn to basic statistical properties of natural language reflected in word frequency distributions. We will explore a simple application that have gained huge popularity in recent years, namely, dictionary-based sentiment analysis that can be used to estimate affective content of texts.

Ngrams and associations

Because words meaning depends critically on their context, this episode will introduce several techniques to model word context and word associations. We will look at word collocations, co-occurrence matrix and similarity-based measures for vector spaces.

Visualization

Visual inpspection of data is a valuable tools both for discovering patterns and communicating results. In this episode we will work with the matplotlib and seaborn libraries in order to check our data integrity, compare patterns and visualize findings.

Machine learning

Machine learning is 'the new black' in data analysis especially when we deal with large data sets with a lot of noise. Instead of classical analysis, we can train models to discover patterns in our data set and generalize to unseen cases. This episode is a lecture on application of machine learning in text analytics.

Document classification

Classification is a supervised learning task, which is one of the central tasks in machine learning. In this episode we will train several types of supervised learning algorithms to classify a set of documents according to some classes or labels (e.g., genre or author features).

Document clustering

While supervised learning assumes labeled data, clustering, which is an unsupervised learning task, can be used to discover meaningful classes in the data. In this episode we will use hard and hierarchical clustering to partition our data sets.

Latent variable models

Latent variable models cover both geometrical and probabilistic approached to modelling latent themes or semantic structures in collections of documents. In this episode we will cover both latent semantic analysis (LSA) and latent Dirichlet allocation (LDA) also called topic modeling in digital humanities.

Entity extraction

Named entities are unique identifiers for entities in texts (a proper noun serving as a name for someone or something). We can use entity extraction to detect proper nouns in texts and classify them into categories such as person, localization and organization.

Statistics#1

Introduction to descriptive statistics and graphics in Python or R. Before going into the actual statistical modelling and analysis of a text data set, it is often useful to make some simple characterizations of the data interms of summary statistics and graphics.

Business analytics

One of AU's leading specialists in Business Analytics will explain how we can use text data to explore and investigate business performance in order to gain insight and drive business planning. This episode is a lecture.

Statistics#2

A look at simple statistical tests for comparing data between two classes or a prior stipulated class and techniques for predicting variation in a response variable based on one and multiple explanatory variables.

Word embedding

Word embedding algorithms have become popular tools for uncovering the full associative structure of large text collections. Using an artificial neural network, we can construct high-dimensional vector representations of words, sentences, and documents and explore their similarity structure.

Kristoffer L. Nielbo
Associate Professor, MA, PhD,
Interacting Minds Centre, Department of Culture and Society.
Jens Chr. Skous Vej 4, building 1483, 326, 8000 Aarhus C., DK.
Email: **kln@cas.au.dk**
Phone: **+45 8716 2903**
Mobile: **+45 2683 2608**