Unstructured Data Analysis (Text Analytics)

2019 Spring

School of Industrial Management Engineering

1. Overview

- ✓ This module aims to provide students with the theoretical and practical knowledge and skills to collect, modify, and analyze a large amount of unstructured data, especially texts, from various sources.
- ✓ Topics covered in this module include data collection methods from various sources, preprocessing methods including natural language processing, document representation & summarization, feature selection and extraction, document clustering, document classification, and topic models.
- ✓ The students are assessed by one final exam at the end of the semester, three presentations (proposal, interim, and final) and the final manuscript for their term projects.

2. Lecturer & Course homepage

- ✓ Pilsung Kang, Assistant professor at School of Industrial Management Engineering, Korea University
  - ■ E-mail: pilsung_kang@korea.ac.kr
  - ■ Course homepage: https://github.com/pilsung-kang/text-mining

3. Textbook and additional resources (not mandatory)

- ✓ Weiss, S.M., Indurkhya, N., and Zhang, T. (2010). Fundamentals of Predictive Text Mining. Springer.
- ✓ Feldman, R. and Sanger, J. (2007). The Text Mining Handbook. Cambridge University Press.
- ✓ Kao, A. and Poteet, S.R. (2007). Natural Language Processing and Text Mining. Springer.
- ✓ Manning, C.D., Raghavan, P., and Schutze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
- ✓ Jurafsky, D. and Martin, J.H. (2008). Speech and Language Processing, 2nd Ed. Prentice Hall. (Free online course available: https://www.youtube.com/playlist?list=PL6397E4B26D00A269)
- ✓ Socher, R. (2016). CS224d @Stanford: Deep learning for natural language processing (course homepage: http://cs224d.stanford.edu/, video lectures are available at Youtube)
- ✓ Blunsom, P. et al. (2017). Deep natural language processing @Oxford (course homepage: https://github.com/oxford-cs-deepnlp-2017/lectures)

4. Assessments

- ✓ Final exam (50%): Closed book
- ✓ Term project (50%): three presentations
  1. Group project: maximum 4 students in a group
  2. Proposal (10%): purpose of the project (task), data description, expected effects, etc.
  3. Interim presentation (15%): data collection/preprocessing, feature extraction, issues to be discussed.
  4. Final presentation (25%): employed/developed models, experimental results including interesting patterns discovered, limitations and future research directions.

5. Introduce yourself

- ✓ Submit your self-introduction slide (max. 5 pages) to the lecturer via E-mail by the end of the 2nd week.

Schedule & Topics

Week	Date	Contents
1	3/5	Orientation
	3/7	Introduction to Text Analytics ✓ The usefulness of large amount of text data and the challenges
2	3/12	Text Preprocessing: Natural Language Processing 1 ✓ Introduction to NLP, Lexical Analysis 1
	3/14	Text Preprocessing: Natural Language Processing ✓ Lexical Analysis 2, Other Topics in NLP
3	3/19	Neural Network Basics 1 ✓ Neural Network: Overview
	3/21	Neural Network Basics 2 ✓ Convolutional Neural Network, Recurrent Neural Networks
4	3.26	Neural Network Basics 3 ✓ Auto-Encoder, Some Practical Techniques
	3/28	Document Representation 1 ✓ Bag-of-Words, TF-IDF, N-Gram
5	4/2	Document Representation 2 ✓ Word Embedding: NNLM, Word2Vec
	4/4	Document Representation 3 ✓ GloVe, FastText, Sentence/Document Embedding
6	4/9	Dimensionality Reduction: Feature Selection and Extraction 1 ✓ Supervised feature selection: index term selection, information gain
	4/11	No Class (IE conference)
7	4/16	Dimensionality Reduction: Feature Selection and Extraction 2 ✓ Unsupervised feature selection: latent semantic analysis (LSA)
	4/18	Document Similarity & Clustering 1 ✓ Document similarity measures: cosine similarity, Euclidean distances, etc.
8	4/23	No class (Midterm exam break)
	4/25	Document Similarity & Clustering 2 ✓ Clustering algorithms: K-means clustering, hierarchical clustering, DBSCAN, etc.
9	4/30	Topic Modeling 1 ✓ Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (pLSA)
	5/2	Topic Modeling 2 ✓ Latent Dirichlet Allocation (LDA) 1: Document Generation Process
10	5/7	Topic Modeling 3 ✓ Latent Dirichlet Allocation (LDA) 2: Inference based on Gibbs Sampling
	5/9	Document Classification 1 ✓ Naïve Bayesian classifier, k-nearest neighbor classifier ✓ Classification performance evaluation
11	5/14	Document Classification 2 ✓ CNN/RNN-based Document Classification
	5/16	Sentiment Analysis 1 ✓ Dictionary-based sentiment analysis
12	5/21	Sentiment Analysis 2 ✓ Model-based sentiment analysis
13-14		Term project
15		**Final Exam**
16		**Term Project Final Presentation**