Unstructured Data Analysis (Text Mining)

2017 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: http://dsba.korea.ac.kr → Courses → Unstructured Data Analysis (Graduate, 2017 Spring)

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)

4. Assessments

- ✓ Final exam (30%): Closed book
- ✓ Term project (70%): 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 (10%): data collection/preprocessing, feature extraction, issues to be discussed.
  4. Final presentation (20%): employed/developed models, experimental results including interesting patterns discovered, limitations and future research directions.
  5. Research paper (20%): each team **must** write a research paper based on the results of the term

project. The final manuscript should be completed by the end of the 16th week. The type of the manuscript determines the maximum grade.

1. A+: international journal paper (must be written in English)
2. A: domestic journal paper (written in either English or Korean)

5. Introduce yourself

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

6. Schedule & Topics

Week	Contents	Exercises
1	Orientation	
2	Introduction to Text Mining ✓ The usefulness of large amount of text data and the challenges ✓ Overview of text mining methods	R Exercise
3	From Texts to Data ✓ Obtain texts to analyze: text files, databases, Facebook APIs, and web scraping	R Exercise
4	**Term Project Proposal**	
5	Natural Language Processing I ✓ Morphological analysis: tokenization, stemming/lemmatization, POS tagging, parsing, chunking, named entity recognition, language model, etc.	
6	Document Summarization ✓ Summarize a large amount of texts to understand at a glance (Wordcloud) ✓ Interpret the relationship between features (Association rules, Graphs)	R Exercise
7	Dimensionality Reduction I: Feature Selection ✓ Supervised feature selection: index term selection, information gain, cross entropy, etc.	R Exercise
8	Dimensionality Reduction II: Feature Extraction ✓ Unsupervised feature selection: latent semantic analysis (LSA), Word Embedding, etc.	R Exercise
9	**Term Project Interim Presentation**	
10	Document Similarity & Clustering ✓ Document similarity measures: cosine similarity, Euclidean distances, etc. ✓ Clustering algorithms: K-means clustering, hierarchical clustering	
11	Document Classification ✓ Naïve Bayesian classifier, k-nearest neighbor classifier ✓ Classification performance evaluation	R Exercise
12	Topic Modeling I: Probabilistic Latent Semantic Analysis (pLSA)	R Exercise
13	Topic Modeling II: Latent Dirichlet Analysis (LDA)	R Exercise
14	Sentiment Analysis ✓ Dictionary-based sentiment analysis ✓ Model-based sentiment analysis	R Exercise
15	Prepare the final presentation and write the manuscript	
16	**Final Exam** **Term Project Final Presentation**	