Unstructured Data Analysis (Text Analytics)

2020 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, Associate 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)

✓ Manning, C. (2020). CS224n: Natural language processing with deep learning

- Course homepage: http://web.stanford.edu/class/cs224n/

✓ Socher, R. (2017). 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 (40%): Closed book

✓ Term project (40%): 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-minutes Youtube video (20%)

1. Students must upload a short video (max 5 minutes) that reviews the lecture within 24 hours after the class.
2. A student explains what he/she learns in the class to his/partner.

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

Week	Date	Contents
1	3/3	Orientation
	3/5	Introduction to Text Analytics ✓ The usefulness of large amount of text data and the challenges
2	3/10	Text Preprocessing ✓ Tokenization (Stemming, Lemmatization), POS Tagging
	3/12	Text Preprocessing ✓ Parsing, etc.
3	3/17	Text Representation 1 ✓ Bag-of-Words, N-Grams
	3/19	Text Representation 2 ✓ Word Embedding: NNLM, Word2Vec
4	3/24	Text Representation 3 ✓ GloVe, FastText
	3/26	Text Representation 4 ✓ Skip-thought, Doc2Vec
5	3/31	Topic Modeling (can be used as a document representation) 1 ✓ Latent Semantic Analysis (LSA), probabilistic LSA (pLSA)
	4/2	Topic Modeling (can be used as a document representation) 2 ✓ Topic Modeling: Latent Dirichelet Allocation (LDA) 1
6	4/7	Topic Modeling (can be used as a document representation) 3 ✓ Topic Modeling: Latent Dirichelet Allocation (LDA) 2
	4/9	Topic Modeling (can be used as a document representation) 4 ✓ Topic Modeling: Latent Dirichelet Allocation (LDA) 3
7	4/14	Language Modeling and Pretrained Models 1 ✓ Language Models Overview, Transformer 1
	4/16	Language Modeling and Pretrained Models 2 ✓ Transformer 2
8	4/21	Language Modeling and Pretrained Models 3 ✓ ELMo, GPT
	4/23	Language Modeling and Pretrained Models 4 ✓ BERT
9	4/28	Text Classification & Sentiment Analysis 1 ✓ Text Classification Overview, Naïve Bayesian Classifier
	4/30	No Class
10	5/5	No Class
	5/7	Text Classification & Sentiment Analysis 2 ✓ CNN-based Model, RNN-based Model
11	5/12	Text Classification & Sentiment Analysis 3 ✓ Sentiment Classification
	5/14	Sequence to sequence (Seq2seq) Model 1 ✓ Question Answering 1
12	5/19	Sequence to sequence (Seq2seq) Model 2 ✓ Question Answering 2
	5/21	Sequence to sequence (Seq2seq) Model 3 ✓ Open Information Extraction
13-14		Term project
15		**Final Exam**
16		**Term Project Final Presentation**