Unstructured Data Analysis (Text Analytics)

2021 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
  - Course Youtube playlist: https://youtube.com/playlist?list=PLetSlH8YjIfVzHuSXtG4jAC2zbEAErXWm
  - Term project presentation and Youtube summary video:
  - https://docs.google.com/spreadsheets/d/118AMJWTO90ecanvUadqqEOIHj9bVCNlcMHZkPOjBpnc/edit?usp=sharing

3. Course Structure

- ✓ Online video lecture
  - Students are required to watch the assigned videos before the discussion and Q & A session
- ✓ Tuesday: Discussion and Q & A session
- ✓ Thursday: Term project meeting

4. 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

5. Assessments

- ✓ Final exam (20%): 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.
- ✓ Discussion and QA participation (20%)
  1. How many good questions does a student ask?
  2. How many good answers does a student prepare?

6. Introduce yourself

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

7. Schedule

Week	Topics
1	Orientation Introduction to Text Analytics Text Preprocessing: Tokenization (Stemming, Lemmatization), POS Tagging, Parsing, etc.
2	Text Representation 1: Bag-of-Words, N-Grams, Word Embedding: NNLM, Word2Vec
3	Text Representation 2: GloVe, FastText, Skip-thought, Doc2Vec
4	Topic Modeling 1: Latent Semantic Analysis (LSA), probabilistic LSA (pLSA)
5	Topic Modeling 2: Latent Dirichelet Allocation (LDA)
6	Language Modeling and Pretrained Models 1: Overview and Transformer
7	Language Modeling and Pretrained Models 2: ELMo, GPT, BERT
8	Text Classification 1: Overview, Count-based Models
9	Text Classification 2: CNN-based Models, RNN-based Models
10	Text Analytics Task 1: Sentiment Analysis
11	Text Analytics Task 2: Text (Extractive) Summarization
12	Text Analytics Task 2: Text (Abstractive) Summarization
13	Text Analytics Task 3: Question Answering 1
14	Text Analytics Task 3: Question Answering 2
15	Text Analytics Task 4: (Open) Information Extraction