Introduction to Text Mining

Evangelos Theodoridis
Data Engineer
Knowledge Graph Team
6th April 2017
Topics

1. Key Concepts in Text Mining
2. Use Cases & Applications
3. Challenges
Data Mining :: Key Concepts

**Definition:** “The efficient discovery of previously unknown, valid, potentially useful, understandable patterns in large datasets”

**Alt. Definition:** “The analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful”

**Application Domains:**
- Business Intelligence
- Science
- Web/Social Networks
- and many more

{Prediction, Description} Methods
Data Mining Tasks
Text Mining :: Key Concepts

**Definition:** extracting meaningful, useful, previously unknown and ultimately comprehensible information from **textual document repositories**

**Text data sources:** Surveys, Emails, Call Center, Comments, complains, social network posts, abstracts, web pages/blogs, news feeds, ebooks, scientific articles ...

Structured vs Semi-structured and Unstructured Data

**Text Mining = Data Mining + Computational Linguistics (NLP)**
Text Mining :: Applications & Use cases

1. Business Intelligence
2. Scientific Discovery
3. E-Government
4. Social Media monitoring
5. Cybercrime
6. Sentiment Analysis & Opinion Mining
7. Recommendations
8. Automated translation
9. Information Retrieval
10. Publishing
Text Mining :: Techniques

Linguistic Analysis:
1. Syntactical Analysis
2. Grammatical Analysis
3. Part-of-speech Recognition
4. Word sense disambiguation
5. Thesaurus Construction

Document Summarization
Clustering Algorithms
Classification Algorithms

Topic/Subject/Concept & Relationship Extraction
Named Entity Recognition
Meaning Extraction
Sentiment Analysis
Association rules Extraction
Example 1 :: Text Mining in Biomedical Literature


Example 1:: Text Mining in Biomedical Literature

http://nactem7.mib.man.ac.uk/geniatagger
Example 2 :: Web/Content Recommendation

Data sources:
• Content of the documents (Web pages)
• User behavior captures sequences of document(web page) views

Model:
• Similar content in groups of documents
• Similar user behavior/preferences in groups of users
• Combine models to provide recommendations

Charu Aggarwal *Content-Based Recommender Systems* Recommender Systems pp 139-166, Springer, 29 March 2016. [https://link.springer.com/chapter/10.1007%2F978-3-319-29659-3_4](https://link.springer.com/chapter/10.1007%2F978-3-319-29659-3_4)
Example 3 :: Springer Nature SciGraph

www.springernature.com/scigraph

Springer Nature SciGraph: “the new Linked Open Data platform aggregating data sources from Springer Nature and key partners from the scholarly domain”
Challenges

Computationally demanding tasks (Huge number of dimensions)

Dealing with natural language

• Unstructured content
• Language is ambiguous
• Homonymy, synonymy, polysemy, hyponymy
• Misspellings, abbreviations, spelling variants, slang
• Multilingual context

Application driven – domain knowledge integration

Ethical and legal issues – privacy and intellectual property rights
References

Reading Material


Software:

https://opennlp.apache.org/
http://www.nltk.org/
https://gate.ac.uk/
http://alias-i.com/lingpipe/
https://community.rapidminer.com/
https://nlp.stanford.edu/software/CRF-NER.shtml
Thank you

Evangelos Theodoridis

evangelos.theodoridis@springernature.com

www.springernature.com/scigraph