Text Analysis and Retrieval: Learning Outcomes

UNIZG FER, Academic Year 2021/2022

Published: 15 March 2022

This document defines the Text Analysis and Retrieval AY 2021/2022 course learning outcomes. The learning outcomes are statements that describe what you should learn in this course and what you should be capable of doing by the end of the course. They serve to increase your awareness of your own learning and to make it easier for you to prepare for knowledge and skill assessments.

Course-Level Learning Outcomes

1. Summarize the application areas, trends, and challenges in text analysis and retrieval
2. Explain the fundamental techniques of text analysis and retrieval
3. Describe, analyze, and criticize the main text analysis methods and results from recent scientific papers
4. Use linguistic preprocessing tools
5. Apply machine learning algorithms to text analysis/retrieval tasks
6. Design, implement, and evaluate a system for a concrete text analysis/retrieval task
7. Structure and write a text analysis system description paper or a preliminary research paper

To the above learning outcomes are achieved through lectures (theory part), team project, and paper reading sessions. In particular, the lectures contribute to learning outcomes 1 and 2, paper reading sessions contribute to learnign outcomes 3 and 7, while the team project contributes to learning outcomes 4–7.

Lecture-Level Learning Outcomes

Below we list the specific learning outcomes for each class session, which jointly contribute to course-level outcomes 1 and 2.

1 Lecture: Introduction

1.1. Define information retrieval

1.2. Define text mining and explain how it differs from information retrieval and data mining

1.3. List the main reasons why text analysis and retrieval (TAR) is challenging

1.4. List the various types of language ambiguity and give an example for each

1.5. Explain what Natural Language Processing (NLP) is

2 Lecture: Basics of Natural Language Processing

2.1. Basic NLP pipeline

- 2.1.1. Describe the components of a basic NLP pipeline
- 2.1.2. Describe what POS tagging is and why we need it
- 2.1.3. Explain stemming and lemmatization, why we need it, and the difference between them
- 2.1.4. List the main tools available

2.2. Syntactic parsing

- 2.2.1 Describe what parsing is and why we need it
- 2.2.2 Differentiate between phrase-based and dependency-based parsing
- 2.2.3 Describe what chunking is and why we need it
- 2.2.4 List the main tools available for parsing/chunking

2.3. Corpora & language models

- 2.3.1. Describe what a corpus is, explain why we need it, and name a few
- 2.3.2. Describe what a language model is and what it is used for
- 2.3.3. Define and explain the probablity modeled by an n-gram language model
- 2.3.4. Differentiate between statistical and neural language models

3 Lecture: Basics of Information Retrieval

3.1. Main IR models

- 3.1.1. Explain the two typical IR preprocessing steps
- 3.1.2. Describe the three main components of an IR model
- 3.1.3. Describe the vector space model and the TF-IDF weighting scheme
- 3.1.4. Explain the probability ranking principle and the BM25 ranking function
- 3.1.5. Describe the LM information retrieval model
- 3.1.6. List the main IR tools available

3.2. IR evaluation

- 3.2.1. Explain what an IR test collection consist of and what it's used for
- 3.2.2. Define and calculate the standard IR evaluation metrics (P, R, F1, MAP, P@K, R-precision)

3.3. Link analysis with PageRank

- 3.3.1. State the PageRank hypothesis
- 3.3.2. Define and explain the PageRank iterative update formula in matrix form

4 Lecture: Machine Learning for NLP

4.1. Framing NLP tasks as ML problems

- 4.1.1. List at least three advantages and disadvantages of machine-learning-based NLP systems
- 4.1.2. Illustrate how to frame standard NLP tasks as ML problems and what features to use
- 4.1.3. Explain what are lexical features are and how to encode them using one-hot encoding
- 4.1.4. Explain approaches to feature design and analysis

4.2. Sequence labeling

4.2.1. Explain what sequence labeling is and why we need it

4.2.2. Explain the basic idea behind HMM and its main weakness

4.2.3. Explain the basic idea behind MEMM and how it differs from HMM

4.2.4. Explain the basic idea behind CRF and how it differs from MEMM

4.3. Data annotation

4.3.1. Describe the typical annotation workflow

4.3.2. Compute and interpret the kappa coefficient for a given confusion matrix

5 Lecture: Semantics

5.1. Lexical semantics

5.1.1. Define and exemplify polysemy and the main lexical relations

5.1.2. Describe WordNet and give an example of synsets involving a polysemous word

5.1.3. Describe the purpose and the main approaches to word sense disambiguation

5.1.4. Describe frame semantics and FrameNet, and give an example of a frame

5.2. Distributional semantics

5.2.1. State the distributional hypothesis and give an example

5.2.2. Explain what a distributional semantic model is, how it's constructed, and what it's used for

5.2.3. Differentiate between sparse/dense and count-based/predictive vector representations

5.2.4. Define distributional semantic composition and the simplest approach to it

5.3. Semantic parsing

5.3.1. Define semantic parsing and explain how it differs form syntactic parsing

5.3.2. Differentiate between deep and shallow semantic parsing

5.3.3. Define semantic role labeling and illustrate how it can be framed as a machine learning task

5.3.4. Differentiate between FrameNet and PropBank semantic roles

6 Lecture: Neural NLP

6.1. Neural word representations

6.1.1. Explain what word embeddings are and what they can be used for

6.1.2. Describe the skip-gram training setup and provide an example of a training instance

6.1.3. Describe the CBOW training setup, compare it to skip-gram and provide an example of a training instance

6.1.4. Define negative sampling and explain what we use it for

6.2. Neural models for natural language processing

6.2.1. Explain why traditional neural models cannot be used for NLP

6.2.2. Describe four sequence processing use-patterns and name a typical NLP task for each

6.2.3. Describe the vanilla RNN and explain and exemplify its deficiencies when used for NLP

6.2.4. Describe the LSTM model and exemplify its mechanisms on a sentence processing task

6.2.5. Describe the transformer model and its advantages over an RNN

6.3. Contextualized word representations

6.3.1. Explain what contextualized embeddings are and the motivation for their use

6.3.2. List and compare training setups for learning contextualized representations

6.3.3. List the main RNN- and transformer-based contextualized representation models

7 Lecture: Recap & Prep

7.1. List the main TAR tasks

7.2. List the main conferences in the NLP community

7.3. Explain the main functional structure of an NLP paper

7.4. Explain how to read and summarize a scientific paper

7.5. Describe how to structure the project report paper

8 Paper Reading: Information Extraction

8.1. Describe the main information extraction tasks and give examples of each

8.2. Describe, review, analyze, and criticize the methods and results present in a research paper on the selected topic

9 Paper Reading: Question Answering

9.1. Describe what question answering is and give an example

9.2. Describe, review, analyze, and criticize the methods and results present in a research paper on the selected topic

10 Paper Reading: Text Summarization

10.1. Describe what text summarization is and give an example

10.2. Describe, review, analyze, and criticize the methods and results present in a research paper on the selected topic

11 Paper Reading: Sentiment Analysis

11.1. Describe what sentiment analysis is and give an example

11.2. Describe, review, analyze, and criticize the methods and results present in a research paper on the selected topic

12 Paper Reading: Author Analysis; Wrap-up

12.1. Describe what author profiling is and give an example

12.2. Describe, review, analyze, and criticize the methods and results present in a research paper on the selected topic

12.3. Name three hot topics in NLP/IR