# Computational Linguistics course materials

Course credits: 3

Materials:

* Textbook: Jurafsky, D., Martin, J. H. (**J&M**), “Speech and Language Processing”, 3rd edition (online, 2019) <https://web.stanford.edu/~jurafsky/slp3/>
* One chapter from: Eisenstein, J. (**E**), “Natural Language Processing” (online, 2018) <https://github.com/jacobeisenstein/gt-nlp-class/tree/master/notes>

(Section 4 and 5 (lecture 8-11) could be removed for a smaller number of credits)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **DS&AI course building blocks** | COMPUTATIONAL LINGUISTICS | Literature | Slides | Tutorial or assignment |
| **Section 1** | **Introduction to Computational Linguistics and Natural Language Processing** | | |  |
| Lesson 1 | Introduction: challenges of text data | None | L01\_NLP-introduction.pdf (Felipe Bravo-Marquez)  L01\_introduction.pptx (Suzan Verberne) |  |
| **Section 2** | **Pre-processing** | | |  |
| Lesson 2 | Tokenization and sentence splitting, morphology | J&M chp 2.2, 2.3, 2.4 | L02\_TextProc.pdf (Jurafsky & Martin) | Language processing with Spacy |
| Lesson 3 | Regular expressions and edit distance | J&M chp 2.1, 2.5 | L03\_EditDistance.pdf, L03\_Regex.pdf (Jurafsky & Martin) |  |
| **Section 2** | **Language processing** | |  |  |
| Lesson 4a | POS tagging | J&M chp 8 | L04\_POS-Tagging.pdf (Nathan Schneider) |  |
| Lesson 4b | Sequence labelling and recurrent neural networks | J&M chp 9 | L04b\_RNNs.pdf (Christopher Manning) | CRFsuite: sequence labelling for Named entity recognition |
| Lesson 5 | Sequence-to-sequence transformation | None | L05\_NLP-seq2seq.pdf (Felipe Bravo-Marquez) | GPT2-2 text generation tutorial |
| **Section 3** | **Word representations (vector semantics and embeddings)** | |  |  |
| Lesson 6 | Vector space model and classification | None | L06\_text\_categorization.pptx (Suzan Verberne) | Scikit-learn: Working with text data. |
| Lesson 7 | Word embeddings | J&M chp 6 | L07\_vector2.pdf (Jurafsky & Martin) | Word embeddings in Python with gensim |
| **Section 4** | **Syntactic processing** | |  |  |
| Lesson 8 | Constituent grammar and parsing | J&M chp 12 & 13 | L08\_parsing\_algo.pdf (Diyi Yang) |  |
| Lesson 9 | Dependency parsing | J&M chp 15 | L09\_dependency1.pdf (Diyi Yang) |  |
| **Section 5** | **Semantic analysis** | |  |  |
| Lesson 10 | Word sense disambiguation | J&M chp 19 | L10\_Chapter18.wsd.pdf (Jurafsky & Martin) |  |
| Lesson 11 | Semantic role labelling | J&M chp 20 | L11\_22\_SRL.pdf (Jurafsky & Martin) |  |
| **Section 6** | **Classification Models** | |  |  |
| Lesson 12 | Naïve Bayes | J&M chp 4 | L12\_Naive\_Bayes.pptx (Suzan Verberne) | **assignment 1**: Assignment\_1\_text\_categorization.pdf |
| Lesson 13 | Feedforward neural networks | J&M chp 7 | L13\_feedforward\_NN.pptx (Suzan Verberne) |  |
| **Section 7** | **Neural NLP and transfer learning** | |  |  |
| Lesson 14 | Contextual word representations (BERT language models) | None | L14\_contextual-representations.pdf (Christopher Manning) | BERT word embeddings in Huggingface |
| **Section 8** | **Applications of Computational Linguistics** | |  |  |
| Lesson 15 | Information Extraction | J&M chp 18 | L15\_information\_extraction.pptx (Suzan Verberne) | **assignment 2**: Assignment\_2\_sequence\_labelling.pdf |
| Lesson 16 | Sentiment Analysis | None | L16\_sentiment\_analysis.pptx (Suzan Verberne) |  |
| Lesson 17 | Dialogue systems /Conversational agents | J&M chp 26 | L17\_convagents1.pdf (Jurafsky & Martin) | **assignment 3**: Assignment\_3\_sentiment\_analysis.pdf |
| Lesson 18 | Machine translation | E chp 18 | L18\_machine\_translation\_1.pdf (Diyi Yang) |  |

Sources for slides:

* <https://web.stanford.edu/~jurafsky/slp3/> (Jurafsky & Martin)
* <http://web.stanford.edu/class/cs224n/> (Christopher Manning)
* <http://tmr.liacs.nl/TM.html> (Suzan Verberne)
* <https://www.cc.gatech.edu/classes/AY2020/cs7650_spring/> (Diyi Yang)
* <http://www.cs-114.org/course-schedule/> (Nathan Schneider)
* <https://github.com/dccuchile/CC6205/> (Felipe Bravo-Marquez)

Tutorials:

* Language processing with Spacy: <https://realpython.com/natural-language-processing-spacy-python/>
* Scikit-learn: Working with text data. <https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html>
* CRFsuite: sequence labelling for Named entity recognition <https://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html>
* Word embeddings in Python with gensim: <https://machinelearningmastery.com/develop-word-embeddings-python-gensim/>
* BERT word embeddings in Huggingface: <https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/>
* GPT2-2 text generation tutorial: <https://minimaxir.com/2019/09/howto-gpt2/>