1. Annotation

Course Description

This course gives introductory insights into statistical methods that are used in natural language processing systems. The goal of this course:
= understand statistical methods for language processing in detail
= feeling for language tech applications, avoiding pitfalls
= ability to plan technology requirements for a language tech project
= analyze and evaluate the use of NLP in applications
= see the beauty of language technology, be ready to write your thesis in language tech.

This is an introductory NLP course dedicated to classic algorithms based on the Jurafsky and Martin textbook. Later, we will offer also an additional course covering neural methods for NLP based on the Goldberg textbook.

Course Prerequisites

Required: No knowledge beyond general computer science on BA-level
Advantageous:
- introductory knowledge of machine learning
- introductory knowledge of statistics

2. Structure and Content

Course Academic Level

Master-level
<table>
<thead>
<tr>
<th>Topic</th>
<th>Summary of Topic</th>
<th>Lectures (# of hours)</th>
<th>Seminars (# of hours)</th>
<th>Labs (# of hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction into Natural Language Processing</td>
<td>Formal Grammar, Derivation, Formal Language, Automaton, Generation and Acceptance, FSA, Pushdown Automaton, Finite State Automaton, The Chomsky Hierarchy of Formal Languages</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Formal Languages and Automata</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computational Morphology</td>
<td>Morphology with FSAs, Finite State Transducer, Regular Relations, Applications of FSTs in Language Technology, Tries, DAWGs</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Graph-Based Methods</td>
<td>Motivation for Graph Representation, Page Rank, Random Surfer Model, LexRank, Sentence Similarity Graph, Graph Clustering, Chinese Whispers</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Syntactic Parsing</td>
<td>Chunking, Syntactic Parsing with CFGs, Dynamic Programming Parsing: CYK, Earley, Probabilistic CYK Parsing, PCFG, Dependency parsing, Nivre’s algorithm</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Distributional Semantics</td>
<td>LSA, pLSA, LDA, Neural Word Embeddings, Word2Vec</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Word Senses and their Disambiguation</td>
<td>WordNet, Word Sense Disambiguation, Word Sense Induction</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

### 3. Assignments

### 4. Grading

<table>
<thead>
<tr>
<th>Type of Assessment</th>
<th>Graded</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Grade Structure</th>
<th>Activity Type</th>
<th>Activity weight, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Homework Assignments</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Final Exam</td>
<td>50</td>
</tr>
</tbody>
</table>

### Grading Scale
 Attendance Requirements  Mandatory with Exceptions

## 5. Basic Information

### Maximum Number of Students

<table>
<thead>
<tr>
<th></th>
<th>Maximum Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall:</td>
<td>40</td>
</tr>
<tr>
<td>Per Group (for seminars and labs):</td>
<td></td>
</tr>
</tbody>
</table>

### Course Term (in context of Academic Year)

Term 4

### Course Delivery Frequency

Every year

### Students of Which Programs do You Recommend to Consider this Course as an Elective?

<table>
<thead>
<tr>
<th>Masters Programs</th>
<th>PhD Programs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Science</td>
<td></td>
</tr>
</tbody>
</table>

### Course Tags

- Math
- Programming
- Statistics

## 6. Textbooks and Internet Resources
7. Facilities

Software

Python 3.7, R 3.5.

8. Learning Outcomes

Knowledge

Understand statistical methods for natural language processing in detail.

Skill

Select the proper language models and computational methods for a variety of NLP tasks.
Analyze and evaluate statistical NLP methods in various applications.
Develop statistical models and methods for various NLP applications.
Conduct methodological research in natural language processing.

Experience

Implementation of a range of computational methods and linguistic models for various NLP-problems.

9. Assessment Criteria
<table>
<thead>
<tr>
<th>Select Assignment 1 Type</th>
<th>Homework</th>
</tr>
</thead>
</table>
| Input Example(s) of Assignment 1 (preferable) | = Construct FSA/FST  
= Using given HMM model, derive probability of a sequence  
= Train HMM model for several iterations  
= Train CRF using an external program package  
= Parse sequence with a given CFG and algorithm  
= Train a model for statistical parsing  
= Train a topic model  
= Implement a clustering algorithm  
= Use pretrained word embeddings for word sense disambiguation  
= etc. |
| Assessment Criteria for Assignment 1 | Yes/no assessment.  
50% of homework assignments are required. |

10. Additional Notes