Skoltech
Skolkovo Institute of Science and Technology

Course Syllabus

Course Title (in English)	Statistical Natural Language Processing
Course Title (in Russian)	Статистические методы обработки естественного языка
Lead Instructor(s)	Alexander Panchenko

Contact Person	Alexander Panchenko
Contact Person's E-mail	a.panchenko@skoltech.ru

Course Description

This course gives introductory insights into methods that are used in natural language processing systems. This is an introductory NLP course dedicated to classic algorithms and models for NLP yet with the coverage of some more recent neural models. The course is largely based on the Jurafsky&Martin textbook, but also features lectures on graph-based models for NLP and data annotation for NLP.

If you would like to get a course on purely "modern" neural NLP methods similar to Stanford's CS224n, then you shall enroll in the "Neural Natural Language Processing" course at Skoltech. Thus, given a very broad scope of NLP, we decided to split the sheer volume of material into these two complementary 3 credit courses.

Goals of this course:

- understand methods for language processing in detail
- the 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.

Course Prerequisites / Recommendations

Required: No knowledge beyond general computer science on BA-level

Advantageous:
- introductory knowledge of machine learning
- introductory knowledge of statistics

Аннотация

Данный курс дает вводную информацию о методах, которые используются в системах обработки естественного языка (NLP). Этот вводный курс, посвящен классическим алгоритмам и моделям NLP, однако он включает и обзор более современных нейронных моделей. Курс в значительной степени основан на учебнике Джурафски и Мартина, но также включает лекции по графовым моделям для NLP и аннотации данных для NLP.

Если вы хотите пройти курс с фокусом на «современные» методы нейронного NLP, подобный CS224n Стэнфордского университета, вам следует записаться на курс «Neural Natural Language Processing» в Сколтехе. Таким образом, учитывая очень широкий охват области NLP, мы решили разделить весь объем материала на эти два взаимодополняющих 3-х кредитных курса.

Цели этого курса:

- подробно разбираться в методах автоматической обработки текста
- способность планировать технологические требования для NLP проекта
- анализировать и оценивать использование NLP в приложениях
- познать красоту языковых технологий, захотеть написать диссертацию по технологиям обработки языка.

Course Academic Level	Master-level
Number of ECTS credits	3

Topic	Summary of Topic	Lectures (# of hours)	Seminars (# of hours)	Labs (# of hours)
Introduction into Natural Language Processing and Morphology	Introduction to the field of Natural Language Processing (NLP). Basic tasks. Morphological analysis.	2	2	
Distributional Semantics and Word Sense Disambiguation	Sparse count-based vector spaces. Word Embeddings, Word2Vec, GloVe and related models. WordNet, Word Sense Disambiguation, Word Sense Induction.	2	2	
Sequence Tagging	Sequence tagging task and its applications such as Named Entity Recognition. Conditional Random Field and related models.	2	2	
Language Models and Machine Translation	Language Models: unigram/bigram/n-gram, Markov Chain, Zipf's law, Hidden Markov Models. Probabilistic and neural models for machine translation.	2	2	
Graphs for NLP: Networks and Clustering	Motivation for Graph Representation in NLP. Types of graphs in NLP tasks. Graph Clustering, Chinese Whispers and related algorithms. Applications.	2	2	
Graphs for NLP: Embeddings and GCNs	An overview of neural approaches for learning dense representations of nodes in graphs which can be useful for processing of linguistic network data such as node2vec and GCNs. Applications to taxonomy enrichment and entity linking.	2	2	
Data Annotation and Crowdsourcing for NLP	The majority of the successful NLP models rely on linguistically annotated data in one form or another. Often in practical applications for a given language and domain such a dataset is not available not making it possible to apply the supervised models. In this lecture you will learn how to setup creating of the required data.	2	2	
Syntax	An overview of various types of syntactic parsing common in NLP: chunking, dependency parsing, constituency parsing and others.	2	2	

Assignment Type	Assignment Summary
Homework Assignments	Lexical semantics. In this assignment, the student will be able to select among two task where distributional models, such as word2vec, are to be used and possibly improved: word sense induction and taxonomy enrichment. The solutions shall be submitted to the respective online competitions based on the CodaLab platform with a public leaderboard.
Homework Assignments	Named Entity Recognition (NER). In this assignment NER a related sequence tagging task will be addressed. The students will be asked to develop an NLP for solving the NER or a similar task on their own. The result will be a report in the structure of a paper and the code of the solution. Quality of the obtained results and innovativeness of the used model will be judged. he solutions shall be submitted to the respective online competitions based on the CodaLab platform with a public leaderboard.
Final Project	The students will be asked to select a topic related to NLP from a list of selected proposed research projects. A supervisor who is working on the related topic in Skoltech will be assigned to a small group of students (up to 5 persons). Alternatively a student is free to select a topic of her choice. The assignment will be similar to the final projects of the Deep Learning course at Skoltech and will be judged according to the similar criterion. Namely, the quality of the final report, presentation, code, results as compared to the baselines, and the other usual criteria which emulate a standard NLP research project evaluation.

Type of Assessment: Graded

Grade Structure

Activity Type	Activity weight, %
Homework Assignments	60
Final Project	40

A: 80

B: 70

C:	60
D:	50
E:	40
F:	0
Attendance Requirements	Mandatory with Exceptions

Maximum Number of Students

	Maximum Number of Students
Overall:	40
Per Group (for seminars and labs):	

Course Term (in context of Academic Year)	Term 2
Course Delivery Frequency	Every year

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

Masters Programs	PhD Programs
Data Science	Computational and Data Science and Engineering

Course Tags	Math Programming Statistics

Required Textbooks	ISBN-13 (or ISBN-10)
-	

Recommended Textbooks	ISBN-13 (or ISBN-10)
Jurafsky, D. and Martin, J. H. (2009): Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Second Edition. Pearson: New Jersey	0130950696
Manning, C. and Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.	9780262133609
Manning, C. D., Raghavan, P., and Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.	9780521865715

Equipment
Access to a computational server with GPUs (e.g. Nvidia 2080 Ti or similar) may be useful but the Google CodaLab platform should be normally sufficient.

Software
Python 3

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.

Select Assignment 1 Type	Homework Assignments
Input Example(s) of Assignment 1 (preferable)	Lexical semantics. In this assignment, the student will be able to select among two task where distributional models, such as word2vec, are to be used and possibly improved: word sense induction and taxonomy enrichment.
Assessment Criteria for Assignment 1	The solutions shall be submitted to the respective online competitions based on the CodaLab platform with a public leaderboard.
Select Assignment 2 Type	Homework Assignments
Input Example(s) of Assignment 2 (preferable)	Named Entity Recognition (NER). In this assignment NER a related sequence tagging task will be addressed. The students will be asked to develop an NLP for solving the NER or a similar task on their own. The result will be a report in the structure of a paper and the code of the solution.
Assessment Criteria for Assignment 2	Quality of the obtained results and innovativeness of the used model will be judged. The solutions shall be submitted to the respective online competitions based on the CodaLab platform with a public leaderboard.
Select Assignment 3 Type	Final Project

Input Example(s) of Assignment 3 (preferable)	Using graph convolutional networks for taxonomy enrichment
Assessment Criteria for Assignment 3	Technical report, Presentation, Code, Novelty, Comparisons to the baselines