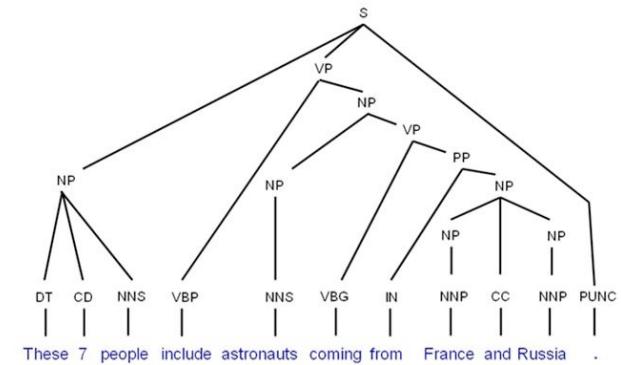
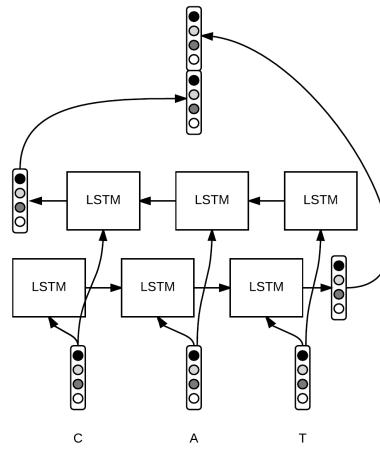
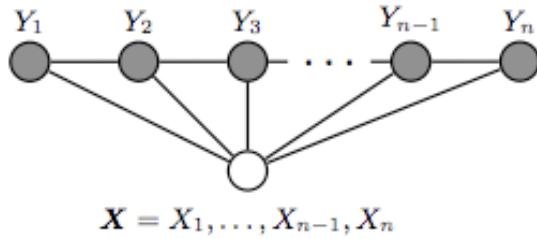
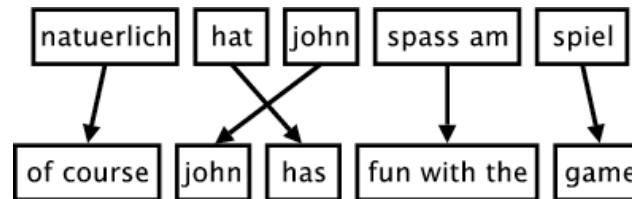
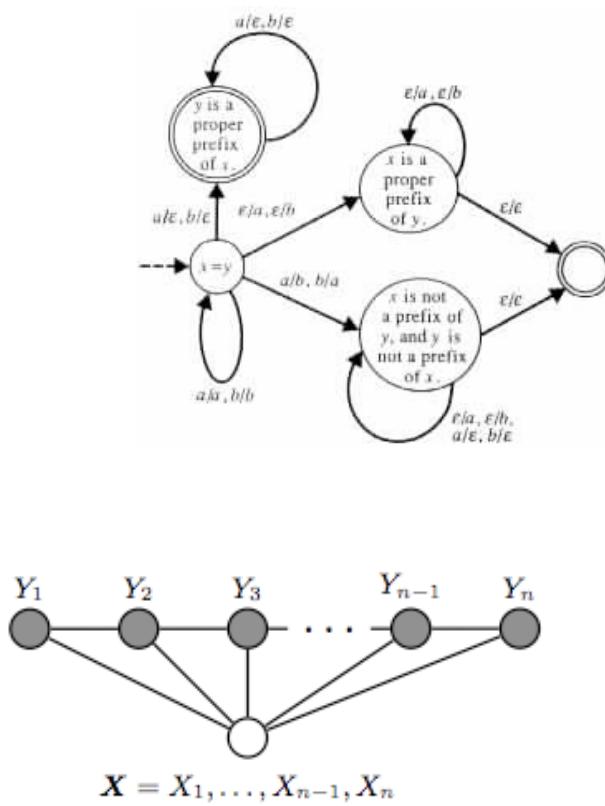


# Introduction to Natural Language Processing



# Teaching team

**Alexander Panchenko**, Associate Professor @ Skoltech AI

- Head of the NLP group at Skoltech and the joint laboratory of MTS- Skoltech in the field of AI
  - <https://faculty.skoltech.ru/people/alexanderpanchenko>
- **PhD** in NLP, University of Louvain, Belgium
- **Postdoc**, University of Hamburg, TU Darmstadt, Germany
- Author of many visible works in the field of NLP at leading conferences in this field (ACL, NAACL, EMNLP, ...)
- **NLP lab** at Skoltech yielded a substantial amount of NLP research publication since its creation in April 2019:
  - <https://sites.skoltech.ru/nlp>

# The team of course instructors

- Prof. Dr. Alexander Panchenko (Skoltech)
- Dr. Irina Nikishina (U Hamburg, Skoltech graduate)
- Anton Razzhigaev (Skoltech/AIRI, PhD student)
- Artem Vazhentsev (Skoltech/AIRI, PhD student)
- Daniil Moskovskii (Skoltech, PhD student)
- Daniil Larionov (Skoltech/U Mannheim, PhD student)
- Elisey Rykov (Skoltech, PhD student)
- Andrey Sakhovskiy (Skoltech/Sber AI, PhD student)
- Maria Lysyuk (Skoltech, PhD student )



# Acknowledgements

---

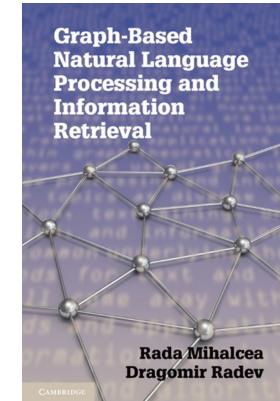
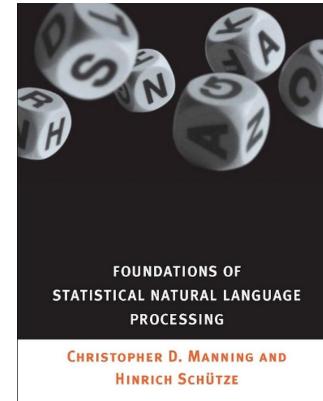
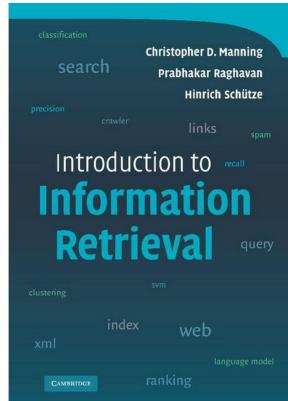
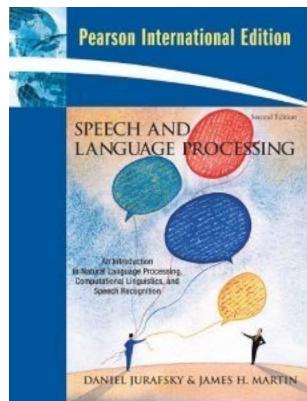
This course is based and relies on multiple sources, inter alia:

- Course of Chris Biemann (University of Hamburg)
- Lecture of David Dale (Meta)
- Course of Daniel Jurafsky (Stanford University)
- Tutorials of Dmitry Ustalov (JetBrains)
- Course of Jury Leskovec (Stanford University)

We try to provide a specific link where possible.

# Textbooks

- Jurafsky, D. and Martin, J. H. (2023): **Speech and Language Processing. An Introduction to Natural Language Processing.** Third Edition (in progress): <https://web.stanford.edu/~jurafsky/slp3/>: **The main reference!**
- Manning, Christopher D., Hinrich Schütze, and Prabhakar Raghavan (2008). **Introduction to information retrieval.** Cambridge university press.
- Manning, C. D. and Schütze, H. (1999): **Foundations of Statistical Natural Language Processing.** MIT Press: Cambridge, Massachusetts
- Mihalcea, Rada, and Dragomir Radev (2011): **Graph-based Natural Language Processing and Information Retrieval.** Cambridge university press.

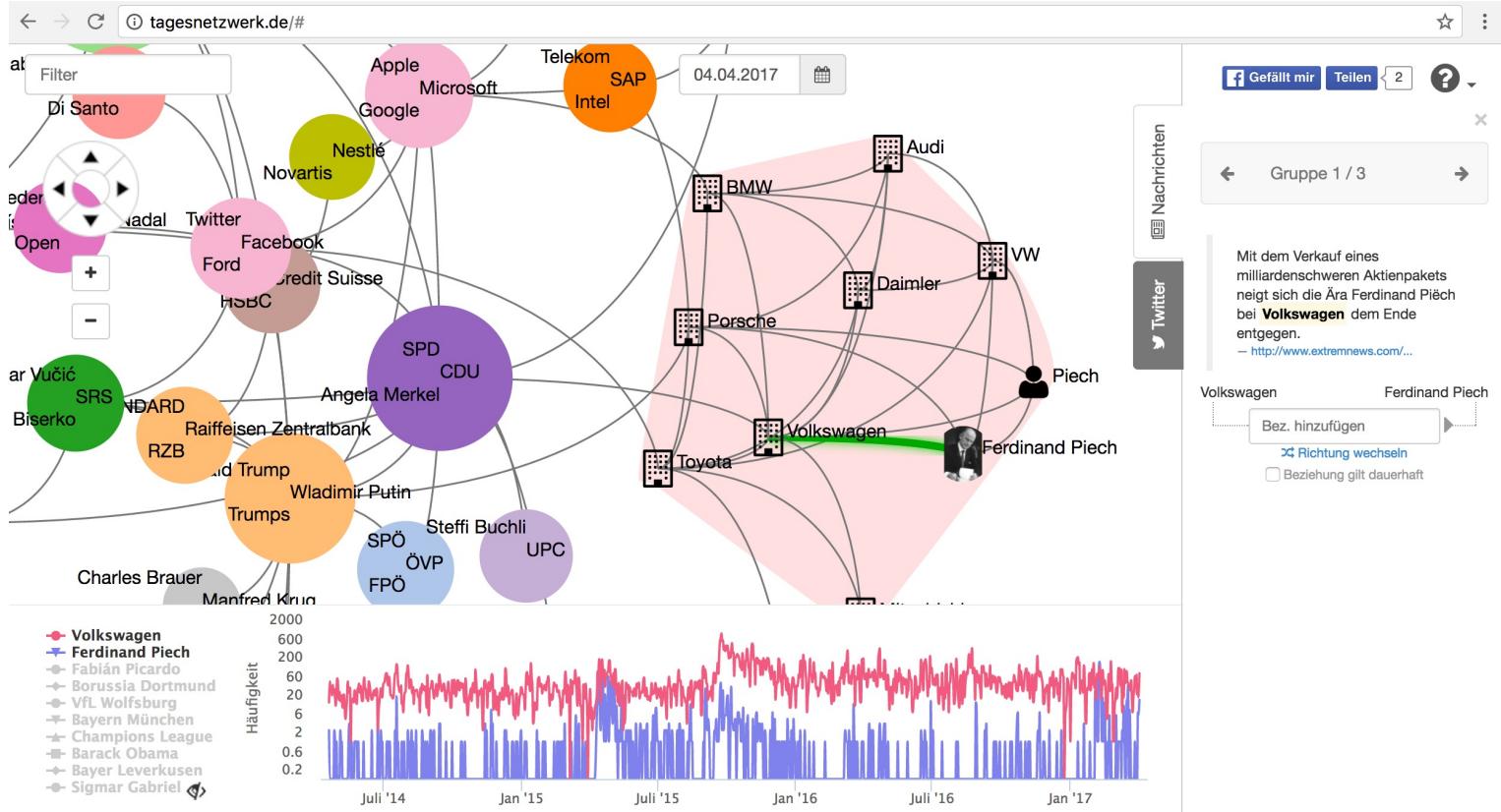


# Learning Goals

---

- understand methods for language processing
- 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 master thesis in language tech**

# Network of the day



- Student project, up since 2014.
- [www.tagesnetzwerk.de](http://www.tagesnetzwerk.de)

# Comparative Argumentative Machine



Wow, Python much faster than MatLab .

RE: Wow, Python much faster than MatLab .

Remember that Python with NumPy tend to be faster than Matlab.

Python might be faster Click, to show context. I'm not good at MATLAB so I don't know how to get computational times (or in Python, for that matter).

As you can see from the results- Matlab is significantly faster than python.

Right, exactly; but "flat" Matlab (that is, Matlab with few looping constructs) has been shown to me to be faster than Python+NumPy for intensive calculations.

But I also tested with 64 bit float maxtrix and on my machine, Matlab 2010b is still faster than Python 3.2 with Numpy-MKL

- MA project, a paper at SIGIR CHIIR'19 conference in UK
- <http://ltdemos.informatik.uni-hamburg.de/cam/>
- <https://arxiv.org/abs/1901.05041>

# Word sense disambiguation

Sentence

Jaguar is a large spotted predator of tropical America similar to the leopard.

(A)

Model

Word Senses based on Cluster Word Features

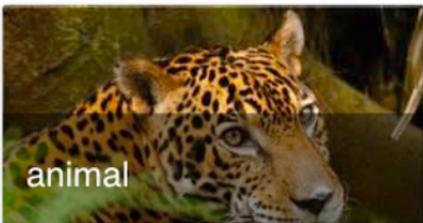
(C)

[DISAMBIGUATE SENTENCE](#)

[RANDOM SAMPLE](#)

## Detected Entities

The system has detected these entities in the given sentence.



Jaguar

(D)



is a large spotted

predator

(D)



America

(D)

- MA project, a paper at EMNLP'17 conference in Denmark
- <http://ltbev.informatik.uni-hamburg.de/wsd/>
- <https://aclweb.org/anthology/papers/D/D17/D17-2016/>

# Text Detoxification

Dale, D., Voronov, A., Dementieva, D., Logacheva, V., Kozlova, O., Semenov, N. and Panchenko, A. (2021): [Text Detoxification using Large Pre-trained Neural Models](#). In Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP-2021). Punta Cana, Dominican Republic. ([Master project of Anton Voronov](#))

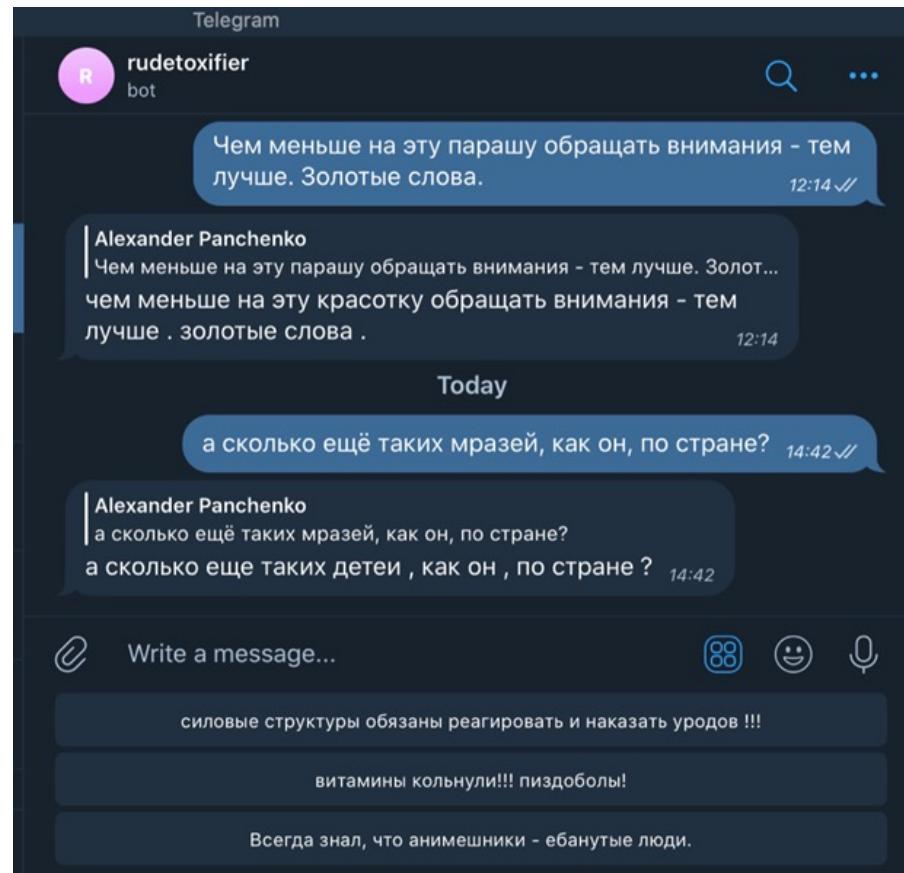
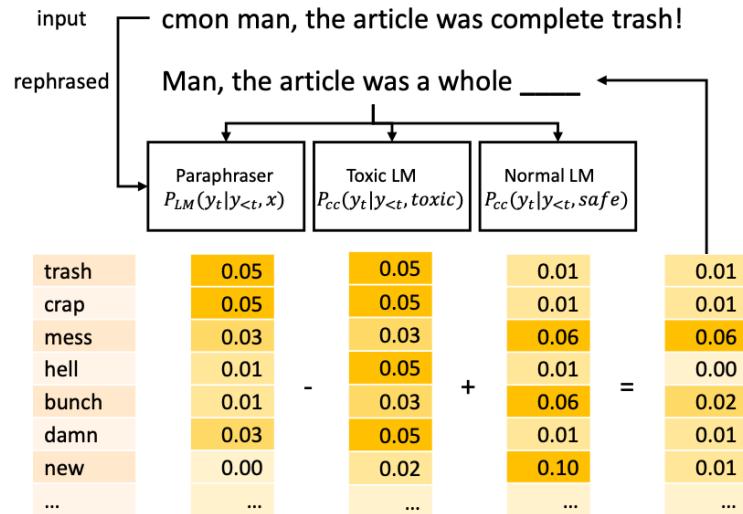


Figure 1: The overview of ParaGeDi model.

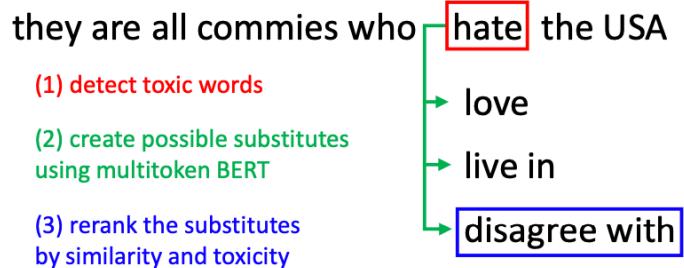
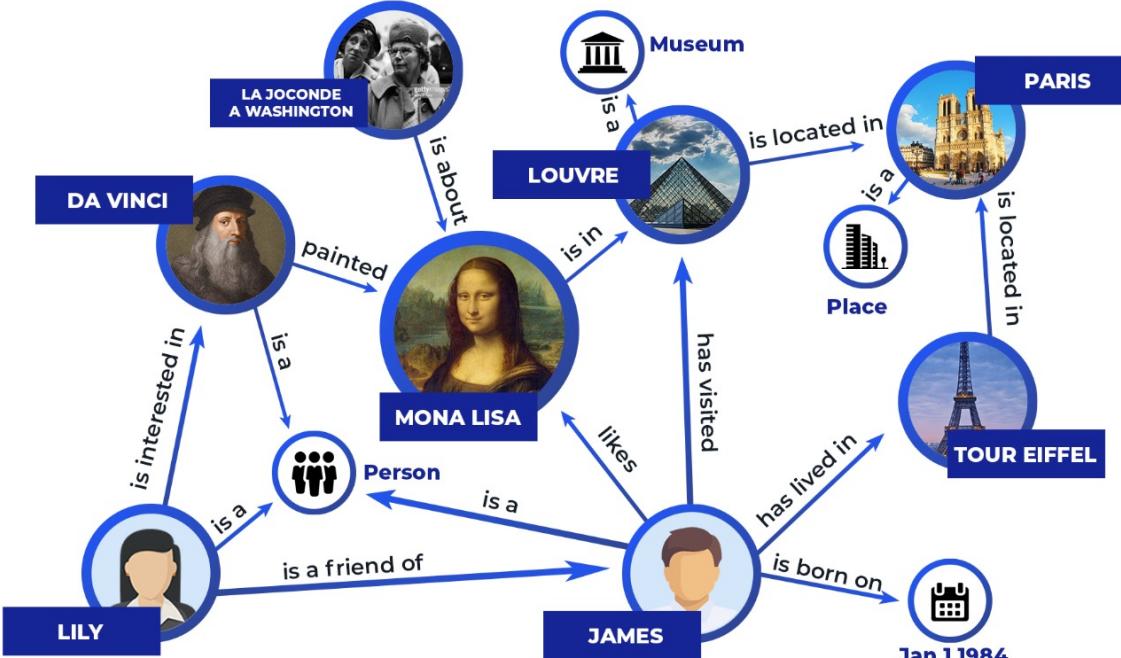


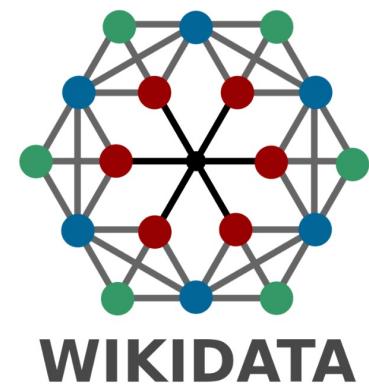
Figure 2: The overview of the CondBERT model.

# Knowledge Graph Question Answering



## Verifiable QA:

- Reliable => Knowledge Graphs (KG)
- Relevant => WikiData
- Precise => Uncertainty Estimation (UE)



# Knowledge Graph Question Answering

Anton Razzhigaev, Mikhail Salnikov, Valentin Malykh, Pavel Braslavski, and Alexander Panchenko. 2023. [A System for Answering Simple Questions in Multiple Languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 524–537, Toronto, Canada. Association for Computational Linguistics.

22 Назад WikiData QA system ... печатает

Кто сражался с немейским львом?  
18:17 ✓

Heracles, 0.4418  
Hydra, 0.3959  
Orthrus, 0.3957  
Chimera, 0.387  
Labours of Hercules, 0.3079

[www.wikidata.org](http://www.wikidata.org)  
Heracles  
divine hero in Greek mythology,  
son of Zeus and Alcmene

18:18

В каком архитектурном стиле  
построено здание Большого театра в  
Москве?  
18:27 ✓

Art Nouveau, 0.0325  
neoclassicism, 0.032  
Neoclassical architecture, 0.0293  
constructivist architecture, 0.0283  
Victor Louis, 0.0197

[www.wikidata.org](http://www.wikidata.org)  
Art Nouveau  
international philosophy and  
style of art, architecture and  
applied art

18:29

Кто написал картину «Мона Лиза»?  
03:43 ✓

Leonardo da Vinci, 0.0541  
Neil Jordan, 0.0447  
Michael Kamen, 0.0446  
George Harrison, 0.0446  
Mona Lisa, 0.0445

[www.wikidata.org](http://www.wikidata.org)  
Leonardo da Vinci  
Italian Renaissance polymath  
(1452-1519)

03:44

Что лучше кофе или чай?  
18:02 ✓

Sorry, uncertainty is too high.  
18:02

Где живёт Дональд Трамп?  
16:32 ✓

Manhattan, 0.0438  
Greencastle, 0.0436  
Jamaica Hospital, 0.0434  
Falls Church, 0.043  
Hudson, 0.0419

[www.wikidata.org](http://www.wikidata.org)  
Manhattan  
borough of New York City, New  
York, United States

16:35

If the precision control mechanism decides that  
uncertainty is too high, the bot won't return the answer

10

# Course Schedule

Week	Date	Time	Type of Activity	Topic	Lecture	Seminar
1	31.10.2023					
	02.11.2023					
2	07.11.2023	16-19	Lecture + Seminar	Introduction and Morphological Analysis	Alexander Panchenko	Sergey Petrakov
	10.11.2023	16-19	Lecture + Seminar	Introduction and Morphological Analysis (cont'd)	Alexander Panchenko	Sergey Petrakov
3	14.11.2023	16-19	Lecture + Seminar	Distributional Semantics and Word Sense Disambiguation	Alexander Panchenko	Daniil Moskovskiy
	16.11.2023	16-19	Lecture + Seminar	Distributional Semantics and Word Sense Disambiguation (cont'd)	Alexander Panchenko	Daniil Moskovskiy
	17.11.2023	EoD	Assignment 1: Release	Word Sense Induction	Irina Nikishina	
4	21.11.2023	16-19	Lecture + Seminar	Sequence Labelling	Alexander Panchenko	Artem Vazhentsev
	23.11.2023	16-19	Lecture + Seminar	Sequence Labelling (cont'd)	Alexander Panchenko	Artem Vazhentsev
5	28.11.2023	16-19	Lecture	Language Models and Machine Translation	Alexander Panchenko	Elisey Rykov
	30.11.2023	16-19	Invited lecture	Invited lecture: Multilingual GPT (Sber Devices)	Maria Tikhonova	--
	1.12.2023	EoD	Assignment 1: Deadline	Word Sense Induction	Everybody	
	1.12.2023	EoD	Assignment 2: Release	Semantic Role Labelling	Irina Nikishina	
6	05.12.2023	16-19	Seminar	Language Models and Machine Translation (cont'd)	--	Elisey Rykov
	07.12.2023	16-19	Lecture + Seminar	Graphs for NLP: Clustering	Alexander Panchenko	Anton Razzhigaev (online)
7	12.12.2023	16-19	Lecture + Seminar	Graphs for NLP: Vector Representations	Alexander Panchenko	Andrey Sakhovskiy
	14.12.2023	16-19	Lecture + Seminar	Graphs for NLP: Vector Representations (cont'd)	Alexander Panchenko	Andrey Sakhovskiy
	15.12.2023	EoD	Assignment 2: Deadline	Semantic Role Labelling	Everybody	
	15.12.2023	EoD	Assignment 3: Release	Taxonomy Enrichment	Irina Nikishina	
8	19.12.2023	16-19	Lecture + Seminar	Syntactic Parsing	Alexander Panchenko	Daniil Larionov
	21.12.2023	16-19	Lecture + Seminar	Data Annotation and Crowdsourcing for NLP	Alexander Panchenko	Maria Lysyuk
	21.12.2023	16-19	Assignment 3: Deadline	Taxonomy Enrichment	Everybody	

# Course Schedule

time\day	TUESDAY	WEDNESDAY	THURSDAY
11-12			
12-13			
13-14			
14-15			
15-16			
16-17			
17-18	16:00 – 19:00 Lecture + Seminar		16:00 – 19:00 Lecture + Seminar
18-19			

# Assessment

- Three individual assignments 33.33% each (worth 100%)
  - We are going score by report + code (Jupyter) and results (Codalab)
  - **75 points** for 100% of the course

Information		Technical report		Code		Results		<b>Penalty:</b> late submission 1 day = 1 point	Total
Student	Codalab notebook	Methodology	Discussion of results	Readability	Reproducibility	Improved over the baseline	Bonus: top-1 - 10 points, top-20% - 5 points		
100 %									
Ivan Petrov	<a href="http://colab.com">http://colab.com</a>	5	5	5	5	5	0 or 5 or 10	0 to 25	25 + Bonus - Penalty

- You can also gain **extra points** for
  - Top results on the shared tasks within assignments (see above)
  - Top results on quiz questions (see below)
- Quizzes:
  - For lectures (2 points for each):  $7 * 2 = 14$
  - For seminars (4 points for each):  $7 * 4 = 28$
  - In summary: you can learn additional points with quizzes
- No exam

# Code and technical report are submitted as a single Jupyter notebook (following a pre-defined format)

The screenshot shows a Jupyter Notebook interface with two panes. The left pane is titled "ASSIGNMENT-TEMPLATE.IPYNB" and contains a table of contents:

1. Information about the submission
1.1 Name and number of the assignment
1.2 Student name
1.3 Codalab user ID
1.4 Additional comments
2. Technical Report
2.1 Methodology
2.2 Discussion of results
3. Code
3.1 Requirements
3.2 Download the data
3.3 Preprocessing

The right pane is titled "assignment-template.ipynb" and shows the content of the notebook:

## 2. Technical Report

Use Section 2 to describe results of your experiments as you would do writing a paper about your results. DO NOT insert code in this part. Only insert plots and tables summarizing results as needed. Use formulas if needed to describe your methodology. The code is provided in Section 3.

### 2.1 Methodology

Enter here a detailed description of the method used in your submission(s) to Codalab. The description should be at least 2-4 paragraphs featuring the following: type of the model, meta-parameters, how did you select meta-parameters, any further modifications of the out-of-the-box solutions, etc. The text is markdown and you can use math environment to write formulas:

$$\hat{y} = \beta_0 + \sum_{j=1}^n x_j \beta_j$$

Also you can insert images as needed:

This part of the should contain description of all methods that you tried and, most importantly, that worked the best for you. Here you can include some tricks of your preprocessing, description of the models and motivation of their usage, the description of the training process details (train-test split, cross-validation, etc.). So, everything valuable that will help us to understand the scope of your work

# Assignments are Shared Tasks

- A Kaggle-style competition for the best score

CodaLab

RESULTS

#	User	Entries	Date of Last Entry	Team Name	MAP ▲	MRR ▲
1	RefalMachine2	334	08/28/20		0.5858	0.6241
2	vvyadrincev	25	03/17/20		0.5163	0.5506
3	mkunilovskaya	29	03/07/20	KuKuPl	0.4901	0.5301
4	aleksei_pronkin	9	11/23/20	Aleksei_Pronkin	0.4218	0.4521
5	vadim_bz	8	12/20/20		0.4210	0.4518
6	lilaspourpre	3	03/04/20	baseline (fasttext bugfixed)	0.4210	0.4518
7	iامdenay	7	03/12/20		0.4208	0.4512
8	Igor_Karpikov	2	11/23/20		0.0928	0.1019



# Focus of the course and relation to the Neural NLP course

- **Introduction to NLP (~ NLP part I):**
  - Term 2 (November-December)
  - 3 credits
  - Grouping by tasks
  - Wider scope of various approaches
- **Deep Learning for NLP (~ NLP part II)**
  - Term 4 (April-May)
  - 3 credits
  - Grouping by model.
  - More advanced and only neural NLP models

