UNIVERSITY OF WARSAW

UNIVERSITY'S INTEGRATED DEVELOPMENT PROGRAMME

Course syllabus

Course title	**Introduction to natural language processing**
Instructor(s)	Adam Zadrożny PhD
Contact details	adam.zadrozny.students.cogni@gmail.com
Affiliation	National Centre for Nuclear Research
Course format	lecture + class
Number of hours	30 hours + 30 hours
Number of ECTS credits	**5 ECTS credits**
Brief course description	The course is focused on natural language processing. It will cover both theoretical and practical aspects of natural language processing, but will put more emphasis on practicals skill that could be used in scientific and technological projects.
Full course description	The course is focused on natural language processing. It will cover both theoretical and practical aspects of natural language processing, but will put more emphasis on practicals skill that could be used in scientific and technological projects. During classes current state-of-art models will be used more than well established packages like NLTK. There will be a three ways of passing this course: 1/ large project 2/ exam 3/ assignments and exam. Since NLP is very fast developing field it is important to note that there might be slight changes in course material to reflect latest developments.
Learning outcomes	Course enables student to: - understand the mechanisms and applications of most commonly used methods in natural language processing for cognitive research (K_W01, K_W02) - have a practical knowledge on natural language processing (K_U03, K_U04) - know the limits and advantages of each method (K_K02) - knows the limits of current knowledge in the field (K_K01)
Learning activities and teaching methods	Lectures and exercises will be closely related. More focus will be given to technical tasks and exercises that give a possibility to extract information. For example, if during the lecture a task on text summarisation is presented, students might be asked to write a simple programme on text summarisation. More focus will be given to state-of-the-art methods like for BERT, ELMo or Flair than historical ones as Glove. During the course

	basic programming skills in Python will be very helpful. Data analysis tasks will be done in Jupyter notebook. Programs for data processing will be written in Python 3 language. To complete a course student will spend: - 30 hours attending lectures - 30 hours attending exercises - 90 hours doing assignments and reading material
List of topics/classes and bibliography	1. Introduction to natural language processing (NLP). Historical background. Basic terms. Mathematical background. Examples of contemporary nap applications. (1 lecture) 2. Basic text processing: tokenising, stemming, lemmatisation. Regular expressions. (1 lecture) 3. Text Corpuses. N-gram models. (1 lecture) 4. Bag-of-words, continuous bag-of-words, Naive Bayes (1 lecture) 5. Word vectorisation. Glove and Word2Vec. (1 lecture) 6. Contextual word embeddings BERT, ELMo, Flair. (1 lecture) 7. Estimating next word or letter and where it is leading us. (1 lecture) 8. Sentiment recognition (1 lecture) 9. Named Entity Recognition (NER). Part of speech tagging. (1 lecture) 10. Machine learning translation (1 lecture) 11. Automatic text generation and GPT-2 model (1 lecture) 12. Text summarization (1 lecture) 13. Example of use cases of NLP. Information extraction, question answering, bots (1 lecture) 14. State-of-the art methods of NLP processing. The last two lectures will be address methods published in the second half of 2019 and the most used current methods. (2 lectures) Bibliography: Highly recommended: - Delip R., Natural Language Processing with PyTorch, O'Reilly, 2019 - Bird S., Klein E., Loper E., Natural Language Processing with Python, O'Reilly, 2009

	- Shaw Z., Learn Python 3 the Hard Way, Addison-Wesley, 2017
	- Kaiser Ł., Deep Learning: The Good, the Bad and the Ugly, PhDOpen 2018/2019, [at least first lecture] http://phdopen.mimuw.edu.pl/index.php?page=l18w5
	Recommended:
	- OpenAI (and related arxiv papers)
	  - GPT-2 https://openai.com/blog/better-language-models/
	  - GPT-2 https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf
	  - Unsupervised Sentiment Neuron https://openai.com/blog/unsupervised-sentiment-neuron
	- BERT https://arxiv.org/abs/1810.04805
	- Flair https://github.com/zalandoresearch/flair
	Optional:
	- Articles from Arxiv.org from cs.CL section

Assessment methods and criteria	There will be three paths to pass the course: 1) prepare a project on NLP 0-100% 2) exam 0-100% 3) assignments (40%) and exam (60%) Since this course has a practical application I will strongly encourage students to try with the project. Examples of projects will be given at the beginning of the course.
Attendance rules	2 unexcused absences are allowed. More than 2 absences might result in penalty points during grading.
Prerequisites	Prior to the course the student shall: - know how to write programmes in python - have basic knowledge on neural networks - have some linguistic knowledge on how language is structured
Academic honesty	Students must respect the principles of academic integrity. Cheating and plagiarism (including copying work from other students, internet or other sources) are serious violations that are punishable and instructors are required to report all cases to the administration.
Remarks	The main goal is to get practical experience that could be used for solving research and real life problems.