**Nazwa przedmiotu:**

Natural Language Processing

**Koordynator przedmiotu:**

Dr inż. Anna Wróblewska

**Status przedmiotu:**

Obowiązkowy

**Poziom kształcenia:**

Studia II stopnia

**Program:**

Data Science

**Grupa przedmiotów:**

Współny

**Kod przedmiotu:**

.

**Semestr nominalny:**

3 / rok ak. 2022/2023

**Liczba punktów ECTS:**

6

**Liczba godzin pracy studenta związanych z osiągnięciem efektów uczenia się:**

1. godziny kontaktowe – 80 h; w tym a) obecność na wykładach – 15 h b) obecność na zajęciach projektowych – 45 h c) obecność na ćwiczeniach – 15 h d) konsultacje – 5 h 2. praca własna studenta – 75 h; w tym a) przygotowanie do zajęć projektowych – 20 h b) przygotowanie do ćwiczeń – 15 h c) przygotowanie raportu i prezentacji projektu – 20 h d) zapoznanie się z literaturą – 20 h Razem 158 h, co odpowiada 6 pkt. ECTS

**Liczba punktów ECTS na zajęciach wymagających bezpośredniego udziału nauczycieli akademickich:**

a) obecność na wykładach – 15 h
b) obecność na zajęciach projektowych – 45 h
c) obecność na ćwiczeniach – 15 h
d) konsultacje – 5 h
Razem 80 h, co odpowiada 3 pkt. ECTS

**Język prowadzenia zajęć:**

angielski

**Liczba punktów ECTS, którą student uzyskuje w ramach zajęć o charakterze praktycznym:**

a) przygotowanie do zajęć projektowych i ćwiczeń – 20 h
b) obecność na zajęciach projektowych – 45 h
c) przygotowanie raportu i prezentacji projektu – 20 h
Razem 85 h, co odpowiada 3 pkt. ECTS

**Formy zajęć i ich wymiar w semestrze:**

|  |  |
| --- | --- |
| Wykład: | 15h |
| Ćwiczenia: | 15h |
| Laboratorium: | 0h |
| Projekt: | 45h |
| Lekcje komputerowe: | 0h |

**Wymagania wstępne:**

Introduction to Machine Learning, Programming in R and Python

**Limit liczby studentów:**

Number of groups: no limits Tutorial – the number of students in a group matches the limits defined by the Warsaw University of Technology Project classes – the number of students in a group matches t

**Cel przedmiotu:**

The aim of the course is to present the basic theory, models and methods of processing, analysis and knowledge discovery in various text data sets. During the course the students will learn about basic and advanced techniques of information systems and solutions for natural language processing by deep learning methods.
As part of the course, students should acquire the practical ability to use selected methods to model and analyse textual data.

**Treści kształcenia:**

Lecture: During the course the students will learn about the basic and advanced techniques of information systems: information extraction, text indexing, measuring the quality of search and exploration systems, the architecture of modern knowledge search systems and information management. Contemporary natural language processing solutions using deep learning methods will also be provided, including the latest architectures, i.e. recursive networks, word representation models - word embeddings.
The lecture and projects will include selected topics from the list below:
1. Introduction to information retrieval: information theory, NLP / text mining methods, statistics, linguistics, issues in the text processing process: lemmatization, stemming
2. Words and sentences: regular expressions, tokenization, editorial distance, linguistics: levels of language description, morphology: word formation, inflection, stamps, parts of speech, soundex algorithm, spelling errors
3. Data sources: text corps
4. Models of text representation, incl. vector models (word embeddings) - word2vec, gloVe, FastText, ELMO, FLAIR, BERT
5. Architecture of deep neural models - recursive, recursive, convolutional and dynamic networks for text recognition applications: language modeling, opinion analysis, text parsing, sentence classification
6. Statistical methods of natural language processing, language modeling, n-grams, collocations, disambiguation (word sense disambiguation)
7. Grammar analysis (POS tagging, parsing)
8. Information extraction, NER (named-entity recognition), relationship extraction, information semantics (ontologies, construction of ontologies from a text)
9. Information search: inverted index, similarity measures, result ranking, link analysis (PageRank, HITS), commercial system architectures, measuring the quality of returned results, visualization of search results, architecture of commercial information search systems / knowledge bases
10. Applications:
a. Document categorization and grouping (hierarchical grouping, LDA - latent dirichlet allocation)
b. Analysis of the emotional color of the text (sentiment analysis)
c. Answering queries (question answering)
d. chatbots
e. Summarizing documents
f. Automatic translations
g. Content-based recommendations
Exercises and project classes:
Development of methods and applications in the field of NLP, the applications will be aimed at processing text and data from websites or text corpora.

**Metody oceny:**

Completion of the course is based on the results of the tasks carried out under the projects. For tasks carried out in the laboratory, both the quality of solutions and the timeliness of their implementation are assessed. Detailed rules for awarding points are presented to students during the first project classes (regarding passing the project).
The final grade depends on the total number of points obtained from projects and their presentations and is determined in accordance with the following rules: 0-50 points - 2.0, 51-60 points - 3.0, 61-70 points - 3.5, 71-80 points - 4.0, 81- 90 points - 4.5, 91-100 points - 5.0. A prerequisite for obtaining a positive grade is obtaining at least 50 points.

**Egzamin:**

nie

**Literatura:**

1. D. Jurafsky, J. Martin, Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, Prentice Hall Series in Artificial Intelligence, Pearson/Prentice Hall 2009
2. H. Schutze, C. Manning, P. Raghavan, Introduction to information retrieval, 2008 http://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf
3. J. Hirschberg, C. Manning, Advances in natural language processing”, Science (New York, N.Y.), 17 July 2015, Vol.349(6245), pp.261-6
4. M. Kłopotek, Inteligentne wyszukiwarki internetowe, Warszawa: Akademicka Oficyna Wydawnicza EXIT, 2001
5. Python NLTK (Natural Language Toolkit) http://nltk.sourceforge.net
6. Open NLP http://opennlp.sourceforge.net/
7. Deep Learning (Adaptive Computation and Machine Learning series) – November 18, 2016, Ian Goodfellow, Yoshua Bengio , Aaron Courville

**Witryna www przedmiotu:**

.

**Uwagi:**

.

## Charakterystyki przedmiotowe

### Profil ogólnoakademicki - wiedza

**Charakterystyka W01:**

Student knows the theoretical basis of knowledge discovery methods in text data sets (e.g. text data mining methods), including the use of machine learning methods and deep learning networks

Weryfikacja:

Grades of projects and their presentations

**Powiązane charakterystyki kierunkowe:** DS2\_W03, DS2\_W04, DS2\_W06

**Powiązane charakterystyki obszarowe:**

**Charakterystyka W02:**

Student knows the basic methods of representing non-structural textual data and the way they are used in combination with machine learning methods and deep learning networks

Weryfikacja:

Grades of projects and their presentations

**Powiązane charakterystyki kierunkowe:** DS2\_W03, DS2\_W06

**Powiązane charakterystyki obszarowe:**

### Profil ogólnoakademicki - umiejętności

**Charakterystyka U01:**

Student is able to design and evaluate algorithms solving a specific problem of data analysis of practical importance using inter alia classification methods and deep learning networks

Weryfikacja:

Grades of projects and their presentations, Grades of projects

**Powiązane charakterystyki kierunkowe:** DS2\_U01, DS2\_U04, DS2\_U05, DS2\_U14

**Powiązane charakterystyki obszarowe:**

**Charakterystyka U02:**

Student can select the appropriate programming tools to design an algorithm for the classification of text data, including the tools suitable for the use of deep learning networks

Weryfikacja:

Grades of projects

**Powiązane charakterystyki kierunkowe:** DS2\_U01, DS2\_U04, DS2\_U05

**Powiązane charakterystyki obszarowe:**

**Charakterystyka U03:**

Student has the ability to use literature and online resources in English related to the task being solved

Weryfikacja:

Grades of projects

**Powiązane charakterystyki kierunkowe:** DS2\_U15

**Powiązane charakterystyki obszarowe:**

### Profil ogólnoakademicki - kompetencje społeczne

**Charakterystyka K01:**

Student can work individually and, in a team, and can manage a small team

Weryfikacja:

Grades of projects

**Powiązane charakterystyki kierunkowe:** DS2\_K04

**Powiązane charakterystyki obszarowe:**

**Charakterystyka K02:**

Student has the ability to continue education and awareness of the need for self-education within the lifelong learning process

Weryfikacja:

Grades of projects

**Powiązane charakterystyki kierunkowe:** DS2\_K01

**Powiązane charakterystyki obszarowe:**

**Charakterystyka K03:**

Student is aware of the responsibility for jointly implemented tasks within the teamwork

Weryfikacja:

Grades of projects

**Powiązane charakterystyki kierunkowe:** DS2\_K03, DS2\_K04

**Powiązane charakterystyki obszarowe:**