

INTERDYSCYPLINARNE STUDIA DOKTORANCKIE „HUMANISTYKA CYFROWA”

Przedmiot:	ANALIZA DANYCH TEKSTOWYCH (Text mining)
Prowadzący:	mgr Bartłomiej Balcerzak
Forma zajęć:	wykład
Liczba godzin:	15
Forma zaliczenia:	egzamin ustny (zaliczenie z oceną)

Opis przedmiotu:

W toku wykładu studenci poznają podstawowe terminy i pojęcia związane z automatyczną analizą danych tekstowych oraz automatycznym przetwarzaniem języka naturalnego. Zagadnienia te będą przedstawiane zarówno z perspektywy teoretycznej, jak i w formie praktycznych ich realizacji w postaci bibliotek w języku Python, dedykowanych dla zadań text miningowych. W związku z tym przewidywane jest również krótkie wprowadzenie do języka Python, aby studenci mogli w pełni wykorzystać zdobytą wiedzę. Wśród zagadnień poruszonych na wykładzie znajdować się będą zarówno kwestie reprezentacji danych tekstowych (metoda worka słów, metoda wektorów zagnieżdżeń słów t.zw. word embeddings oraz popularne techniki związane z przetwarzaniem języka naturalnego, takie jak analiza sentymentu, klasyfikacja danych tekstowych oraz wiele innych. Tematyka analiz danych tekstowych zostanie poruszona w kontekście języka angielskiego (ze względu na powszechność narzędzi), oraz polskiego (ze względu na zapotrzebowanie kierunku).

Opis przedmiotu (ang.):

Within this lecture, the will learn about the basic tools and techniques related to text mining and natural language processing. These topics will be presented from the theoretical perspective, as well as from a practical one. This will be

achieved by the use of python libraries dedicated to text mining and NLP. Therefore a brief introduction to the Python programming language will be conducted in order for the students to obtain the skills necessary for the full use of text mining techniques. Within the lecture, topics such as word representation (bag of words, and word embeddings approaches) as well as popular nlp techniques will be covered. These include, sentiment analysis, text classification and much more. Due to the needs of students, and software availability, lectures will be conducted in context the English and Polish languages.

Tematy zajęć:

1. Wprowadzenie do języka Python.
2. Metoda bag of words, podstawowa analiza tekstu.
3. Przetwarzanie tekstu, stemming, lematyzacja (w kontekście typów języków naturalnych).
4. Analiza syntaktyczna, analiza części mowy, przetwarzanie tekstu.
5. N-gramy.
6. Modele językowe i sieci neuronowe w text mining.
7. Analiza sentymentu i klasyfikacja tekstu.
8. Word Embeddings, metody Word2Vec, Glove.

Bibliografia:

Literatura będzie podana przez prowadzącego podczas zajęć.

EFEKTY KSZTAŁCENIA

Wiedza:

Student zna podstawy języka Python

Student zna podstawowe rodzaje reprezentacji danych tekstowych.

Student ma wiedzę o najpopularniejszych narzędziach analiz danych tekstowych.

Umiejętności:

Student potrafi napisać prosty skrypt w języku Python

Student potrafi zastosować do tekstu reprezentacje worka słów oraz word embeddings.

Student potrafi przeprowadzić istotne kroki w przetwarzaniu danych tekstowych.

Student potrafi poprawnie dokonać oraz zinterpretować rezultaty klasyfikacji tekstu.

Kompetencje społeczne:

Student rozumie, jakie znaczenie mają technologie komputerowe dla propagacji i analiz tekstów

Student zna potencjał oraz ograniczenia technologii dedykowanej do analizy języka.

Student jest świadom społecznych oraz etycznych konsekwencji stosowania technik text miningowych.