Утвърдил: ……..

Декан

Дата .....

СОФИЙСКИ УНИВЕРСИТЕТ “СВ. КЛИМЕНТ ОХРИДСКИ”

Факултет: Славянски филологии

Специалност: *(код и наименование)*

.....

Магистърска програма: *(код и наименование)*

.....

УЧЕБНА ПРОГРАМА

Дисциплина:

(код и наименование) Езикови данни и модели

Преподавател: доц. д-р Атанас Атанасов

Асистент:

Учебна заетост	Форма	Хорариум
Аудиторна заетост	Лекции	30
	Семинарни упражнения	
	Практически упражнения (хоспетиране)	
Обща аудиторна заетост		**30**
Извънаудиторна заетост	Реферат	
	Доклад/Презентация	
	Научно есе	
	Курсов учебен проект	30
	Учебна екскурзия	
	Самостоятелна работа в библиотека или с ресурси	30
Обща извънаудиторна заетост		**60**
ОБЩА ЗАЕТОСТ		**90**
Кредити аудиторна заетост		**1**
Кредити извънаудиторна заетост		**2**
ОБЩО ЕКСТ		3

№	Формиране на оценката по дисциплината[1]	% от оценката
1.	Workshops {информационно търсене и колективно обсъждане на доклади и реферати)	40
2.	Участие в тематични дискусии в часовете	
3.	Демонстрационни занятия	
4.	Посещения на обекти	
5.	Портфолио	
6.	Тестова проверка	
7.	Решаване на казуси	
8.	Курсов проект	60
9.		
10.		
11.		
12.	Изпит	

Анотация на учебната дисциплина:

Курсът представлява въведение в областта на моделирането на езикови данни. Обхваща общ спектър от теми: от основите на работата с текст и предварителната му подготовка до обработването му и прилагането на техники от машинното самообучение като логистична регресия и градиентно спускане. Включени са практически приложения, като използване на библиотеки за езикова обработка и предварително обучени езикови модели.

Предварителни изисквания:

Завършен поне един граматически курс. Английски език на работно ниво.

Очаквани резултати:

Очаква се студентите да придобият умения за обработване на езикови данни, както и за подготвянето и използването им при обучаване на езикови модели.

Учебно съдържание

№	Тема:	Хорариум
1	Въведение в езиковите бази от данни	1
2	Подготовка на работната среда. Jupyter Notebook	2
3	Библиотеки за работа с данни. Pandas	1
4	Събиране на данни. Формат на езикови данни	2
5	Подреждане и почистване на данни	2
6	Визуализация на езикови данни	2

[1] В зависимост от спецификата на учебната дисциплина и изискванията на преподавателя е възможно да се добавят необходимите форми, или да се премахнат ненужните.

7	Обработка на текст: извличане на информация	2
8	Предварителна обработка на текст за машинно самообучение. Векторизация.	4
9	Статистически модели: N-грами, "Bag of words", TF-IDF	2
10	Логистична регресия. Градиентно спускане	4
11	Работа с колекции от езикови данни. Библиотеките HuggingFace Datasets и TensorFlow Datasets. WordNet. FrameNet. Universal Dependencies	2
12	Работа с предварително обучени езикови модели	4
13	Фина настройка на езикови модели	2

Конспект за изпит

№	Въпрос
1	Изготвяне и защита на курсов проект, демонстриращ познания в областта на моделирането на езикови данни.

Библиография

Основна:

McKinney, W. 2022. Python for Data Analysis. https://wesmckinney.com/book/

XML Tutorial. https://www.w3schools.com/xml/default.asp

Devlin, J., Chang, M. Lee, K., Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805

Mckinney, W. 2011. pandas: a Foundational Python Library for Data Analysis and Statistics. https://www.researchgate.net/publication/265194455_pandas_a_Foundational_Python_Library_for_Data_Analysis_and_Statistics

Jurafsky, D., Martin, J. H. 2024. Speech and Language Processing (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/

Das, M., Selvakumar, K., Alphonse, P.J.A. 2023. A Comparative Study on TF-IDF feature Weighting Method and its Analysis using Unstructured Dataset. https://arxiv.org/abs/2308.04037

Допълнителна:

Дата:
29.03.2024 г.

Съставил:
доц. д-р Атанас Атанасов