Студијски програм: МАС Рачунарске науке

Назив предмета: Обрада природних језика

Наставник/наставници: Милош Савић

Статус предмета: изборни

Број ЕСПБ: 6

Услов: нема

Циљ предмета

Циљ предмета је да студенте упозна са основним концептима, алгоритмима, техникама и моделима рачунарске обраде природних језика.

Исход предмета

Минимални: Успешан студент би минимално требао бити способан да користи програмске библиотеке за обраду природних језика у развоју корисничких апликација.

Пожељни: На крају курса се очекује да студент разуме алгоритме, технике и моделе за обраду природних језика, те да је у стању да их имплементира и евалуира у широком спектру практичних апликација укључив и напредне системе који алгоритмима дубоког машинског учења обрађују огромне количине неструктурираних текстуалних података у природном језику.

Садржај предмета

Теоријска настава

Основни алгоритми за процесирање сировог текста, нормализација текста (стеминг и лематизација), токенизација текста и регуларни изрази, сегментација текста. Бинарна и TF-IDF репрезентација текстуалних докумената. Пробабилистички језички модели. Корекција речи у тексту. Класификација и кластеринг текста коришћењем традиционалних техника машинског учења за табуларне податке. Текст ембединзи и њихово учење *word2vec* фамилијом модела. Класификација и кластеринг текста коришћењем текст ембединга. Проблем анализе сентимента текста. Препознавање именованих ентитета и екстракција информација из текста. Категоризација речи у тексту ПОС (*part-of-speech*) анотаторима. Анализа структуре реченице конктекстно-слободним и депенденцијалним граматикама. Јавно доступни језички ресурси и корпуси. Језички модели засновани на дубоким рекурентним неуронским мрежама. Трансформери и претренирани језички модели. Машинско превођење. Примене техника обраде природног језика у претраживању информација, сумирању текста, аутоматизованим системима за давање одговора, анализи садржаја на социјалним медијима, језичким роботима и виртуелним асистентима.

Практична настава

Упознавање са Python библиотекама за обраду природног језика (NLTK, SpaCy, TextBlob, gensim). Примена библиотека дубоког машинског учења (tensorflow, PyTorch) у обради природног језика. Студије случаја реализоване коришћењем програма који су засновани на претходно поменутим библиотекама.

Литература

1. Dan Jurafsky and James H. Martin. Speech and Language Processing, 3rd edition, 2023, доступно на https://web.stanford.edu/~jurafsky/slp3/

2. Steven Bird, Ewan Klein and Edward Loper. Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit, доступно на https://www.nltk.org/book/

Број часова активне наставе	**Теоријска настава:** 2	**Практична настава:** 2

Методе извођења наставе

Предавања су базирана на класичном моделу извођења наставе који укључује презентовање наставног материјала коришћењем пројектора. На вежбама, које се одвијају у рачунарској лабораторији, студентима се презентују студијски примери из области обраде природног језика. Додатно, студенти на рачунарима решавају постављене проблеме везане за презентоване студијске примере и тако савладавају релевантне програмске библиотеке. Предиспитне обавезе студент полаже реализацијом практичног пројекта у домену обраде природног језика. На усменом делу испита студент треба да демонстрира суштинско познавање концепата, алгоритама, техника и модела који се користе у рачунарској обради природних језика.

Оцена знања (максимални број поена 100)

Предиспитне обавезе	поена	**Завршни испит**	поена
практични пројекат	60	усмени испит	40