Лабораториска вежба 1 - Зборовни вектори и рекурентни невронски мрежи

Задачи

Во оваа лабораториска вежба ќе се применува податочното множество *Yelp*. Множеството е составено од реченици асоцирани со информација дали содржат позитивен или негативен сентимент.

Задача 1 (25 поени)

Користејќи ги реченициите од податочното множество Yelp креирајте зборовни вектори со методот Word2Vec SkipGram. Вклучете ги само зборовите кои се појавуваат најмалку 15 пати. Големината на векторите треба да биде 100. Визуелизирајте ги добиените зборовни вектори во дво-димензионален простор.

Извршете ги следните операции врз зборовните вектори на соодветните зборови:

```
Paris – France + Italy

Madrid – Spain + France

King – Man + Woman

Bigger – Big + Cold

Windows – Microsoft + Google
```

Кој е најблискиот збор до резултантниот вектор?

Како се менуваат резултатите ако наместо методот SkipGram се користи CBOW? Како се менуваат резултатите ако се промени големината на зборовните вектори?

Задача 2 (25 поени)

Креирајте модел на длабоко учење за препознавање дали даден текст содржи позитивен или негативен сентимент базиран на рекурентни невронски мрежи. Влез во моделот треба да биде секвенца од индекси на соодветните зборови. Потребно е моделот да содржи Embedding слој за мапирање на зборовите во зборовни вектори и LSTM слој за процесирање на секвенците. Тежините на Embedding слојот иницијализирајте ги со зборовни вектори добиени со Word2Vec. Користете ги само зборовите кои се појавуваат најмалку 15 пати (останатите зборови заменете ги со специјален „<UNK>“ токен за непознат збор).

Големината на влезните секвенци треба да биде фиксна. Со користење на функцијата **pad_sequences** може да се промени должината на секвенците (како должина може да ја земете средната вредност на должините на текстовите во податочното множество).

```
from keras.preprocessing.sequence import pad_sequences

pad_sequences(sequences, size)
```

Истренираниот модел евалуирајте го со метриките: точност (**accuracy_score**), прецизност (**precision_score**), одзив (**recall_score**) и F1-мерка (**f1_score**).

Споредете ги функциите на загуба на моделот и евалуационите метрики за различни хиперпараметри на моделот (рата на учење, број на епохи, број на слоеви, број на неврони во слоевите, ...). Како се менуваат перформансите на моделот?