Лабораторная работа №6-7

Анализ тональности коротких текстов с использованием Word2Vec и RNN

Цель работы – Научиться представлять текст в виде векторных эмбеддингов (Word2Vec) и использовать рекуррентную нейросеть (RNN/LSTM/GRU) для классификации тональности коротких сообщений.

Задание: обучить модель для решения задачи определения тональности на примере корпуса коротких отзывов о фильмах:

1. использовать датасет коротких текстов отзывов о фильмах из NLTK movie_reviews: `nltk.download('movie_reviews');`
2. очистить текст и токенизировать;
3. обучить Word2Vec и преобразовать слова в векторы (`torchtext` или `gensim`);
4. построить RNN для определения тональности отзыва на положительный/отрицательный;
5. вывести примеры определения тональности для 10 отзывов;
6. оценить точность модели;
7. составить отчет о проделанной работе в соответствии с требованиями кафедры.

Требования к отчету. Отчет должен содержать постановку задачи, исходные данные, результаты решения задачи, необходимые иллюстративные материалы.

Требования к защите

Защита лабораторной работы происходит индивидуально. Система оценки – рейтинговая.

Критерии оценки:

- корректность выполненного исследования;
- адекватность полученных результатов;
- качество отчета;
- качество ответов на контрольные вопросы;

– срок выполнения работы.

Время выполнения работы – 4 академических часа.

Контрольные вопросы

1. Что такое Word2Vec и для чего он используется в NLP?
2. В чем разница между Skip-gram и CBOW моделями Word2Vec?
3. Зачем применяется nn.Embedding? Чем оно отличается от предобученных эмбеддингов Word2Vec?
4. Что такое паддинг и зачем он нужен при работе с RNN?
5. Объясните, что делает слой LSTM и чем он отличается от обычного RNN.
6. Зачем приводить текст к нижнему регистру и удалять пунктуацию перед обучением?
7. Что такое токенизация и какие бывают её виды?
8. Как можно визуализировать предсказания модели и вероятность тональности текста? Приведите примеры.
9. Зачем необходимо фиксировать веса эмбеддингов Word2Vec (requires_grad=False)?
10. Какие метрики позволяют оценить точность нейронной сети на тестовой выборке?