

Вопрос 1

Выполнен

Баллов: 1,00 из 1,00

Ниже приведены утверждения о свойствах и преимуществах модели, обученной на "замороженных" предобученных embedding-векторах слов.

Какие из них вам кажутся верными?

Выберите один или несколько ответов:

- ☒ a. Она предпочтительна в случае, когда у нас мало обучающих данных.
- ☒ b. Есть шанс, что модель сможет правильно классифицировать слова, синонимы которых она видела при обучении.
- ☒ c. Она сохранит взаимное расположение векторов слов исходной языковой модели.
- ☐ d. Она предпочтительна в случае, когда у нас много обучающих данных.

Ваш ответ верный.

Вопрос Инфо

Следующая модификация, которая позволит нам добиться высоких результатов при аспектной классификации — это взвешенная [функция потерь](#). Из приведённой статистики мы видим, что тэг "other" встречается довольно часто (я бы сказал, почти всегда), а тэг "внешний вид" и тэг "безопасность" встречаются довольно редко. Мы хотим, чтобы взвешенная функция потерь сильнее штрафовала нашу нейросеть за ошибки при предсказании этих редких тэгов. Мы снова запустим процесс обучения прямо в браузере и вы сможете увидеть, как уменьшается функция потерь от эпохи к эпохе. Будем обучаться в течение семи эпох.

Наша нейросеть завершила обучение, мы можем проверить её точность на тестовых и тренировочных данных. Для этого мы опять построим [матрицу ошибок](#) и посмотрим на статистику. Определим уже известную вам функцию "predict_tags". Посмотрим на 10 случайных предложений из тестового датасета — как их будет размечать наша нейросеть. На тренировочных данных матрица ошибок выглядит хорошо. Матрица ошибок, полученная на тестовом датасете, выглядит не так хорошо, как на датасете, на котором мы обучались. Это может быть связано с тем, что, например, в целом, наши датасеты довольно небольшие — это 200 текстов, в каждом по 2000 предложений и, возможно, нам нужно было ещё немного модифицировать [функцию потерь](#), чтобы не так агрессивно взвешивать тэги. Более того, наша нейросеть могла просто [переобучиться](#). В домашнем задании мы попытаемся улучшить полученные результаты, а у меня на сегодня всё. Пока!

Вопрос 2

Выполнен

Баллов: 1,00 из 1,00

И, вот, мы запускаем нашу модель на валидационном датасете, где она... досрочно завершается с ошибкой!

Мы поместили в embedding слой из модели fasttext векторы всех слов, которые входят в train и test выборки, но не подумали о том, как мы будем обрабатывать незнакомые слова.

По ссылке ниже вы можете ознакомиться с решением этой проблемы, использующим возможности fastText:

https://github.com/Samsung-IT-Academy/stepik-dl-nlp/blob/master/task7.1_aspect_sentiment_eval.ipynb

А какие еще решения позволят решить проблему с незнакомыми словами?

Выберите один или несколько ответов:

- ☒ а. Положиться на возможности fastText: убрать из модели embedding слой и сделать отдельный конвертер: слово -> вектор fastText.
- ☒ б. Помимо замороженного embedding слоя сделать тренируемый embedding слой, состоящий из вектора <unknown>
- ☒ в. Для незнакомых слов добавить тэг <unknown> и задать для него случайный вектор в embedding слое; Сделать embedding слой обучаемым.

Ваш ответ верный.