Задание на лабораторную работу № 5

по курсу «Методы искусственного интеллекта»

«Обработка текстов на естественном языке»

В рамках лабораторной работы следует ознакомиться с основными операциями по обработке текстов на естественном языке. В качестве основного набора данных следует использовать подмножество набора данных с новостями Lenta.ru (полная версия). За основу рекомендуется взять блокнот из репозитория курса, изменив и доработав его в соответствии с заданием:

1. Лингвистический анализ:

1.1. Разработать функцию, которая бы выделяла из текста упоминания персоналий (людей). Сопоставить множества персоналий, наиболее часто упоминаемых в новостях экономики за 2000 и 2015 годы (по 10 наиболее упоминаемых персоналий).

1.2. Разработать функцию, которая бы выделяла множество действий, совершенных заданной персоналией («X *поручил* то-то, X *предложил* то-то, что-то *было предложено* X»).

2. Векторная модель документа:

2.1. Провести анализ того, как качество тематической классификации новости без лемматизации зависит от размера обучающего множества (см. кривые обучения в зависимости от размера обучающего множества). Сопоставить с качеством классификации модели с лемматизацией.

3. Вложения (эмбеддинги) слов:

3.1. Загрузить модель эмбеддингов слов, обученную на художественной литературе (см. https://github.com/natasha/navec). Для выбранного набора слов сопоставить схожие слова (расположенные рядом в пространстве вложения) и рассуждение по аналогии. Есть ли разница с моделью вложений, обученной на новостных сообщениях?

4. Нейросетевая обработка текстов:

4.1. Тематический классификатор новостей на основе LSTM демонстрирует очень высокое качество классификации буквально с первой эпохи обучения. Но является ли это «заслугой» LSTM или качественного набора эмбеддингов? Реализуйте простейший классификатор сообщений, в котором на вход полносвязному слою передается просто среднее арифметическое эмбеддингов слов. Сопоставьте качество классификации с моделью на основе LSTM. *Примечание: смысл этой идеи в том, что каждое из измерений эмбеддинга (в случае natasha их 300) соответствует некоторому неявному смысловому оттенку слова. Среднее значение по всей последовательности соответствует тому, в какой мере этот смысловой оттенок присутствует во всех словах сообщения (чем в большем количестве слов он присутствует, тем в большей степени он присущ всему сообщению).*

4.2. Проведите анализ ошибок – найдите новости, на которых модель ошибается, и предложите разумные объяснения этому.

4.3. Усовершенствуйте классификатор, чтобы он осуществлял многоклассовую классификацию (классы «Культура», «Спорт», «Экономика»). Подсказка: скорректируйте размерность выходного слоя, используйте torch.nn.CrossEntropyLoss при обучении и torch.nn.Softmax при формировании предсказаний.