Задание 4. FastText

Шаги работы

- Загрузка предобученной на корпусе CommonCrawl модели
- Из предложенных датасетов берем пары слов, для каждого из них строим векторное представление и находим косинус между векторами
- После работы с файлами имеем два больших векторам: один мы получили извлечением человеческих оценок близости, второй составили из получившихся значений косинусов
- Далее нужно было рассчитать корреляцию Спирмена, то есть упорядочить вектор косинусов по убыванию, при этом совершив перестановку и для вектора человеческих оценок так, чтобы у нас не перемешались между собой значения, относящиеся к одной и той же паре слов
- После упорядочивания наибольшему значению косинуса должен назначиться меньший ранг и далее по убыванию значения идет увеличение ранга. Аналогичная процедура проводится и для второго вектора, после чего на основании различности между рангами идет расчет корреляции Спирмена согласно формуле. Библиотека Spicy предоставляет функцию stats.spearmanr(), которая облегчает рассчеты, т.е. проводит все описанные выше действия, в т.ч. и упорядочивание.

Выводы

Полученные значения коэффициента корреляции:

```
Similarity correlation: 0.8344429673824861
```

```
Relatedness correlation: 0.7355818724712521
```

Проанализируем полученные значения. Для датасета Wordsim Similarity значение переходит рубеж в 0.8, что, согласно традиционной интерпретации, говорит о близости человеческой оценки схожести слов в паре и оценки, основанной на векторном представлении этих слов. В целом, схожесть определена довольно однозначно, т.к. включает в себя не так много видов отношений: в основном, это синонимия, гипо- и гиперонимия, насколько можно судить по данному нам датасету, заданию от прошлого семинара и полученным сейчас значениям косинусов.

Заметим, что для датасета Wordsim Relatedness тесноту корреляционной связи уже нельзя интерпретировать как высокую, скорее как умеренную. Вероятно, это можно объяснить тем, что “родственность” слов — понятие гораздо более широкое и включает в себя ту же схожесть; ассоциативную оценку (некоторые слова могут казаться ближе друг к другу, если часто встречаются в одном контексте, т.е. мы начинаем проводить параллель между ними); другие виды отношений, например, меро- и холонимию, с которыми мы работали в прошлый раз. Из-за этого человеческая оценка может сильно различаться с полученным значением косинуса между парой слов. Например, рассмотрим такую пару:

Wordsim Relatedness: `lad wizard: 0.92`

FastText cosine: `lad wizard: 0.3244911730289459`

Человеческая оценка очень низкая: менее 1 при максимуме в 10.00, то есть нам эти два слова не кажутся родственными. При этом косинус угла лежит в (0; 1), что говорит о том, что векторы указывают в одном направлении — найдена достаточно хорошая близость. Такие значительные различия при расчете корреляции Спирмена дадут нам достаточно большой квадрат разности рангов оценки и в конечном итоге повлияют на итоговый коэффициент.