Домашнее задание № 4
Лексическая семантика
Отчет

Задание С:

На основе уже обученной модели Word2Vec (векторного представления слов) для русского языка, взятой, например, из библиотеки Gensim или с сайта с различными векторными моделями для РЯ, провести экспериментальное исследование семантики нескольких (5-9) выбранных слов (достаточно частотных, разных частей речи): найти семантически близкие и характеризующие слова, определить близость пар слов, а также исследовать другие операции, допускаемые моделью.

За дополнительные баллы: рассмотреть несколько разных (2-3) обученных векторных моделей и сравнить результаты в них для выбранных слов.

Отчет: краткая характеристика использованной модели/моделей векторного представления слов и описание проведенных экспериментов, выводы.

Введение

В данном отчете приведены результаты исследования работы с веб-сервисом RusVectōrēs. В соответствии с заданием выбраны частотные слова для исследования. Сервис позволяет находить вектор для заданного слова, а так же близкие слова (ближайшие по косинусному расстоянию). С веб-сайта проекта можно загрузить несколько вариантов предобученных моделей для русского языка для работы, кроме того, можно работать с ними из браузера или подгружать результаты ответов на запросы по API.

RusVectōrēs

Веб-сервис RusVectores предоставляет предобученные на различных корпусах варианты модели типа Word2Vec. Он позволяет как скачать предобученную модель (эмбединги для слов), так и получить ответы на некоторые частые при работе с дистрибутивной семантикой вопросы прямо в браузере.

Для анализа были выбраны слова, обладающие довольно высокой частотностью (на основе анализа частотности RusCorpora). Выбраны слова разных частей речи (в скобках указаны частоты в корпусе, подсчитаны словоформы): человек (154011), говорит (87581), просто (91658), время (247588), первый (48376), значит (45669).

Семантически близкие слова

В приведенных ниже таблицах указаны слова, семантически близкие к выбранным для исследования. Кроме слов указано косинусное расстояние от исследуемого слова до выбранного. В таблице 1 приведены результаты для анализа с помощью модели Word2Vec предобученной на Национальном Корпусе Русского Языка (НКРЯ) и всех текстов русскоязычной Википедии (на 2018 год).

В предобученной модели вычисляется вектор (эмбеддинг) для запрашиваемого слова, и находятся солова, вектора которых ближе всего по косинусному расстоянию к этому слову. В таблице перечислены по 10 таких наиболее близких к заданному слов. При обучении этой модели использовалась разметка по частям речи, поэтому каждое слово комментируется

частью речи. Это важно, поскольку частеричные омонимы могут встречаться в разных контекстах и иметь разное векторное представляение. Однако, для некоторых похожих частей речи (например, NOUN и PRNOUN в корпусе) использование практически не отличается, поэтому они оказываются самыми близкими словами. Понятно, что ниболее близкие слова обычно той же части речи, но среди первых также могут встречаться и другие части речи.

человек	**говорить**	**просто**	время	первый	значит
человек PROPN 0.69	сказать VERB 0.79	легко ADJ 0.72	времени NOUN 0.60	второй ADJ 0.70	значить ADV 0.82
женщина NOUN 0.64	думать VERB 0.71	просто ADV 0.71	пора NOUN 0.60	третий ADJ 0.61	потому ADV 0.75
люди NOUN 0.63	заговаривать VERB 0.69	простой ADV 0.71	тогда ADV 0.56	последний ADJ 0.59	конечно ADV 0.73
мужик NOUN 0.59	рассуждать VERB 0.69	хорошо ADJ 0.64	момент NOUN 0.54	четвертый ADJ 0.53	наверное ADV 0.72
мужчина NOUN 0.54	толковать VERB 0.67	простой ADJ 0.61	однако ADV 0.53	следующий ADJ 0.51	значить VERB 0.72
человкъ PROPN 0.53	твердить VERB 0.67	несложный ADJ 0.60	потому ADV 0.52	очередной ADJ 0.47	наверно ADV 0.72
народ NOUN 0.53	рассказывать VERB 0.67	легкий ADV 0.60	период NOUN 0.52	первый NOUN 0.47	следовательно ADV 0.70
потому ADV 0.53	разговаривать VERB 0.67	сложно ADJ 0.59	теперь ADV 0.51	пятый ADJ 0.46	действительно ADV 0.69
человеческий ADJ 0.51	уверять VERB 0.65	надо ADV 0.56	поэтому ADV 0.49	первый PROPN 0.46	мол ADV

Таблица 1: НКРЯ и Wikipedia

В таблице 2 указаны наиболее близкие слова, полученные при использовании модели, обученной только на НКРЯ. Они немного отличаются, но в целом очень похожи.

человек	**говорить**	**просто**	**время**	**первый**	**значит**
человек PROPN 0.79	сказать VERB 0.79	простой ADV 0.66	времени NOUN 0.56	третий ADJ 0.60	значить ADV 0.76
человеческий ADJ 0.59	рассуждать VERB 0.63	легко ADJ 0.62	пора NOUN 0.55	второй ADJ 0.60	значит VERB 0.61
существо NOUN 0.57	возражать VERB 0.63	просто ADV 0.60	период NOUN 0.54	последний ADJ 0.56	значить VERB 0.55
народ NOUN 0.54	толковать VERB 0.61	хорошо ADJ 0.57	десятилетие NOUN 0.47	следующий ADJ 0.52	надо ADV 0.51
личность NOUN 0.53	заговаривать VERB 0.59	простой ADJ 0.53	момент NOUN 0.47	четвертый ADJ 0.50	равный ADJ 0.51
человечество NOUN 0.53	отвечать VERB 0.59	легкий ADV 0.52	время PROPN 0.45	предыдущий ADJ 0.48	по-честному ADV 0.5
человкъ PROPN 0.50	разговаривать VERB 0.58	понятно ADJ 0.52	эпоха NOUN 0.45	шестой ADJ 0.46	обязательно ADV 0.4
индивидуум NOUN 0.50	повторять VERB 0.57	примитивно ADV 0.51	кратковременный ADJ 0.41	пятый ADJ 0.46	по-твоему ADV 0.47
нравственный ADJ 0.50	выражаться VERB 0.55	сложно ADJ 0.50	промежуток NOUN 0.39	последующий ADJ 0.45	мол ADV 0.47

Таблица 2: НКРЯ

В таблице 3 наиболее близкие перечислены слова при использовании модели Arenum FastText. Эта модель обучена при использовании алгоритма FastText CBOW на основе 3,4,5-грамм. Для обучения использовался корпус, содержащий около 10 миллиардов слов. Из-за величины корпуса в нем встречаются довольно редкие слова, хотя словарь содержит меньше слов, чем первая модель. Из-за присутствия в корпусе опечаток, наиболее частые из них могут оказаться близки к запрашиваемому слову (например, 'разговоритьься'). В самом деле, такие модели могут использоваться для исправления опечаток. Кроме того, корпус для обучения модели не размечен, и части речи не использовались при обучении тем не менее самые близкие слова все равно обычно той же части речи.

человек	**говорить**	**просто**	**время**	**первый**	значит
человеко 0.82	сказать 0.87	просто-напросто 0.88	времянка 0.63	второй 0.73	значить 0.72
сверхчеловек 0.75	говаривать 0.80	напросто 0.80	вовремя 0.54	третий 0.72	наоборот 0.61
богочеловек 0.72	рассуждать 0.79	запросто 0.74	временно 0.51	первое 0.69	ибо 0.59
человечески 0.71	спорить 0.77	непросто 0.70	стремя 0.47	четвертый 0.66	равно 0.59
	говоритьься 0.77	попросту 0.70	период 0.46	предпоследний 0.64	по-вашему 0.56
	вторить 0.77	банально 0.68	временщик 0.46	последний 0.61	поэтому 0.55
	разговоритьься 0.76	по-простому 0.66	момент 0.46	шестой 0.60	все-таки 0.55
	разговаривать 0.76	вообще 0.64	полгода 0.43	пятый 0.58	вообще 0.53
	поговоритьь 0.76	легко 0.61	пора 0.43	восьмой 0.57	неважно 0.53

Таблица 3: Araneum fastText

Расстояния между выбраными словами

При помощи всех перечисленных выше моделей были посчитаны косинусные расстояния между используемыми словами. В целом оказывается, что вектора не отрицательные, это

	человек	**говорит**	**просто**	**время**	**первый**	**значит**
человек		0.428	0.236	0.382	0.264	0.275
говорит	0.428	0.429	0.378	0.414	0.323	0.429
просто	0.236	0.378		0.227	0.171	0.328
время	0.382	0.414	0.227		0.369	0.291
первый	0.264	0.323	0.171	0.369		0.175
значит	0.275	0.429	0.328	0.291	0.175	

Таблица 4: Косинусные расстояния, модель: НКРЯ и Wikipedia

	человек	**говорит**	**просто**	**время**	**первый**	**значит**
человек		0.156	0.260	0.109	-0.099	0.330
говорит	0.156		-0.006	-0.028	-0.006	0.345
просто	0.260	-0.006		0.108	-0.074	0.459
время	0.109	-0.028	0.108		-0.006	0.161
первый	-0.099	-0.006	-0.074	-0.006		-0.055
значит	0.330	0.345	0.459	0.161	-0.055	

Таблица 5: Косинусные расстояния, модель: Araneum fastText

значит, что эти слова часто встречаются в похожих контекстах. Это неудивительно, поскольку выбраны одни из самых частотных слов. Например, предлоги и местоимения (самые частотные слова в русском языке, где нет артиклей) встречаются в совершенно разнообразных контекстах.

Расстояния между словами приведены в соответствующих таблицах. Названия моделей указаны в названиях таблиц.

Любопытно, что для первой модели не получено отрицательных чисел, а для второй получено. Возможно, это из-за того, что слова высокочастотные и встречаются часто в принципе, а значит часто вместе. А вторая модель обучена на большем корпусе, следовательно учитывает больше контекстов.

По версии модели, обученной на НКРЯ и Википедии, самые близкие слова из перечисленных это 'говорить' и 'значит'. А самые далекие по смыслу 'просто' и 'первый'. По версии Araneum fastText самые близкие слова 'просто' и 'значит', а самые далекие 'человек' и 'первый'. Разница в этих парах может объясняться еще и отсутствием разметки по частям речи в корпусе для Araneum fastText, следовательно все слова, которые пишутся одинаково для нее являются одним и тем же словом.

Заключение

В отчете описаны результаты работы с веб-сервисом RusVectōrēs. С помощью RusCorpora выбраны частотные слова, для которых исследовались близкие по смыслу слова для трех предобученных моделей. Кроме того исследованы косинусные расстояния между выбранными словами.

По приведенным табличкам видно, что близкие по косинусному расстоянию слова действительно близки по смыслу и могут быть синонимами. Любопытным показалось слово 'мол', как синоним к слову 'значит'. Действительно, это слово может интерпретироваться

как глагол и как наречие, вводное слово. И то, и другое значения оказываются близкими по смыслу.

По близким словам, найденным тремя различными моделями, можно сделать вывод, что модель обученная на большем количестве данных, находит близкими довольно редкие слова, попавшие в корпус, плюс для таких больших корпусов нет разметки по частям речи. Поэтому результаты модели Araneum fastText отличаются от стандартных. Кроме того, для ее обучения использовался немного другой механизм, позволяющий обучаться быстрее, и позволяющий обработать такое количество слов в корпусе.