Домашнее задание №4

Дудырев Егор

10.05.2020

1 Задание

Задание С
Для нескольких (5-9) достаточно частотных слов русского языка и семантически связанных с ними слов (синонимов, гипонимов, гиперонимов, антонимов и др.), подобранных из некоторого лингвистического ресурса, в частности,

- тезауруса РуТез: http://www.labinform.ru/pub/ruthes/index.htm или
- системы КроссЛексика: http://www.xl.gelbukh.com

исследовать соотношение их векторных представлений на базе предобученной модели Word2Vec и/или FastText, взятой, например, из ресурса Rusvectores: http://rusvectores.org/ru/models/ или из библиотеки Gensim: https://radimrehurek.com/gensim/models/word2vec.html.

Определить, есть ли какие-либо закономерности в косинусном расстоянии между лексически (тезаурусно) связанными словами? Подобны ли расстояния между парами связанных слов? Выполняются ли алгебраические зависимости?

Отчет: описание использованной модели и проведенных экспериментов, выявленные закономерности (если таковые обнаружились); выводы.

2 Сбор данных

Возмём данные по частоте употребления слов из "Нового частотного словаря русской лексики" О.Н. Ляшевской, С.А. Шарова. В среднем слова употребляются 1 раз на милион. Выберем достаточно частотные слова, которые употребляются 100 раз на милион (1 раз на 10т. употреблений)

Данные по синонимам, гипонимам и гиперонимам возмём из тезауруса РуТез. Антонимы определим по справочнику системы КроссЛексика.

Таким образом выделили следующие достаточно часто употребляемые слова, а также другие слова, связанные с ними:

1. Механизм
    - Синонимы: автомат, аппарат, робот, механизм, машина, установка, прибор, устройство, приспособление
    - Гипонимы: агрегат, комбайн, узел, амортизатор, антенна, сепаратор, смеситель
    - Гиперонимы: приспособление, инструмент, процедура, строение, структура
    - Антонимы: $\emptyset$
2. Столица
    - Синонимы: стольный
    - Гипонимы: абу-даби, будапешт, гавана, дублин, каир, хельсинки, тегеран, токио, москва
    - Гиперонимы: город
    - Антонимы: провинция
3. Мешать
    - Синонимы: воспрепятствовать, помешать, препятствовать, смешать, служить помехой, помешивать, смешиваться
    - Гипонимы: задержать, затруднить, парализовать, путать, примешать, перемешать, размешать, дергать, тормозить, сорвать
    - Гиперонимы: влиять, воздействовать, переместить
    - Антонимы: возбуждать, деблокировать, действовать, настраивать, облегчать, помогать
4. Рассматривать
    - Синонимы: дорассмотреть, проанализировать, разбирать, разобрать, разлядеть, разглядывать, расценивать
    - Гипонимы: анализ, вычислить, критиковать, обозреть, засмотреться, любоваться, смотреться, рефлексия, экспертиза, рассудить

- Гиперонимы: определить, выяснить, оценить, осмотреть, оглядеть
- Антонимы: синтезировать

5. Поведение
    - Синонимы: поведенческий
    - Гипонимы: аскетизм, бесстрашие, геройство, привычка, своеволие, эпатаж, подвижничество, ласкать, скупиться, смелость
    - Гиперонимы: действие
    - Антонимы: ∅
6. Мастер
    - Синонимы: искусница, мастак, мастерица, умелец, умелица
    - Гипонимы: визажист, ювелир, стилист, кулинар, снайпер, ремесленник, кустарь, ювелир
    - Гиперонимы: спортсмен, работник, руководитель, человек
    - Антонимы: дилетант, ломастер, неспециалист, неумеха
7. Газ
    - Синонимы: газовый, газообразный, газик
    - Гипонимы: азот, аммиак, бутан, фтор, хлор, плазма
    - Гиперонимы: вещество, автомобиль
    - Антонимы: ∅
8. Бабушка
    - Синонимы: бабка, бабуля, бабулька, старуха, бабуся
    - Гипонимы: прабабка
    - Гиперонимы: родственница
    - Антонимы: ∅
9. Высота
    - Синонимы: вышина, апофеоз, венец, верх, зенит, разгар, кульминация, высотка
    - Гипонимы: расцвет, процветание

- Гиперонимы: пик, степень, расстояние, возвышенность, пространство, протяженность
- Антонимы: низина

10. Собрать

    - Синонимы: насобирать, пособирать, собирание, собирать, напрягать, напрячь, поднажать, пособирать, убрать
    - Гипонимы: копить, накапливать, набрать, организовать, подобрать, составить, вслушаться, всмотреться, обобрать, перенапрячься
    - Гиперонимы: прикрепить, приделать, соединить, создать, снять, пытаться, стараться
    - Антонимы: израсходовать, надеть, назначить, натянуть, промотать, разбросать, развести, разобрать, растащить

3 Используемая модель

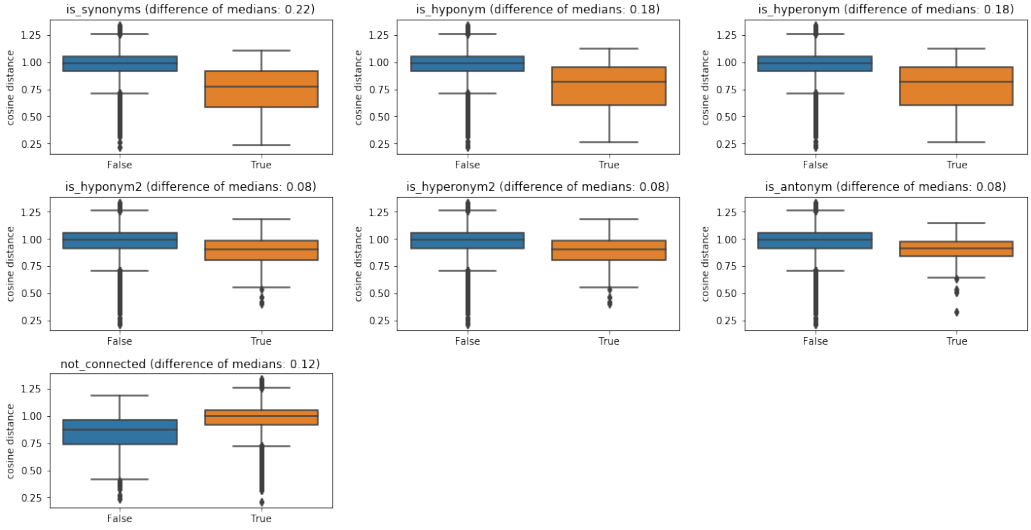
В качестве Word2Vec модели будем использовать модель ruscorpora_upos_cbow_300_20_2019 с сайта Resvectores. Она обучалась на корпусе НКРЯ из 270 миллионов слов. В качестве алгоритма использовался Continuous Bag-of-Words.

4 Поиск зависимостей в векторах

4.1 Похожесть между связанными словами

Вопрос: Являются ли векторы синонимичных слов более похожими друг на друга, чем векторы несвязанных слов? Выполняется ли такая зависимость для других типов связей?
Кроме базовых отношений будем также считать отношения гипонимии и гиперонимии 2ой степени (т.е. такие A и B, что A - гипоним/гипероним C, C - гипоним/гипероним B).

Рис. 1: Расстояния между связанными словами

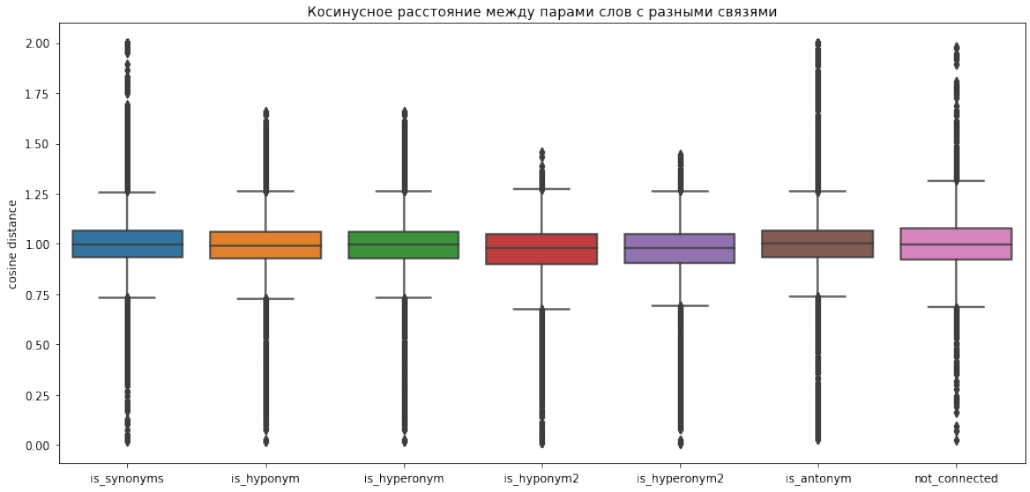


Векторы связанных слов больше похожи друг на друга, чем несвязанных. Расстояния между связанными словами меньше, чем между несвязаннами.

4.2 Векторы связанностей слов

Может быть есть векторы "синонимичности "гипонимичности"и т.д.? Для этого сравним векторы разностей для синонимов, антонимов и т.п. Если векторы разностей синонимов будут похожи между собой, значит можно сказать, что существуют особые "векторы синонимичности". Для каждого типа связи случайным образом выберем 100 пар слов, чтобы можно было сравнивать доверительные интервалы для разных типов связей.

Рис. 2: Расстояния между словами с разным типом связи

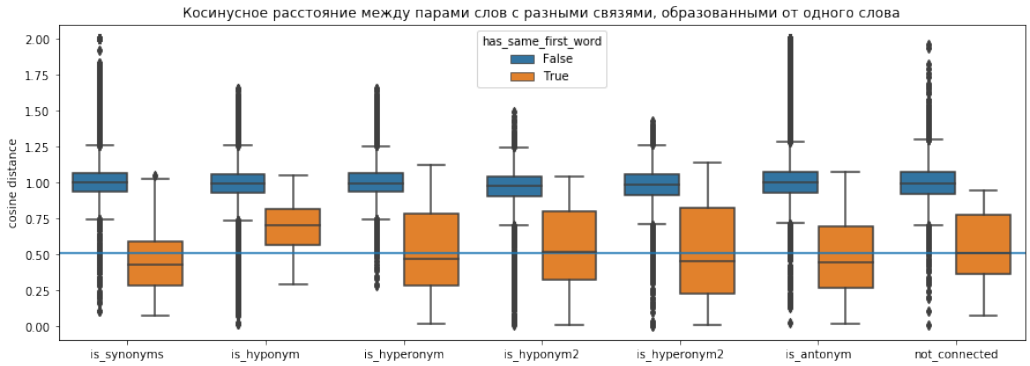


Расстояния между словами с одинаковым типом связи в среднем равны 1. Таким образом, можно сказать, что не существует уникального вектора, отображающего "синонимию "антонимию"и остальные связи.

4.3 Векторы связанностей слов для каждого слова

Может быть векторы "синонимии "антонимии"и других связей существуют в рамках каждого конкретного слова?.

Рис. 3: Расстояния между словами с разным типом связи



Пары слов, синонимичных к одному и тому же слову, похожи друг на друга. Однако настолько же похожи и любые другие две пары, в которых содержатся одинаковые слова. Таким образом, нельзя сказать, что су-

ществуют векторы "антонимичности "синонимичности "гипонимии "гиперонимии"даже в пределах одного слова.

4.4 Алгебраические зависимости

Может быть на векторах слов работают хотя бы алгебраические зависимости?

Классический пример

Для начала рассмотрим известный пример с королем и королевой.
Известно, что

$$\text{король} - \text{королева} \approx \text{муж} - \text{жена}$$

Тогда:

$$\text{король} \approx \text{муж} - \text{жена} + \text{королева}$$

	Слово	Похожесть
0	королева_NOUN	0.920119
1	королева_ADV	0.750072
2	королева_ADJ	0.682138
3	принцесса_NOUN	0.680564
4	королева_PROPN	0.668374
5	герцогиня_NOUN	0.667244
6	мазарини_PROPN	0.640198
7	мария::стюарт_PROPN	0.616200
8	вальер_PROPN	0.616168
9	фаворитка_NOUN	0.609127

Предполагалось, что результатом окажется вектор, похожий на короля (или что-то из королевской тематики мужского рода) но результат оказался совсем другим.

$$\text{муж} \approx \text{король} - \text{королева} + \text{жена}$$

	Слово	Похожесть
0	жена_NOUN	0.729585
1	зять_NOUN	0.658955
2	тесть_NOUN	0.609647
3	муж_NOUN	0.580768
4	сын_NOUN	0.564723
5	супруга_NOUN	0.550519
6	жена_ADJ	0.529578
7	зять_VERB	0.523330
8	отец_NOUN	0.522891
9	племянник_NOUN	0.514432

В этом случае наиболее похожим словом должен был оказаться "муж". Он только на 4ом месте. Однако большую часть похожий слов занимают члены семьи мужского рода, что тоже неплохо.

2 наиболее похожие пары слов

Разницы пары (выяснить, экспертиза), (определить, экспертиза) очень похожи друг на друга (расстояние = 0,007253). Проверим на них алгебраические зависимости.

$$\text{выяснить} - \text{экспертиза} \approx \text{определить} - \text{экспертиза}$$

$$\text{выяснить} \approx \text{определить} - \text{экспертиза} + \text{экспертиза} \approx \text{определить}$$

Если алгебраические зависимости выполняются, то слово "выяснить"очень похоже на слово "определить". Но расстояние между ними равно 0,831424. Т.о. алгебраические зависимости в данном случае не выполняются.
Тогда почему эти слова имеют похожую разность векторов? Вероятно эти пары одинаково часто встречаются в текстах: когда кто-то провёл экспертизу и выяснил что-то или провёл экспертизу и определил что-то.

Похожие пары синонимов

Будем искать похожие пары среди тех, в которых ни одно слово не совпадает. Тогда слова не "сократятся"как в прошлом примере.
Оказывается друг на друга сильно похожи две пары (старуха, бабулька), (бабка, бабуля).
Тогда, при наличии алгебраических соотношений было бы:

$$\text{бабка} \approx \text{старуха} - \text{бабулька} + \text{бабуля}$$

	Слово	Похожесть
0	старуха_NOUN	0.926507
1	старушка_NOUN	0.775208
2	бабка_NOUN	0.762171
3	бабушка_NOUN	0.707106
4	тетка_NOUN	0.692906
5	старушонка_NOUN	0.691257
6	старик_NOUN	0.681647
7	агафья_PROPN	0.647350
8	соседка_NOUN	0.638528
9	лукерья_PROPN	0.631201

Можно сказать, что модель учитывает "анти уменьшительно-ласкательные"отношения и делать слова более грубыми.

Похожие пары антонимов

Посмотрим на похожие пары среди антонимов. Одна из них - (мастер, неспециалист), (умелец, неумеха).

Тогда при наличии алгебраических соотношений было бы:

$$\text{неумеха} \approx \text{неспециалист} - \text{мастер} + \text{умелец}$$

	Слово	Похожесть
0	неспециалист_NOUN	0.360150
1	умелец_NOUN	0.317757
2	вездеход_NOUN	0.301661
3	съедобный_ADJ	0.301216
4	лазера_NOUN	0.294047
5	гош_PROPN	0.293200
6	толком_ADV	0.292733
7	череп_PROPN	0.289082
8	приборчик_NOUN	0.289035
9	гуманоид_NOUN	0.287042
10	поверье_NOUN	0.284934
11	сейд_NOUN	0.279291
12	глушилка_NOUN	0.275649
13	псих_NOUN	0.275639
14	наверняка_ADV	0.275388
15	абориген_NOUN	0.275057
16	штуковина_NOUN	0.273606
17	экстрасенс_NOUN	0.268383
18	холодильник_NOUN	0.266996
19	пиранье_NOUN	0.265510

В качестве ответа должен был получиться кто-то с плохой компетенцией. Но этого не оказалось.
С другой стороны, если рассмотреть следующую зависимость:

$$\text{мастер} \approx \text{мастерица} - \text{неспециалист} + \text{дилетант}$$

	Слово	Похожесть
0	мастерица_NOUN	0.804068
1	рукодельница_NOUN	0.570206
2	дилетант_NOUN	0.538852
3	искусница_NOUN	0.523112
4	любительница_NOUN	0.501103
5	мастер_NOUN	0.495654
6	швея_NOUN	0.474587
7	швь_NOUN	0.472640
8	музыкантша_NOUN	0.464685
9	виртуоза_NOUN	0.457687
10	портниха_NOUN	0.457567
11	искусник_NOUN	0.456476
12	виртуоз_NOUN	0.455557
13	мастериец_NOUN	0.453927
14	вышивальщица_NOUN	0.452901
15	кокетка_NOUN	0.447906
16	портретист_NOUN	0.445566
17	белошвейный_ADJ	0.440653
18	кружевница_NOUN	0.440251
19	охотница_NOUN	0.437364

В этом случае должен был получиться кто-то с хорошей компетенцией. И правда: "искусница "мастер "виртуоза опытные люди.

5 Вывод

Данная модель Word2Vec на базе подхода CBOW, не определяет семантические зависимости между словами. В частности: синонимию, антонимию, гиперонимию, гипонимию.

Интересно, что наиболее похожи друг на друга слова, являющиеся гипонимами гипонимов. Скорее всего такие слова просто часто встречаются в тексте рядом друг с другом, так как они одновременно обозначают и похожие и слегка различные понятия и действия.

Иногда можно заметить, будто модель улавливает различия между грубыми и негрубыми словами, будто она видит разницу между компетентными и некомпетентными людьми. Однако это похоже на исключения из правил. А также похоже на результат проверки множества гипотез: если очень сильно искать моделируемые сложные зависимости, то обязательно можно что-нибудь найти.

6 Используемое ПО

Язык программирования: Python 3.7 (с дистрибутивом Anaconda).
Используемые библиотеки:

- numpy 1.17.1
- pandas 0.23.4
- matplotlib 3.0.2
- seaborn 0.9.0
- re 2.2.1
- gensim 3.6.0
- sklearn 0.20.1
- tqdm 4.43.0

7 Ссылки на код

- Исполняемый код
- Графики