Выявление семантических сдвигов с использованием Word2Vec

Автор: Любовь Беликова

Данные

Тексты выступлений официальных представителей МИД РФ за 2003 – 2020 гг.

- Источник: mid.ru
- Объем корпуса: около 34 тыс. текстов
- Жанры: брифинги, комментарии, ответы на вопросы СМИ, официальные заявления

Исследовательский вопрос

Какие слова поменяли свое значение после 2014 г.?

Метод

1. Обучение Word2Vec на двух корпусах: тексты до 2014 г. и все тексты
2. Сравнение семантических ассоциатов для каждого слова в словаре в первом и втором корпусе с помощью коэффициента Жаккара
3. Выявление слов, для которых набор семантических ассоциатов изменился наиболее сильно

Источник: Kutuzov, A., Kuzmenko E. Two centuries in two thousand words: Neural embedding models in detecting diachronic lexical changes // Quantitative Approaches to the Russian Language (ed. by Kopotev M., Lyashevskaya O., Mustajoki A.). – Routledge New York, 2017. - P. 177-196.

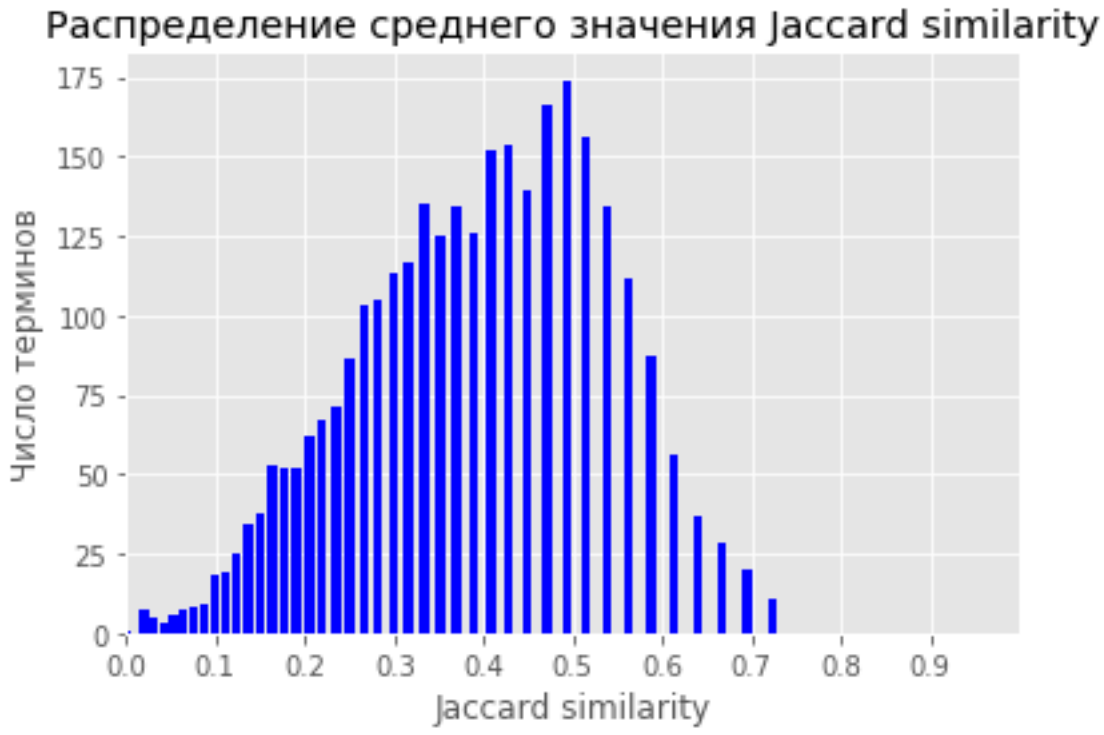
Шаг 1. Обучение моделей Word2Vec

- Корпус до 2014 г. – около 18 тыс. текстов
- Весь корпус – около 34 тыс. текстов
- Модель: `gensim.models.Word2Vec(documents, size=300, window=5, min_count=15)`
- Исключены слова, которые не встречаются в первом корпусе
- Исключены стоп-слова, а также слишком редкие ($p < 0.005$) и частые ($p >= 0.85$) слова

Шаг 2. Измерение коэффициента Жаккара

- 50 семантических ассоциатов для каждого слова
- Коэффициент Жаккара = доля пересечений

Шаг 3. Выявление слов с максимальным семантическим сдвигом



Топ-20:

киевский 0.0
правый 0.010101010101010102
минский 0.02040816326530612
киев 0.02040816326530612
образец 0.02040816326530612
выявлять 0.030927835051546393
умеренный 0.030927835051546393
нелегитимный 0.030927835051546393
кремль 0.030927835051546393
возвращаться 0.04166666666666664
подача 0.04166666666666664
пропагандистский 0.05263157894736842
печально 0.05263157894736842
анкара 0.05263157894736842
избиратель 0.05263157894736842
переворот 0.05263157894736842
крым 0.05263157894736842
предвыборный 0.05263157894736842
недобросовестный 0.06382978723404255
многократно 0.06382978723404255

Результаты

- *киевский, минский, киев, крым, переворот* – от общих слов к теме Украинского кризиса
- *правый* – от темы прав человека (особенно нацменьшинств) к теме националистов, радикалов
- *образец* – от образца в значении «образец для подражания» к образцу какого-либо вещества
- *умеренный* – от общих слов к теме «умеренной» сирийской оппозиции
- *возвращаться* – от общих слов к теме беженцев
- *подача* – от подачи документа (заявления) к подаче информации, особенно ложной
- *предвыборный, недобросовестный, пропагандистский* – к темам дезинформации в медиа, предвзятого отношения к России

Проблемы

- Ряд слов в топе (*выявлять, нелегитимный, кремль, анкара, многократно*) на самом деле не поменяли смысла
- Слова, у которых резко повышается частотность, получают большой семантический сдвиг, хотя фактически контекст их употребления мог остаться прежним
- Слова с реальным семантическим сдвигом могут оказаться внизу списка

ipynb-тетрадка:
https://github.com/meancatt/CL/blob/master/Project3/SemanticShifts2014.ipynb