Национальный исследовательский университет
Высшая школа экономики
Факультет компьютерных наук

Отчет по домашнему заданию №4

Лексическая семантика

студент 1 курса магистратуры
Зуев Кирилл Александрович

Москва, 2019

Содержание

1 Постановка задачи выбранного варианта (C)

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.

2 Уточнение постановки задачи

Для эксперимента я рассмотрел встроенную в **Gensim** модель (обученную на **НКРЯ**) и 3 векторные модели для РЯ с сайта:

- **НКРЯ**;
- **НКРЯ** и **Wikipedia**;
- **Тайга**.

Для исследования я выбрал 9 слов разных частей речи:

- существительные:
  - *кот*;
  - *одежда*;
  - *машина*;
- прилагательные:
  - *красный*;
  - *большой*;
- глаголы:
  - *делать*;
  - *говорить*;
- наречие *хорошо*;
- числительное *мало*.

3 Характеристика использованных моделей

3.1 Встроенная в Gensim модель

- Идентификатор: **word2vec-ruscorpora-300**;
- Корпус: **НКРЯ**;
- Размер корпуса: 250 млн.;
- Объём словаря: 185 тыс.;
- Размерность вектора: 300;
- Размер окна: 10.

3.2 Модель с сайта, обученная на корпусе НКРЯ

- Идентификатор: **ruscorpora_upos_cbow_300_20_2019**;
- Корпус: **НКРЯ**;
- Размер корпуса: 270 млн.;
- Объём словаря: 189 тыс.;
- Размерность вектора: 300;
- Размер окна: 20.

3.3 Модель с сайта, обученная на корпусе НКРЯ и Wikipedia

- Идентификатор: **ruwikiruscorpora_upos_skipgram_300_2_2019**;
- Корпус: **НКРЯ** и **Wikipedia**;
- Размер корпуса: 788 млн.;
- Объём словаря: 249 тыс.;
- Размерность вектора: 300;
- Размер окна: 2.

3.4 Модель с сайта, обученная на корпусе Тайга

- Идентификатор: **tayga_upos_skipgram_300_2_2019**;
- Корпус: **Тайга**;
- Размер корпуса: 5 млрд.;
- Объём словаря: 250 тыс.;
- Размерность вектора: 300;
- Размер окна: 2.

4 Проведённые эксперименты

4.1 Поиск семантически близких слов

Для каждого из выбранных мной слов я нашел наиболее близкие.

- ***кот***

  Для разных моделей получились примерно одинаковые результаты, с разными значениями близости получились слова: *кошка*, *котёнок*, *пёс* и др.

- ***морковь***

  Для разных моделей получились примерно одинаковые результаты, с разными значениями близости получились слова: *капуста*, *помидор*, *картофель* и др.
  Но для модели, обученной на **НКРЯ**, на первых местах оказались слова *сельдерей*, *корнеплод* и *укроп*, которые ещё встречаются во встроенной в **Gensim** модель, но не встречаются в остальных.

- ***ботинок***

  Для разных моделей получились примерно одинаковые результаты, с разными значениями близости получились слова: *сапог*, *туфля*, *башмак* и др.
  Но для модели, обученной на **НКРЯ**, на третьем месте оказалось слово *полуботинок*, которое в целом довольно странное и не встречается ещё только во встроенной в **Gensim** модели.

- ***красный***

  Для разных моделей получились различные похожие слова, обозначающие либо близкие с красным (*алый*, *оранжевый* и *малиновый*), либо другие цвета (*синий*, *белый* и *зелёный*).

- ***большой***

  Для всех моделей получились примерно одинаковые результаты, с разными значениями близости получились слова: *огромный*, *громадный*, *небольшой* и др.

- ***идти***

  Для всех моделей самым близким оказалось слово *пойти*. Также в большинстве моделей встречаются слова *шагать*, *бежать* и *ехать*.

- ***говорить***

  Для разных моделей получились примерно одинаковые результаты, с разными значениями близости получились слова: *сказать*, *рассуждать*, *толковать* и др.

- ***мало***

  Для моделей с сайта получились примерно одинаковые результаты, с разными значениями близости получились слова: *много*, *больше*, *мало* и др. А для встроенной в **Gensim** модели наиболее схожими оказались слова: *мало*, *менее*, *немного* и др.

- ***хорошо***

  Для разных моделей получились примерно одинаковые результаты, с разными значениями близости получились слова: *плохо*, *отлично*, *прекрасно* и др.

4.2 Вычисление близости пар слов

Из рассматриваемых слов я составил всевозможные пары и посчитал их близость для модели, встроенной в **Gensim**. Затем я отсортировал их по значению полученной близости и вывел 10 наиболее похожих. Для этих 10 пар я также посчитал близость для моделей с сайта.

1. *делать* и *говорить*: 0.420 (**НКРЯ**: 0.329, **НКРЯ** и **Wikipedia**: 0.474, **Тайга**: 0.273);
2. *говорить* и *хорошо*: 0.373 (**НКРЯ**: 0.120, **НКРЯ** и **Wikipedia**: 0.394, **Тайга**: 0.215);
3. *мало* и *хорошо*: 0.360 (**НКРЯ**: 0.214, **НКРЯ** и **Wikipedia**: 0.326, **Тайга**: 0.183);
4. *большой* и *мало*: 0.333 (**НКРЯ**: 0.294, **НКРЯ** и **Wikipedia**: 0.386, **Тайга**: 0.218);
5. *кот* и *ботинок*: 0.326 (**НКРЯ**: 0.232, **НКРЯ** и **Wikipedia**: 0.352, **Тайга**: 0.247);
6. *делать* и *хорошо*: 0.301 (**НКРЯ**: 0.081, **НКРЯ** и **Wikipedia**: 0.298, **Тайга**: 0.135);
7. *красный* и *большой*: 0.280 (**НКРЯ**: 0.182, **НКРЯ** и **Wikipedia**: 0.330, **Тайга**: 0.215);
8. *большой* и *хорошо*: 0.241 (**НКРЯ**: 0.218, **НКРЯ** и **Wikipedia**: 0.277, **Тайга**: 0.118);
9. *кот* и *говорить*: 0.239 (**НКРЯ**: -0.088, **НКРЯ** и **Wikipedia**: 0.261, **Тайга**: 0.126);
10. *большой* и *говорить*: 0.231 (**НКРЯ**: -0.036, **НКРЯ** и **Wikipedia**: 0.315, **Тайга**: 0.110).

5 Код программы

Реализация программы для исследования встроенной в **Gensim** модели находится в приложенных файлах **HW4.pdf** и **HW4.ipynb**.

6 Выводы по исследованию

При поиске семантически близких для выбранного набора слов (достаточно частотных) в большинстве случаев с разными значениями близости выдавались примерно одинаковые результаты. В паре случаев для части моделей результаты отличались. Думаю, это связано с тем, что использовались разные корпусы разных объёмов и к тому же разные размеры окон. В целом на результат данного эксперимента это влияло не так сильно.

При вычислении близости пар выбранных слов результаты получились довольно разные. В первую очередь, это связано с тем, что все слова по смыслу отличаются друг от друга, а, во-вторых, значение близости очень зависит от параметров рассматриваемой модели и корпусе, на котором она обучена. Но для нескольких пар различия получились не очень большими, значит, они действительно в части случаев употребляются в одном и том же контексте или похожи по смыслу.