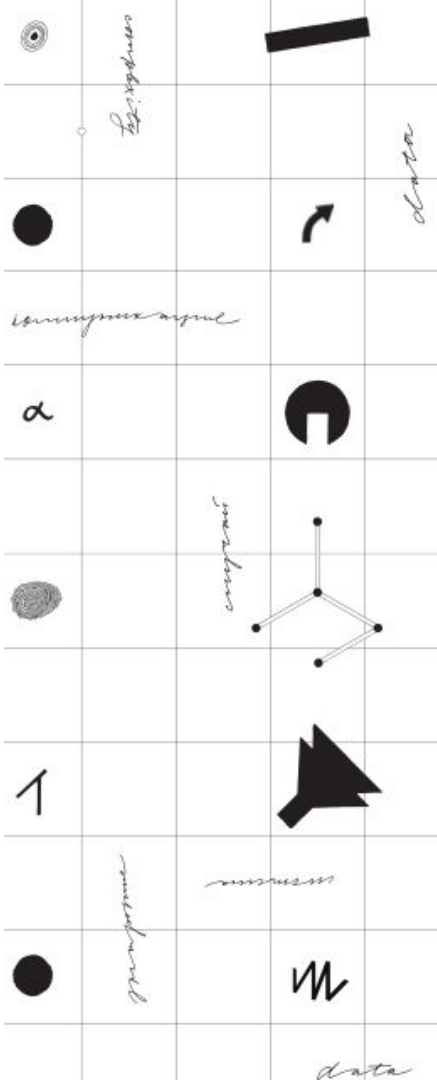
Модель Word2Vec

На основе работы *Mikolov, Tomas, et al. "Distributed Representations of Words and Phrases and their Compositionality."Advances in Neural Information Processing Systems, vol. 26, 2013*

Тимченко Даниил
ИППИ РАН
25.03.24



Содержание

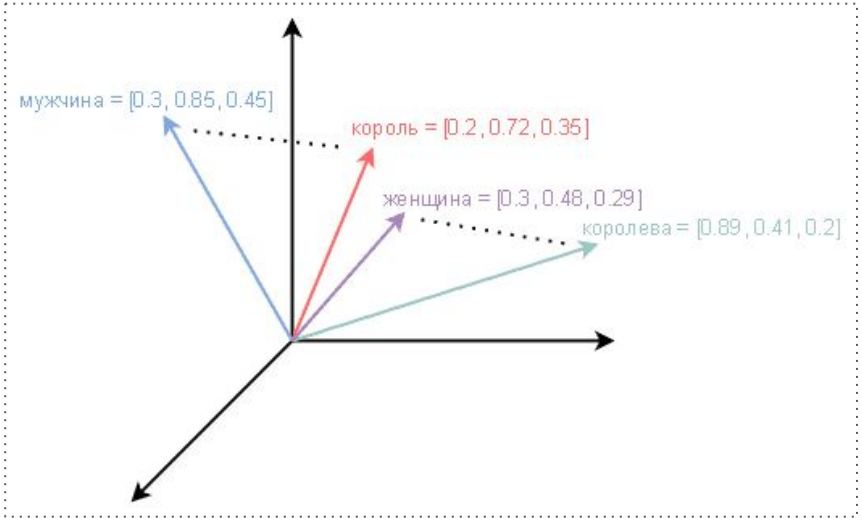
1. Постановка задачи векторизации слов
2. Существующие методы векторизации и их проблемы
3. Описание метода Word2Vec
4. Описание экспериментов для оценки качеств метода
5. Результаты экспериментов
6. Анализ и выводы

Задача векторизации слов

Задача векторизации слов — это **метод получения векторных представлений слов (эмбеддингов)**, которые будут отражать смысл слов, синтаксические и семантические связи между словами.

Эмбеддинги используются в таких задачах как:

- Морфологический анализ
- Составление карты языка
- Анализ тональности текста



Пример эмбеддингов в трехмерном пространстве

Существующие методы

One Hot Encoding [1]

Проблема данного метода:

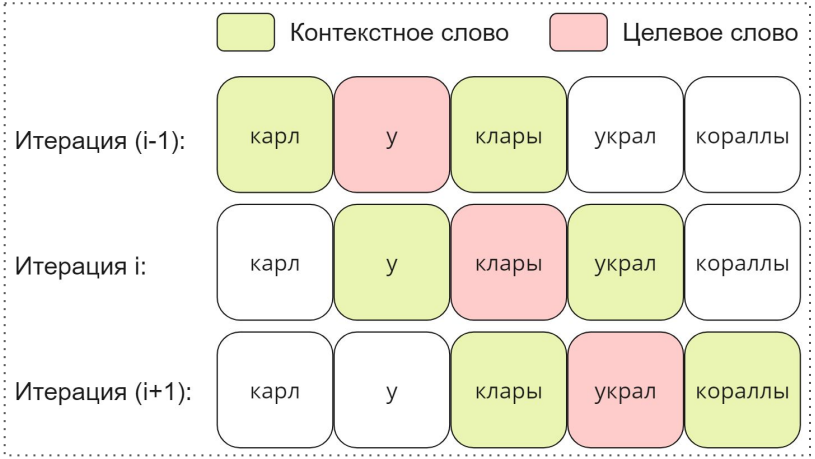
- Векторы не отражают смысл слов
- Сильная разреженность векторов
- Высокая размерность векторов
- Фиксированность словаря
- Слова, не попавшие в словарь, не могут быть обработаны

Пример векторизации One Hot Encoding

[1] Stevens, S. S. (1946). “On the Theory of Scales of Measurement”. Science, New Series, 103.2684,677–680.

Модель Word2Vec

1. Сбор обучающего набора данных из предобработанного текста с помощью функции скользящего окна
   - Пары слов: (целевое, контекстное)
2. Определение loss-функции с учетом использования негативного сэмплирования



Пример нескольких итераций скользящего окна размера 1

$$\log \sigma({v'_{w_O}}^\top v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-{v'_{w_i}}^\top v_{w_I}) \right]$$

Loss-функция, где первое слагаемое отвечает за loss правильного контекстного слова, а второе за loss шумовых (негативных) векторов

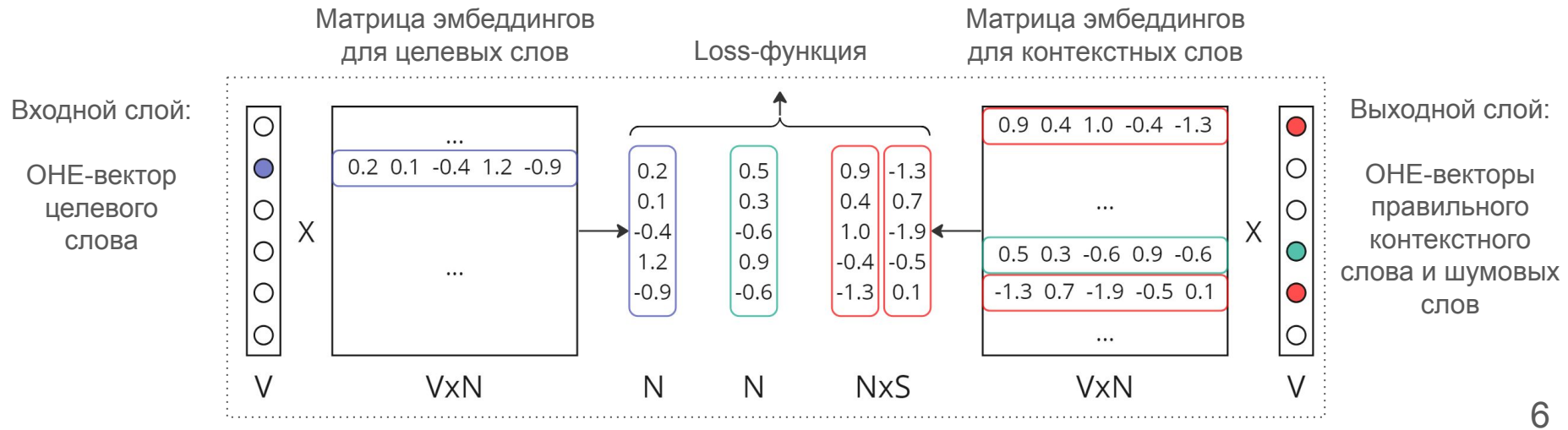
Модель Word2Vec

3. Определение распределение шума в тексте для правильного выбора негативных сэмплов

$$P_n(w) = \frac{U(w)^{\frac{3}{4}}}{Z}$$

Распределение шума, выраженное через вероятностное распределение слов в тексте

4. Задание структуры модели Word2Vec



Описание экспериментов

➔ Качественная оценка способности модели к восприятию синтаксических аналогий

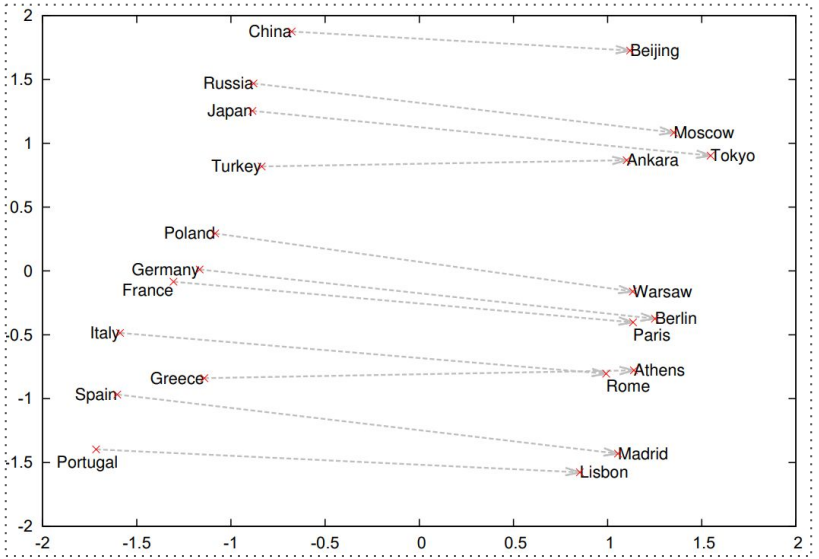
Пример задачи синтаксической аналогии:

- “быстро” : “быстрый” :: “медленно” : ?
  где правильным ответом будет “медленный”

➔ Качественная оценка способности модели к восприятию семантических аналогий

Пример задачи семантической аналогии:

- “Германия” : “Берлин” :: “ Франция” : ?
  где правильным ответом будет “Париж”



Пример результата эксперимента семантических аналогий из оригинальной работы

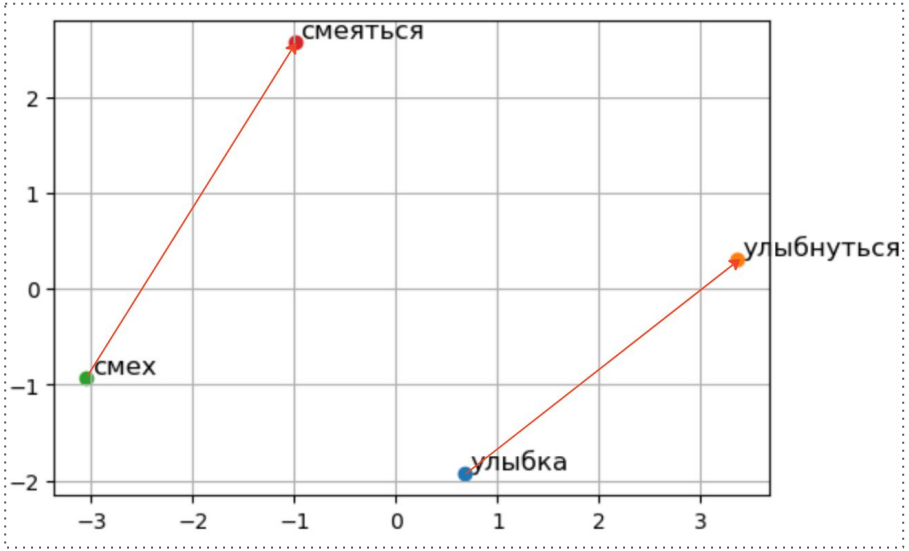
Синтаксические связи

- Синтаксические связи между словами:
  - смех —> смеяться
  - улыбка —> улыбаться

```
Наиболее похожие слова на
vec(смеяться) - vec(смех) + vec(улыбка):

улыбаться:       0.5608
утвердительно:   0.6047
просиять:        0.6134
имя:             0.6438
насмешливый:     0.6541
```

Вывод модели на запрос
“5 наиболее похожих слов на…”



Вектора исследуемых слов, спроецированные методом понижения размерности (PCA)

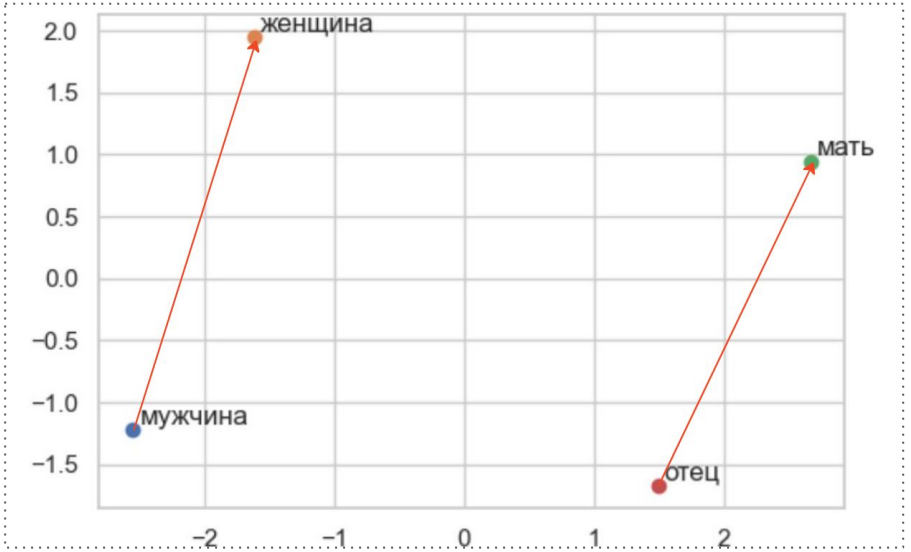
Семантические связи

- Семантические связи между словами:
  - мужчина —> женщина
  - отец —> мать

```
Наиболее похожие слова на
vec(отец) - vec(мужчина) + vec(женщина):

мать:       0.5944
написать:   0.6417
сын:        0.6664
дымок:      0.67
дочь:       0.6806
```

Вывод модели на запрос
“5 наиболее похожих слов на…”

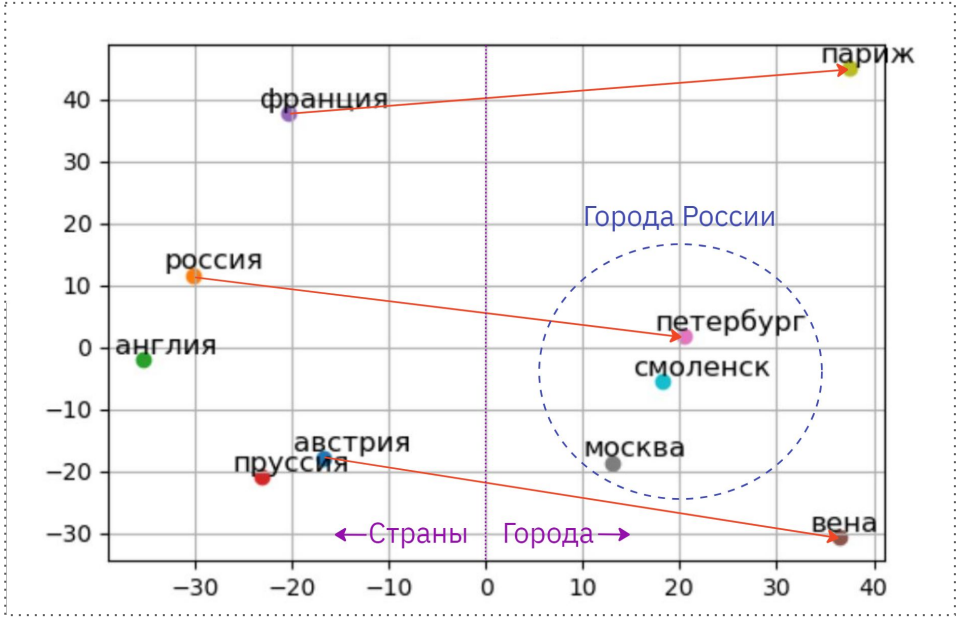


Вектора исследуемых слов, спроецированные методом понижения размерности (PCA)

Семантика в терминах географии

- Способность модели самостоятельно автоматически организовывать концепции и неявно изучать связи между ними

- Во время обучения модели не было предоставлено никакой контролируемой информации о том, что такое столица



Вектора стран и городов, спроецированные методом понижения размерности (PCA)

Анализ и выводы

Преимущества модели Word2Vec:

- Векторы отражают синтаксические и семантические схожести слов
- Возможность дообучения эмбеддингов новыми данными
- Независимость размерности эмбеддингов от размера словаря

Недостатки модели Word2Vec:

- Слова, не присутствующие в обучающей выборке, не могут быть представлены векторами
- Некорректность эмбеддингов для редких слов