Прикладные задачи анализа данных

Семинар 6

Дистрибутивная семантика

Национальныи Исследовательский Университет
Высшая Школа Экономики

Проверочная - `https://goo.gl/forms/NNno4k1A3EHp87KF2`

22 февраля 2018

Определения — 1

- **Word2vec** — способ обучения представлений для слов.
- В лингвистике существует **дистрибутивная гипотеза**: похожие слова имеют похожие смыслы, т.е. в чем более похожих контекстах встречаются два слова, тем ближе должны быть соответствующие им вектора.
- Смысл слова — распределение над контекстами.

Определения — 2

- Мы хотим для каждого слова w из словаря W найти вектор $\vec{w} \in R^d$.
- Пусть дан некоторый текст $x = (w_1, \ldots, w_n)$.
- **Контекст слова** w_j - слова, находящиеся от него на расстоянии не более K, т.е. слова $w_{j-K}, \ldots w_{j-1}, w_j, w_{j+1}, \ldots, w_{j+K}$.
- Вероятность встретить слово w_i в контексте слова w_j:

$$p(w_i|w_j) = \frac{exp(\langle \vec{w}_i, \vec{w}_j \rangle)}{\sum_{w \in W} exp(\langle \vec{w}, \vec{w}_j \rangle)}$$

Обучение представлений

- Теперь рассмотрим выборку текстов $X = \{x_1, \ldots, x_l\}$, где текст имеет длину n_i.
- Можно определить правдоподобие и максимизировать его:

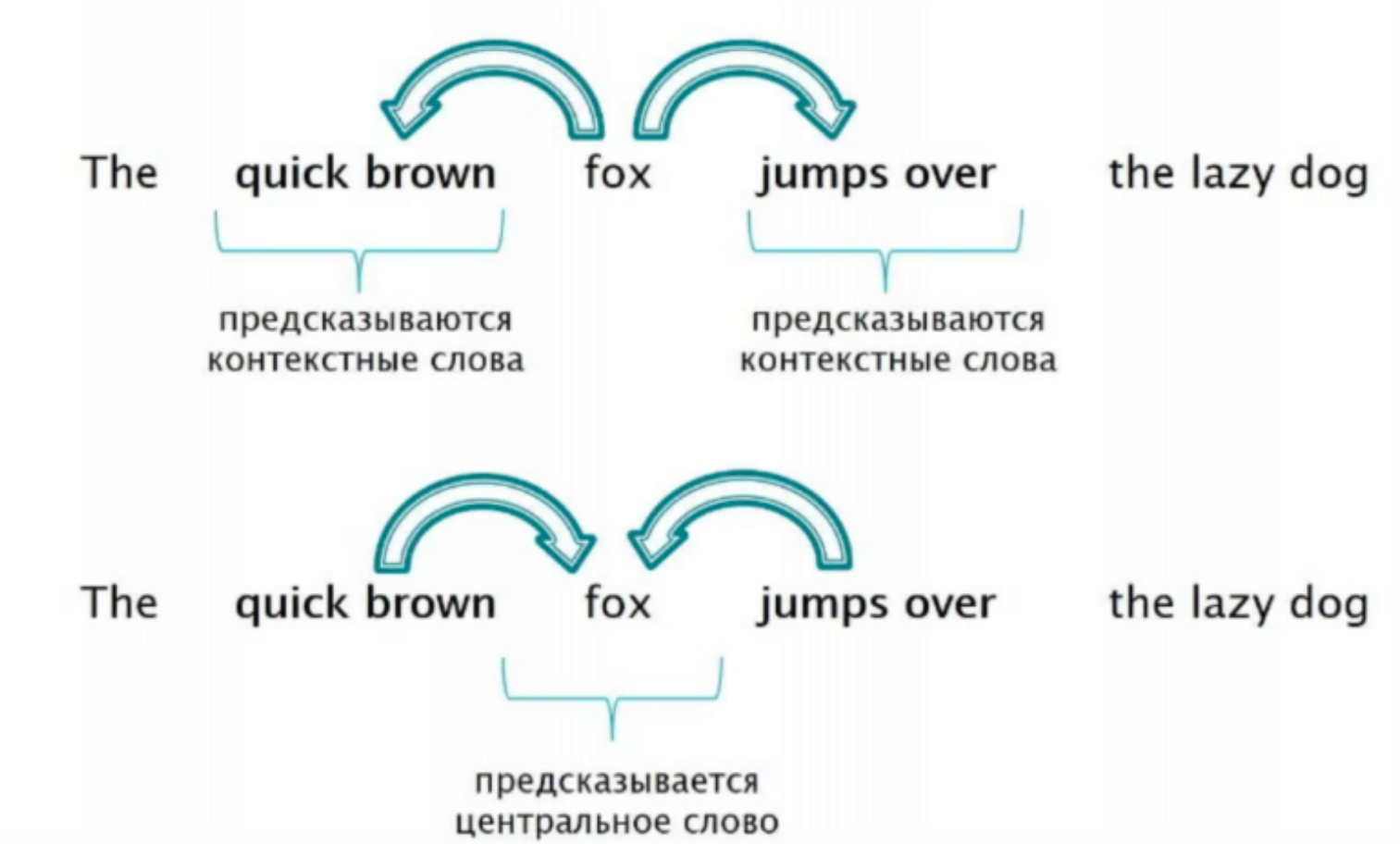
$$\sum_{i=1}^{l} \sum_{j=1}^{n_i} \sum_{k=-K, k\neq 0}^{k=K} \log p(w_{j+k}|w_j) \to \max_{\{\vec{w}\} w \in W}$$

- Этот функционал можно оптимзировать стохастическим градиентным спуском.

Сжатые векторные представления слов

- полезны сами по себе, например, для поиска синонимов или опечаток в поисковых запросах
- используются в качестве признаков для решения самых различных задач:
  - выявление именованных сущностей
  - тэгирование частей речи
  - машинный перевод
  - кластеризация документов
  - ранжирование документов
  - анализ тональности текста

Две модели: Skip-gram и Continuous BOW



Skip-gram и Continuous BOW

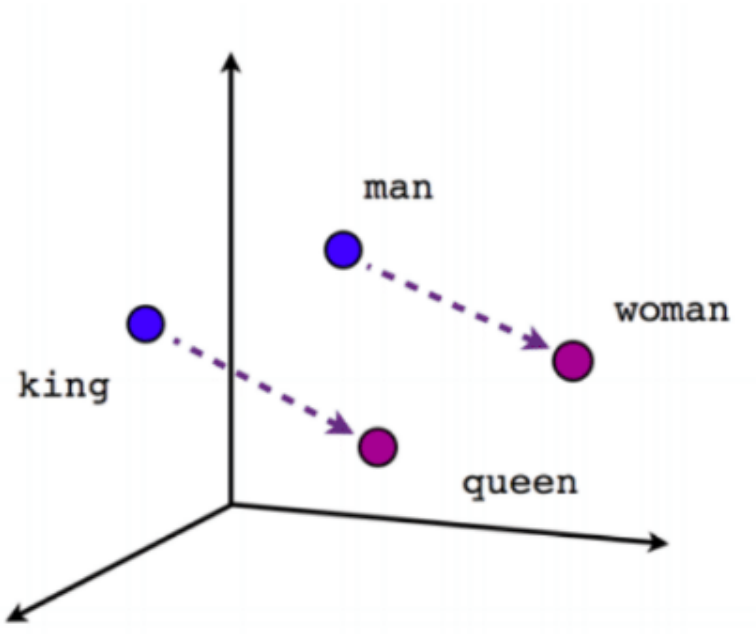
- Skip-gram: какова вероятность встретить соседей при условии данного слова

$$p(w_{i-h}, \ldots, w_{i+h} | w_i) = \prod_{j=i-h}^{i+h} p(w_j | w_i)$$

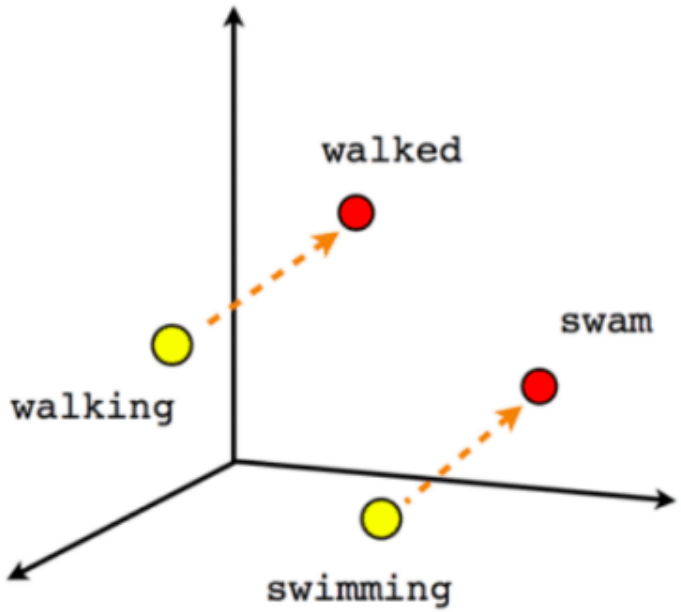
- Continuous BOW: какова вероятность встретить слово в окружении соседей

$$p(w_i | w_{i-h}, \ldots, w_{i+h})$$

Полезные свойства



Male-Female



Verb tense

Миф word2vec

- Это не Deep Learning! А очень простая нейронная сеть.
- Ассоциация с Deep Learning только потому, что на вход глубоких нейронный сетей подаются эмбеддинги

Известные реализации:

- Оригинальный word2vec
- Medallia/Word2VecJava
- FastText
- Spark MLLib Word2Vec
- Gensim word2vec
- и другие

gensim — пакет для тематического моделирования, включает ряд полезных инструментов (часто в качестве удобной обёртки над готовыми реализациями). Предоставляет интерфейс для работы с оригинальным word2vec.

Пример: word2vec в gensim — 1

- Обучем модель на данных английской википедии. ссылка на данные
- Импортируем основные модули:

```
from gensim.corpora import WikiCorpus
from gensim.models import Word2Vec
from gensim.models.word2vec import LineSentence
```

Пример: word2vec в gensim — 2

- Подготовим данные — 100K статей:

```
f = 'enwiki-latest-pages-articles.xml.bz2'
with open('wiki.en.text', 'w') as fout:
    w = WikiCorpus(f, lemmatize=False, dictionary={})
    for i, text in enumerate(wiki.get_texts()):
        fout.write(' '.join(text) + '\n')
        if i == 99999:
            sys.exit()
```

Пример: word2vec в gensim — 3

- Обучим модель:

```
model = Word2Vec(LineSentence('wiki.en.text'),
                 size=200,
                 window=5,
                 min_count=3,
                 workers=8)

# trim unneeded model memory = use (much) less RAM
model.init_sims(replace=True)

model.save('wiki.en.word2vec.model')
```

Пример: word2vec в gensim — 4

- Использование модели:

```
model.most_similar('queen', topn=3)

[(u'king', 0.6691948175430298),
 (u'princess', 0.6487438082695007),
 (u'empress', 0.6162152886390686)]


model.most_similar(positive=['woman', 'king'],
                   negative=['man'], topn=2)

[(u'queen', 0.6960216164588928),
 (u'empress', 0.5979048013687134)]
```

Задание семинара

Скачиваем тут

При подготовке семинара использовались

- материалы семинара Мурата Апишева по курсу МО в ШАД
- лекция Анны Потапенко
- лекция Анны Потапенко по курсу АНД на ФКН ПМИ