

Natural Language Processing

Яковенко Ольга

Natural Language Processing - NLP - Автоматическая обработка естественного языка

- ▶ Распознавание речи
- ▶ Поисковые системы
- ▶ Автоматическое исправление опечаток
- ▶ Обнаружение спама...

Data Science

Table of baby-name data
(baby-2010.csv)

name	rank	gender	year
Jacob	1	boy	2010
Isabella	1	girl	2010
Ethan	2	boy	2010
Sophia	2	girl	2010
Michael	3	boy	2010

Field names

One row
(4 fields)

2000 rows
all told

Объекты NLP

- ▶ Слово
- ▶ Фраза (поисковый запрос, ФИО, адрес, заголовок, ...)
- ▶ Текст
- ▶ Звук



Lorem ipsum

Dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam

Токенизация

Строка -> набор токенов (П: предложение -> слова)

‘Привет, мир!’ → [‘Привет’, ‘,’, ‘мир’, ‘!’]

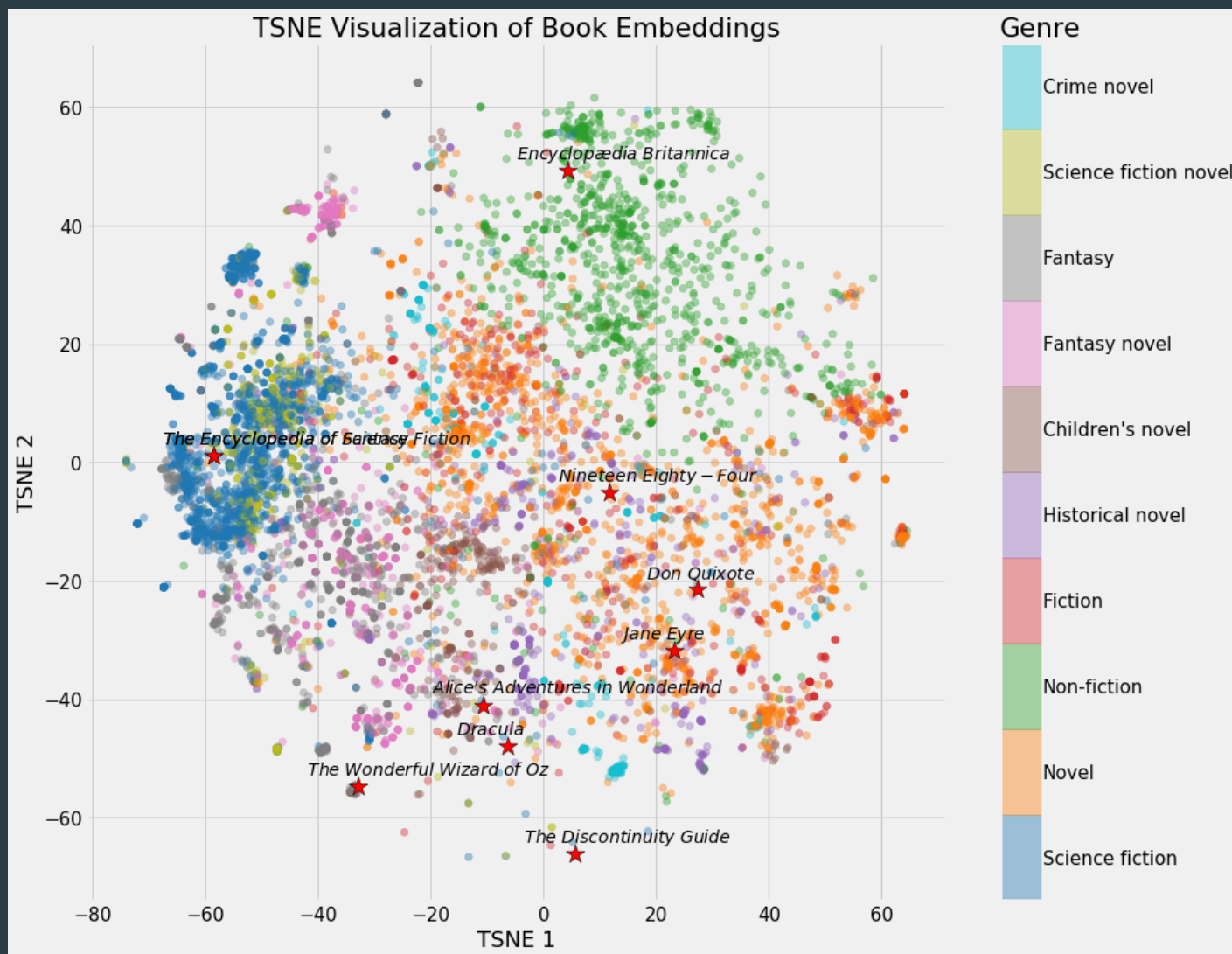
`nltk.word_tokenize`

Векторные представления (embeddings)

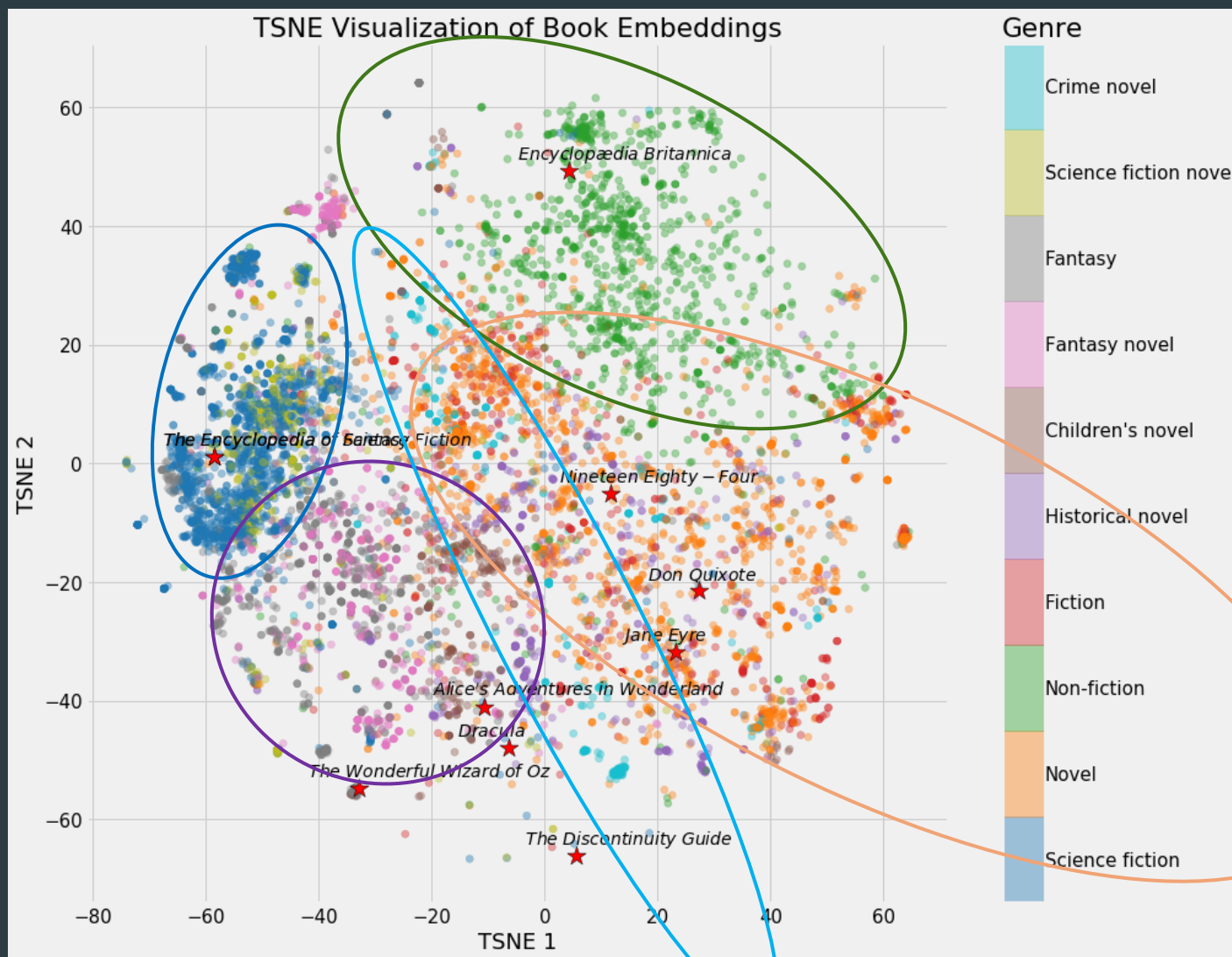
Результат трансформирования текстовых данных
в векторное пространство

‘Привет, мир!’ → [0 1 3 8 2 9 0 7]

Векторные представления (embeddings)



Векторные представления (embeddings)



Bag of Words (BoW) или «мешок слов»

['я еду',
'медленно по шоссе еду',
'я еду еду еду по Бердскому шоссе']



я	медленно	еду	по	Бердскому	шоссе
1	0	1	0	0	0
0	1	1	1	0	1
1	0	3	1	1	1

`sklearn.feature_extraction.text.CountVectorizer`

Tf-idf (*term frequency-inverse document frequency*)

Большой вес в TF-IDF получают слова:

с высокой частотой в
пределах конкретного
документа

&

с низкой частотой
употреблений в других
документах

`sklearn.feature_extraction.text.TfidfVectorizer`

Tf-idf (*term frequency-inverse document frequency*)

$$\text{tf}(t, d) = \frac{n_t}{\sum_k n_k}$$

Сколько раз слово
встретилось в
рамках одного
документа

Количество документов
(текстов) в датасете

$$\text{idf}(t, D) = \log \frac{|D|}{|\{d_i \in D \mid t \in d_i\}|}$$

Число документов из датасета D ,
в которых встречается слово t .

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$

Произведение tf и idf

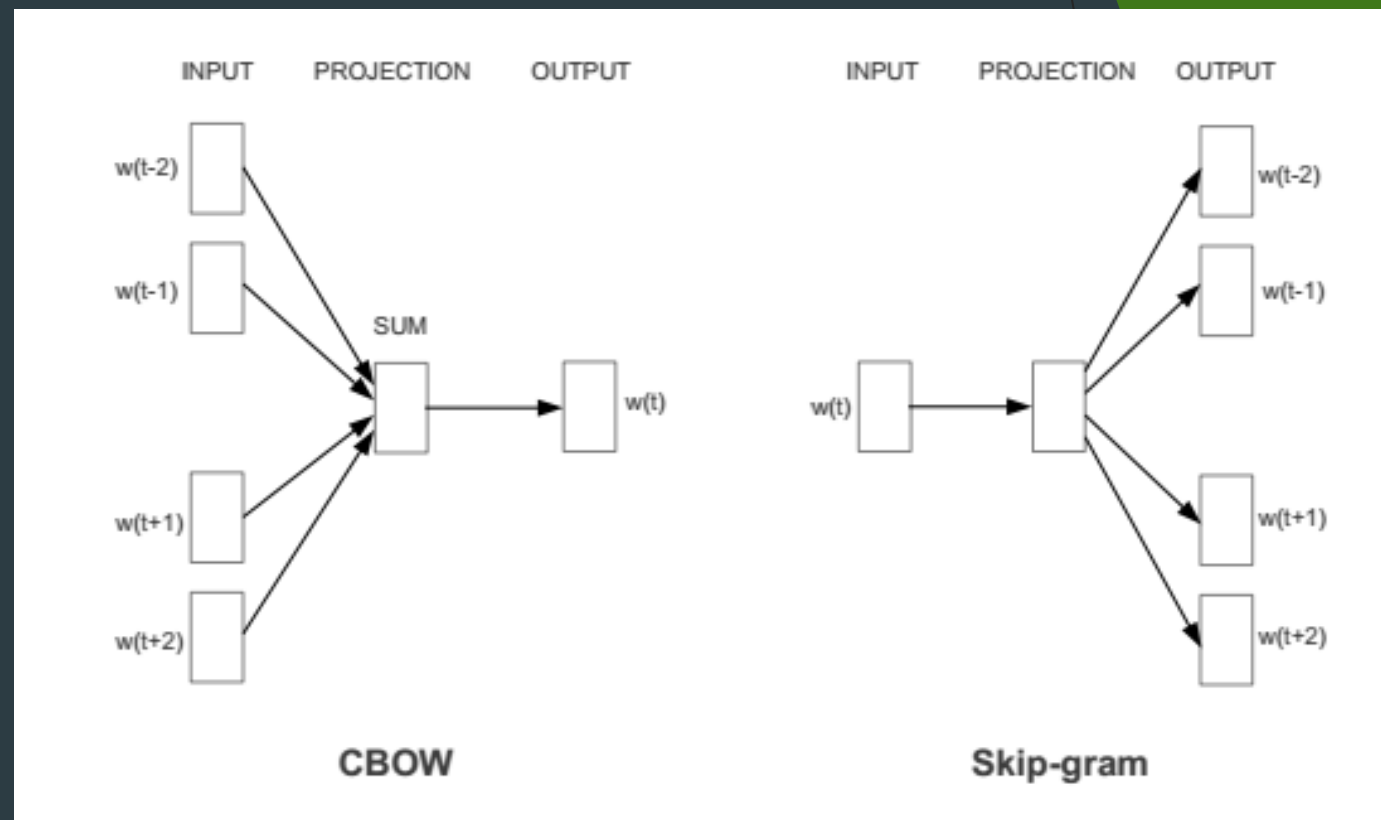
`sklearn.feature_extraction.text.TfidfVectorizer`

Векторные представления

- ▶ Word2Vec
- ▶ FastText
- ▶ ELMO
- ▶ BERT
- ▶ ULMFiT

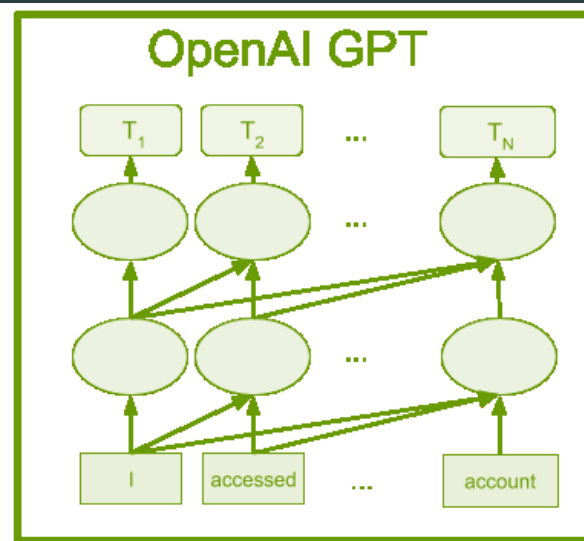
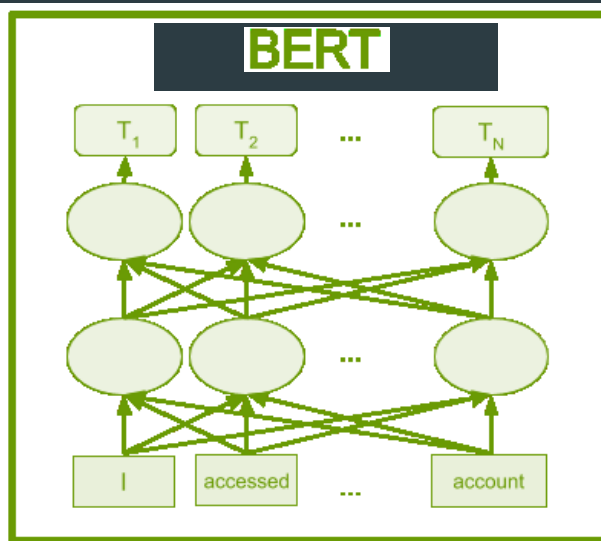
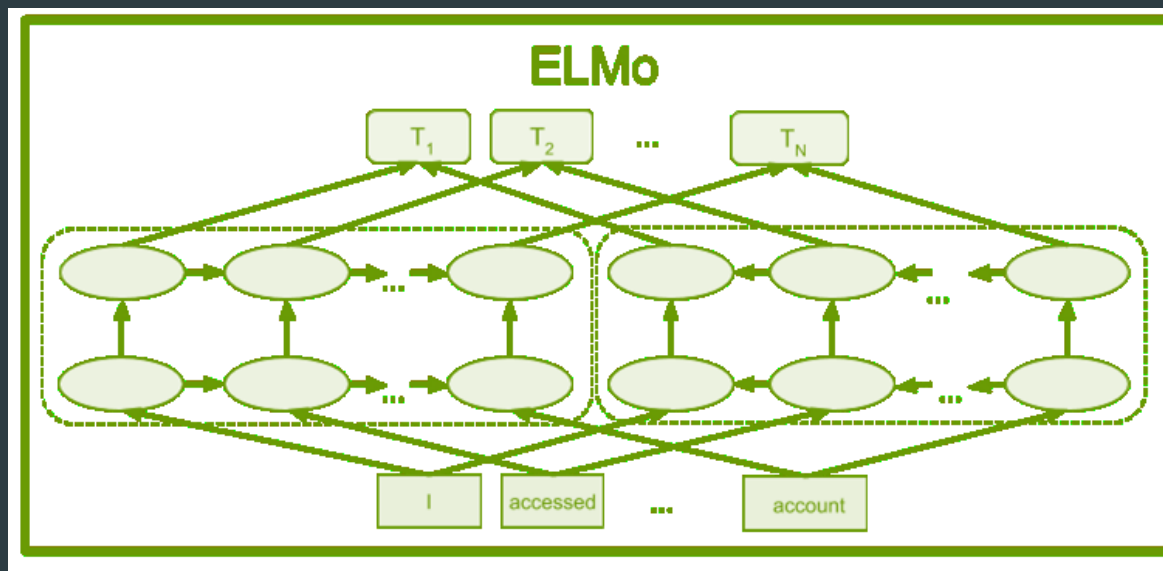
Векторные представления

- ▶ Word2Vec
- ▶ FastText
- ▶ ELMo
- ▶ BERT
- ▶ ULMFiT



Векторные представления

- ▶ Word2Vec
- ▶ FastText
- ▶ ELMo
- ▶ BERT
- ▶ OpenAI
- ▶ ULMFiT



Практика

https://github.com/DinoTheDinosaur/russian_sentiment_edu/blob/master/notebooks/Features_word_level.ipynb