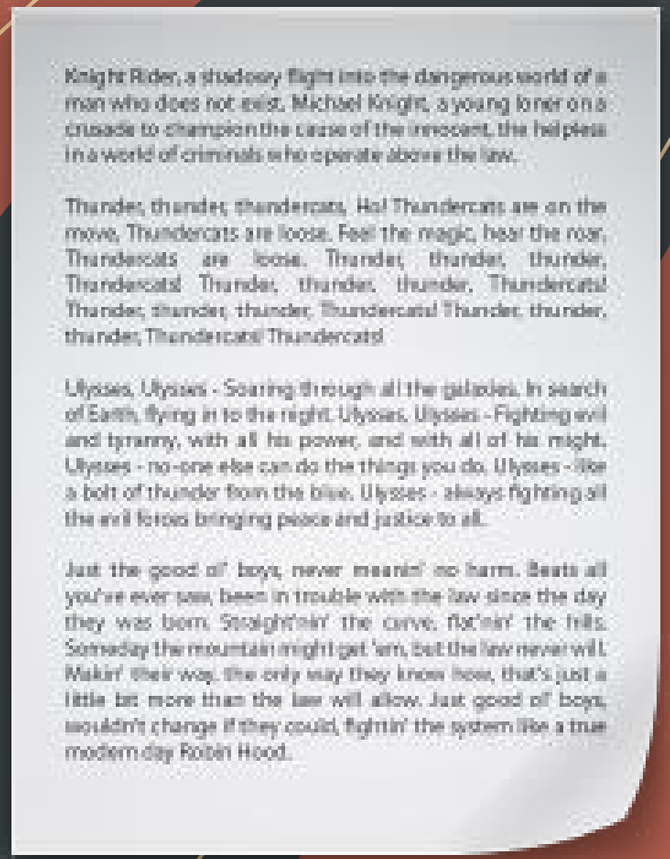
Natural Language Processing

Яковенко Ольга

Автоматическая обработка естественного языка

- направление искусственного интеллекта и математической лингвистики;

- изучает проблемы компьютерного анализа и синтеза человеческой речи.



Knight Rider, a shadowy flight into the dangerous world of a man who does not exist. Michael Knight, a young loner on a crusade to champion the cause of the innocent, the helpless in a world of criminals who operate above the law.

Thunder, thunder, thundercats, Ho! Thundercats are on the move, Thundercats are loose. Feel the magic, hear the roar, Thundercats are loose. Thunder, thunder, thunder, Thundercats! Thunder, thunder, thunder, Thundercats! Thunder, thunder, thunder, Thundercats! Thunder, thunder, thunder, Thundercats! Thundercats!

Ulysses, Ulysses - Soaring through all the galaxies. In search of Earth, flying in to the night. Ulysses, Ulysses - Fighting evil and tyranny, with all his power, and with all of his might. Ulysses - no-one else can do the things you do. Ulysses - like a bolt of thunder from the blue. Ulysses - always fighting all the evil forces bringing peace and justice to all.

Just the good ol' boys, never meanin' no harm. Beats all you've ever saw, been in trouble with the law since the day they was born. Straightnin' the curve, flat'nin' the hills. Someday the mountain might get 'em, but the law never will. Makin' their way, the only way they know how, that's just a little bit more than the law will allow. Just good ol' boys, wouldn't change if they could, fightin' the system like a true modern day Robin Hood.

Natural Language Processing – NLP –Автоматическая обработка естественного языка

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

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

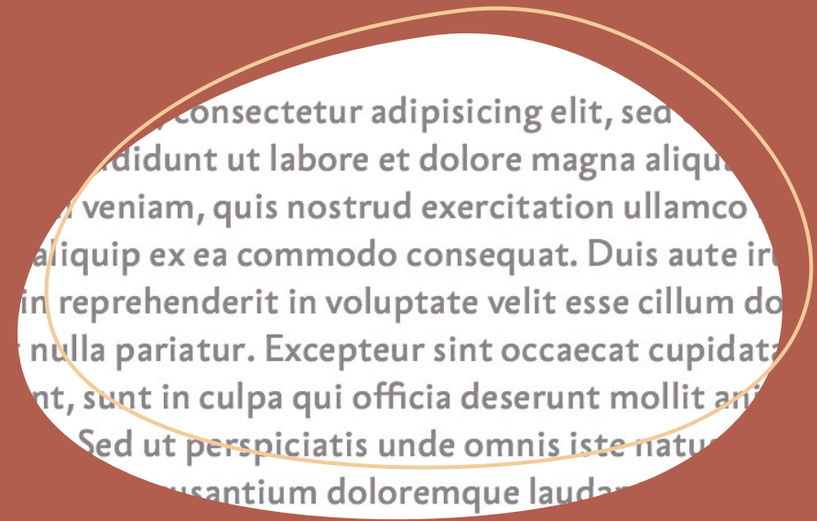
Field names

One row (4 fields)

2000 rows all told

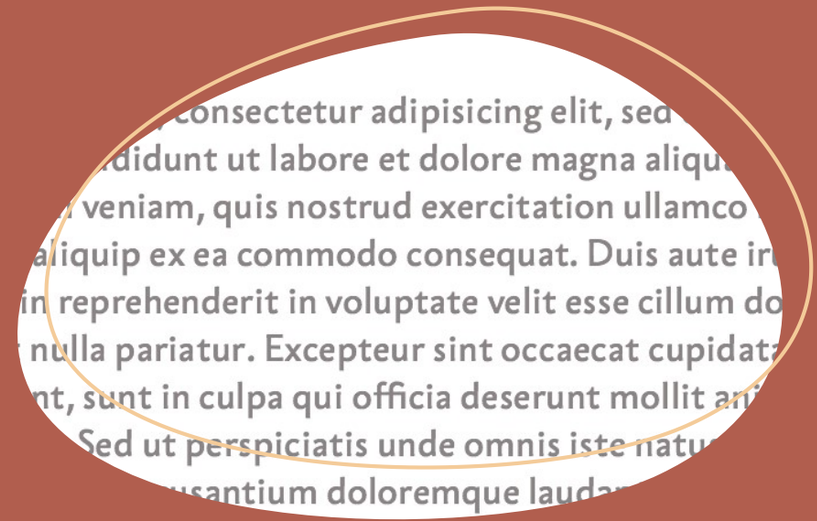
Анализ Данных

Объекты NLP


consectetur adipisicing elit, sed
didunt ut labore et dolore magna aliqu
veniam, quis nostrud exercitation ullamco
aliquip ex ea commodo consequat. Duis aute ir
in reprehenderit in voluptate velit esse cillum do
nulla pariatur. Excepteur sint occaecat cupidat
nt, sunt in culpa qui officia deserunt mollit an
Sed ut perspiciatis unde omnis iste natu
usantium doloremque laud

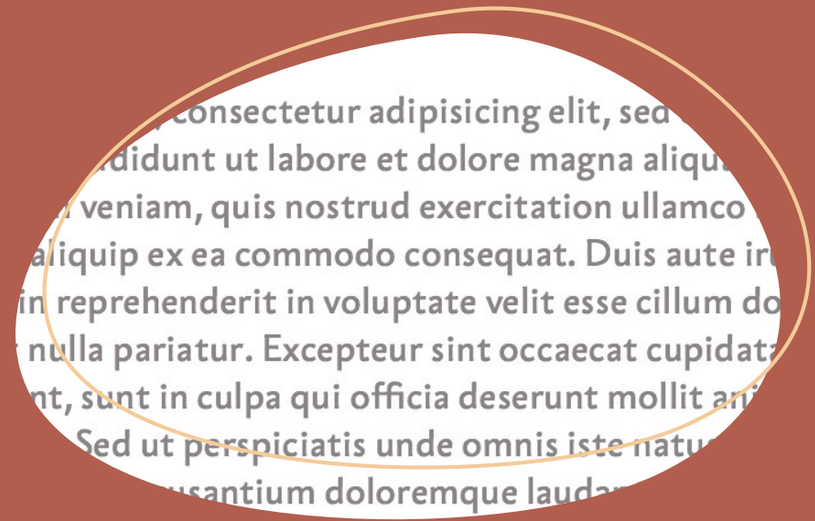
Объекты NLP

- Слово



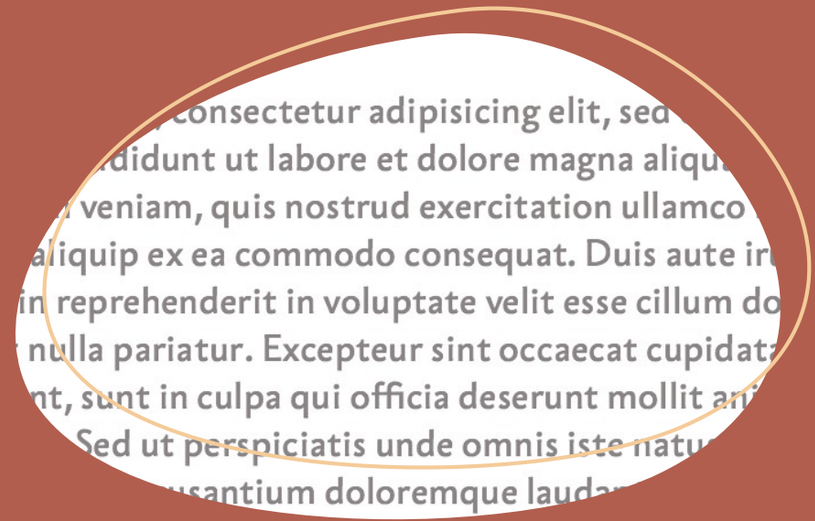
Объекты NLP

- Слово
- Фраза (поисковый запрос, ФИО, адрес, заголовок, …)



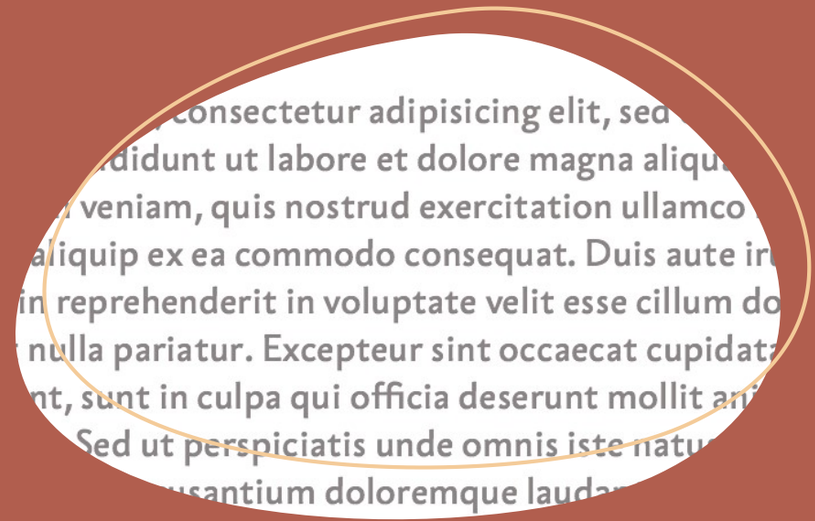
Объекты NLP

- Слово
- Фраза (поисковый запрос, ФИО, адрес, заголовок, …)
- Текст



Объекты NLP

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



Токенизация

Строка -> набор токенов

(П: предложение -> слова)

Токенизация

'Привет, мир!'

['Привет', ',', 'мир', '!']

Строка -> набор токенов

(П: предложение -> слова)

Токенизация

'Привет, мир!'

['Привет', ',', 'мир', '!']

Строка -> набор токенов

(П: предложение -> слова)

nltk.word_tokenize

Векторные представления (embeddings)

Векторные представления (embeddings)

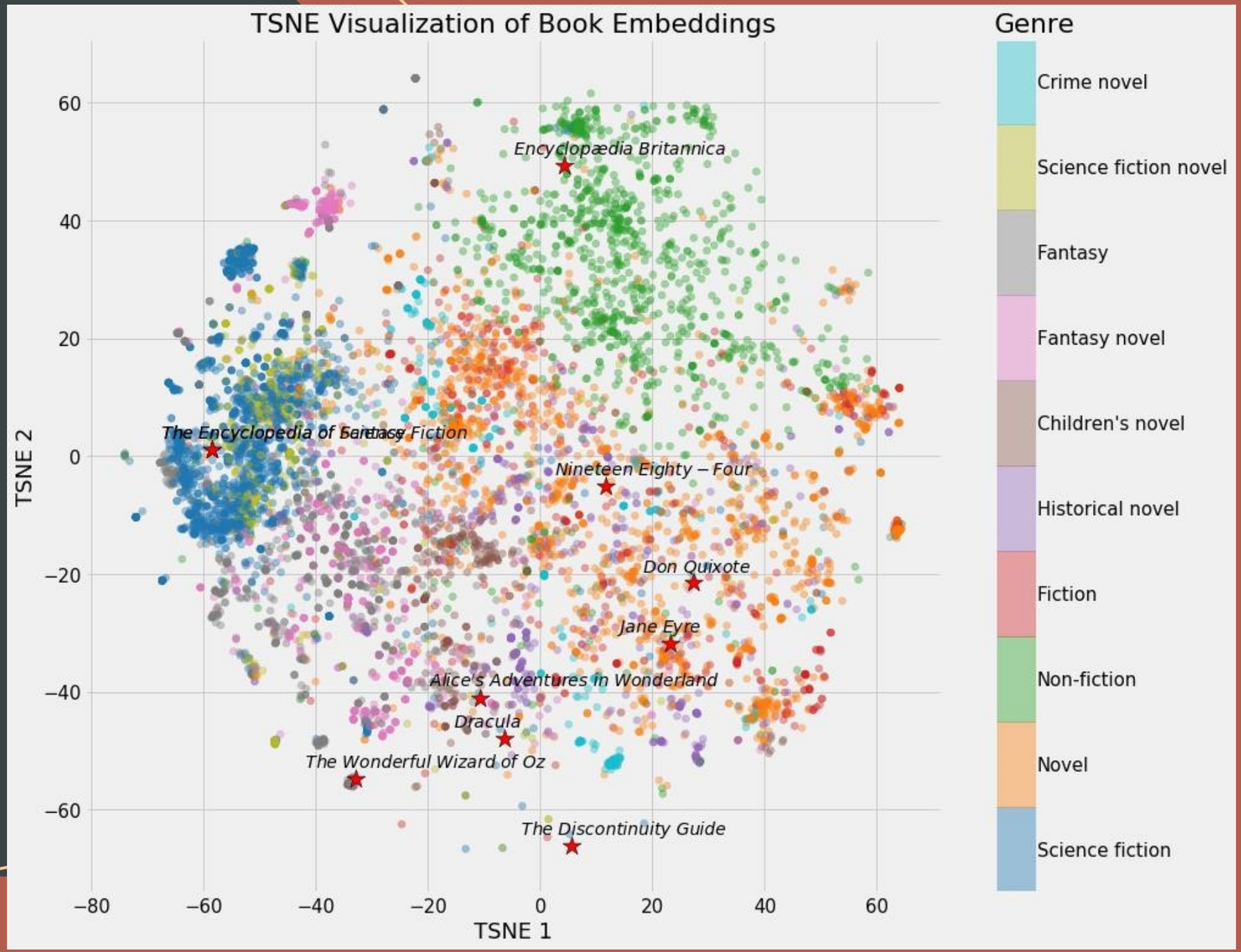
Результат трансформирования текстовых данных в векторное пространство

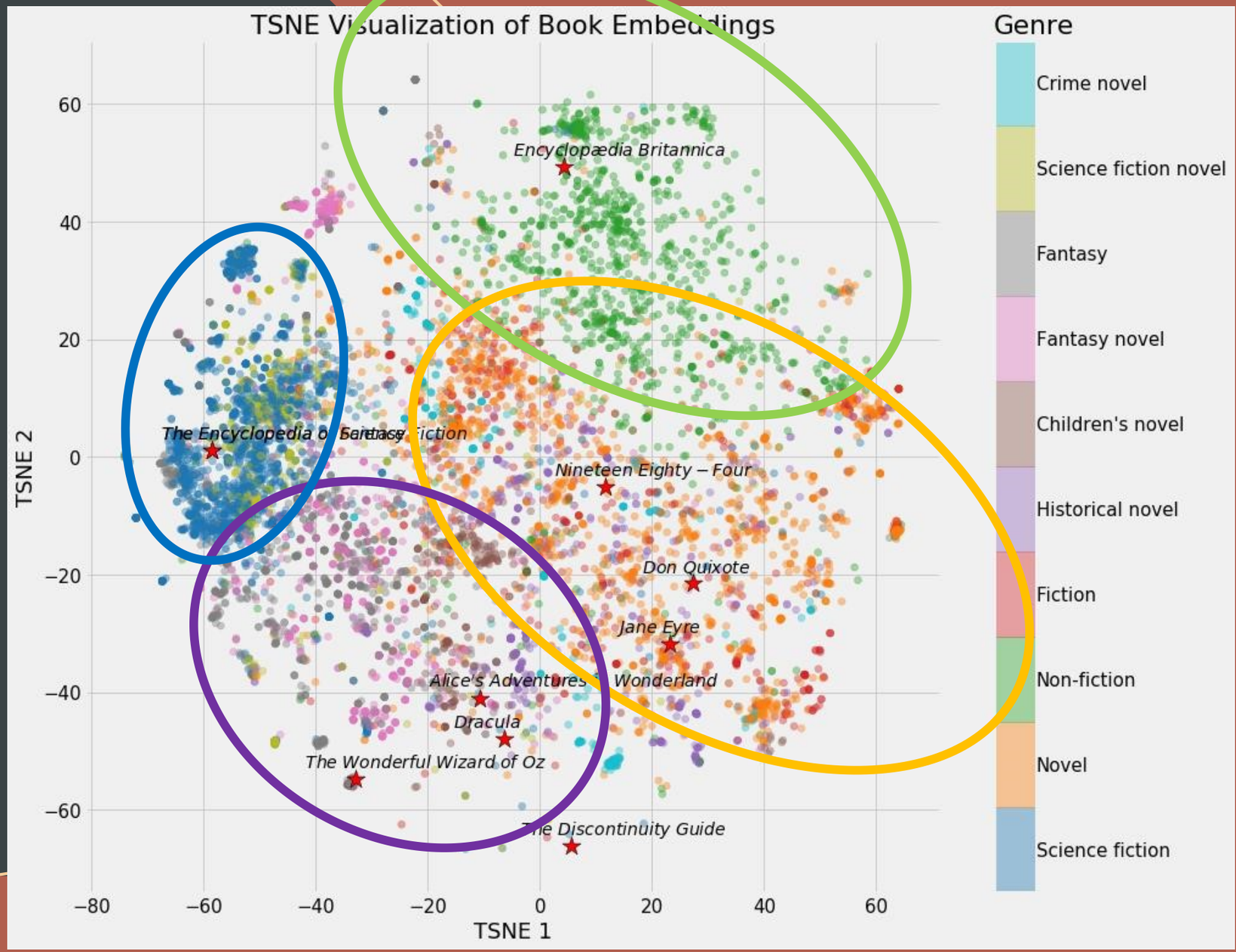
Векторные представления (embeddings)

Результат трансформирования текстовых данных в векторное пространство

'Привет, мир!'

[0 1 3 8 2 9 0 7]


TSNE Visualization of Book Embeddings
Genre
Crime novel
Science fiction novel
Fantasy
Fantasy novel
Children's novel
Historical novel
Fiction
Non-fiction
Novel
Science fiction
TSNE 1
TSNE 2
60
40
20
0
−20
−40
−60
−80
−60
−40
−20
0
20
40
60
Encyclopædia Britannica
The Encyclopedia of Fantasy
Nineteen Eighty – Four
Don Quixote
Jane Eyre
Alice's Adventures in Wonderland
Dracula
The Wonderful Wizard of Oz
The Discontinuity Guide


TSNE Visualization of Book Embeddings
Genre
Crime novel
Science fiction novel
Fantasy
Fantasy novel
Children's novel
Historical novel
Fiction
Non-fiction
Novel
Science fiction
Encyclopædia Britannica
The Encyclopedia of Fantasy
Nineteen Eighty – Four
Don Quixote
Jane Eyre
Alice's Adventures in Wonderland
Dracula
The Wonderful Wizard of Oz
The Discontinuity Guide
TSNE 1
TSNE 2
60
40
20
0
−20
−40
−60
−80
−60
−40
−20
0
20
40
60

Bag of Words (BoW) или «мешок слов»

Bag of Words (BoW) или «мешок слов»

[‘я еду’,

‘медленно по шоссе еду’,

‘я еду еду еду по Бердскому шоссе’]

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

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)

Tf-idf (term frequency-inverse document frequency)

Большой вес в TF-IDF получат слова:

Tf-idf (term frequency-inverse document frequency)

Большой вес в TF-IDF получат слова:

с высокой частотой в пределах одного документа

Tf-idf (term frequency-inverse document frequency)

Большой вес в TF-IDF получат слова:

с высокой частотой в пределах одного документа

&

с низкой частотой употреблений в других документах

Tf-idf (term frequency-inverse document frequency)

Большой вес в TF-IDF получат слова:

с высокой частотой в пределах одного документа

&

с низкой частотой употреблений в других документах

sklearn.feature_extraction.text.TfidfVectorizer

Tf-idf (term frequency-inverse document frequency)

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

Сколько раз слово встретилось в рамках одного документа

Количество документов (текстов) в датасете

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

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

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

Произведение tf и idf

sklearn.feature_extraction.text.TfidfVectorizer

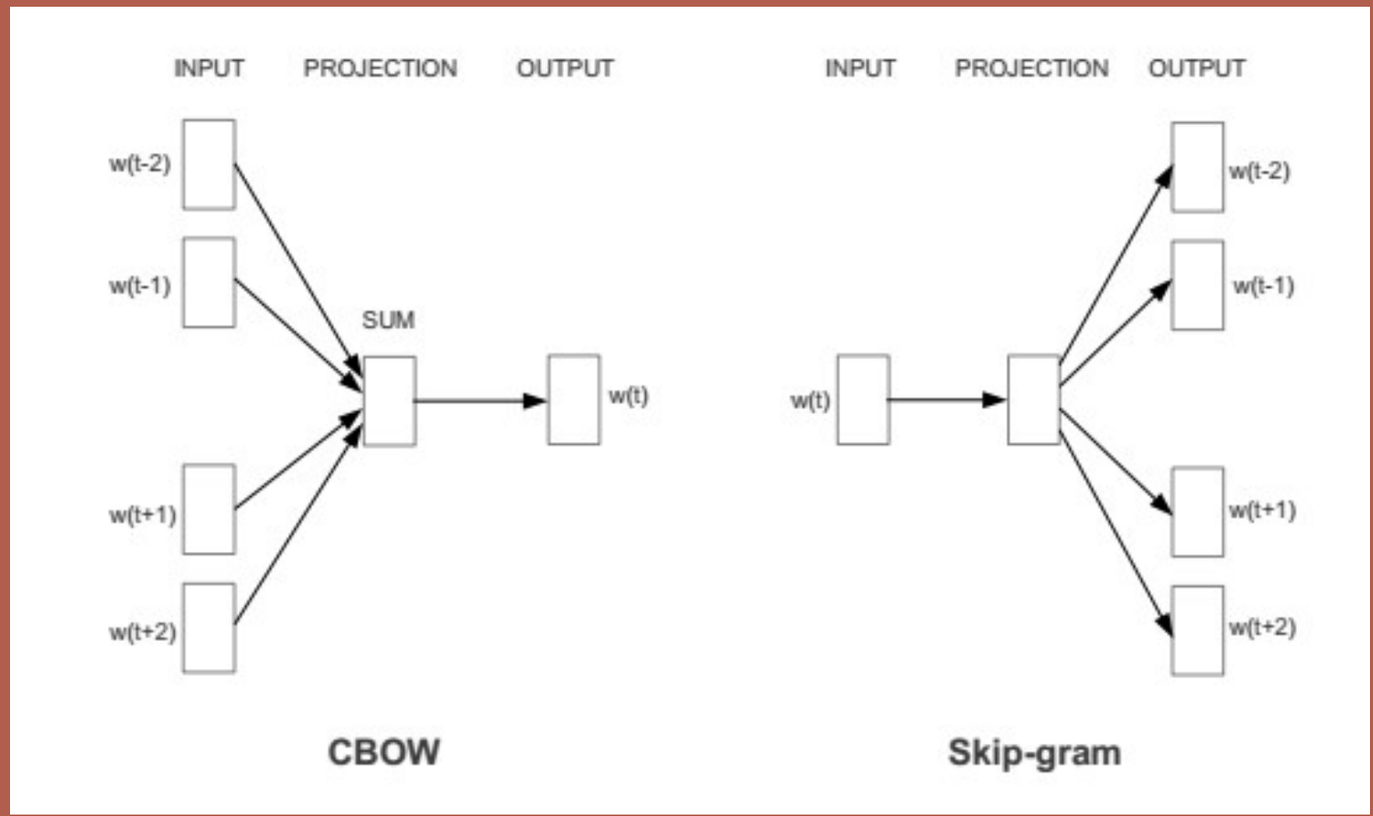
Я очень люблю конфеты! Любить конфеты - моё призвание.	А я ем фрукты вместо конфет. Я на диете.	Кто-то в этом мире ненавидит конфеты. А я ненавижу этих людей.

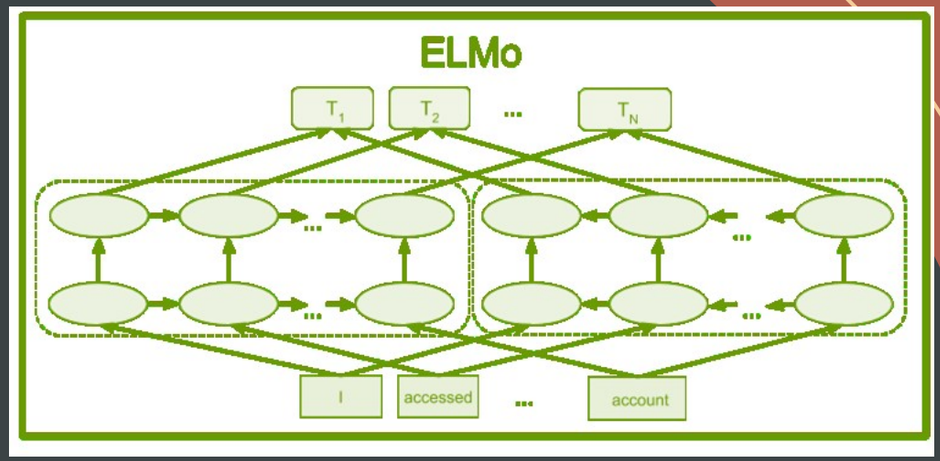
Векторные представления

- Word2Vec
- FastText
- ElMO
- BERT
- ULMFiT

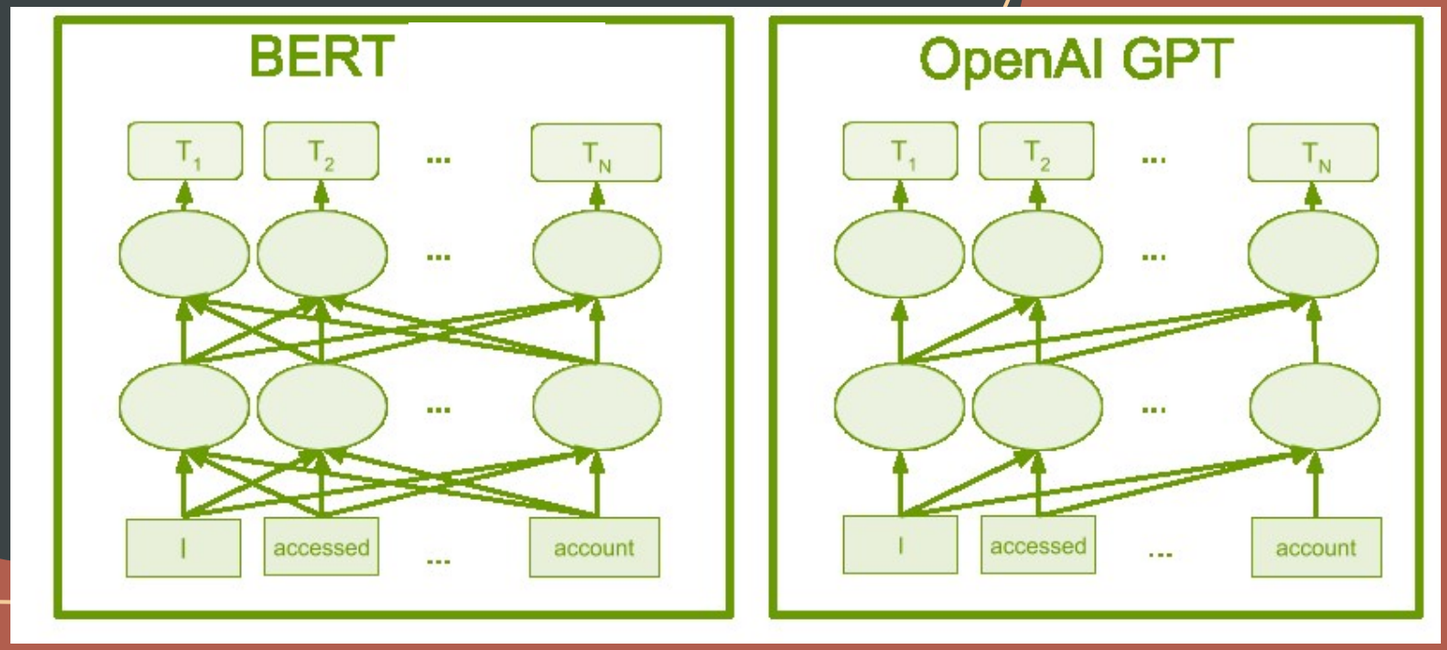
Векторные представления

- Word2Vec
- FastText
- ElMO
- BERT
- ULMFiT





Векторные представления



- Word2Vec
- FastText
- ElMO
- BERT
- ULMFiT
- GPT

SentiRuEval_2016

- Формат xml
- Train–10725 твитов
  - Нейтральные (класс 0): 7158
  - Отрицательные (класс -1): 2807
  - Положительные(класс 1): 760
- Test-3418 твитов
- Метрики соревнования: F1 micro & F1 macro по классам -1 и 1
- Использовать колонки ‘text’в качестве признаков, ‘answer’в качестве меток класса.

http://www.dialog-21.ru/evaluation/2016/sentiment/

Практика

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