

O · T U S

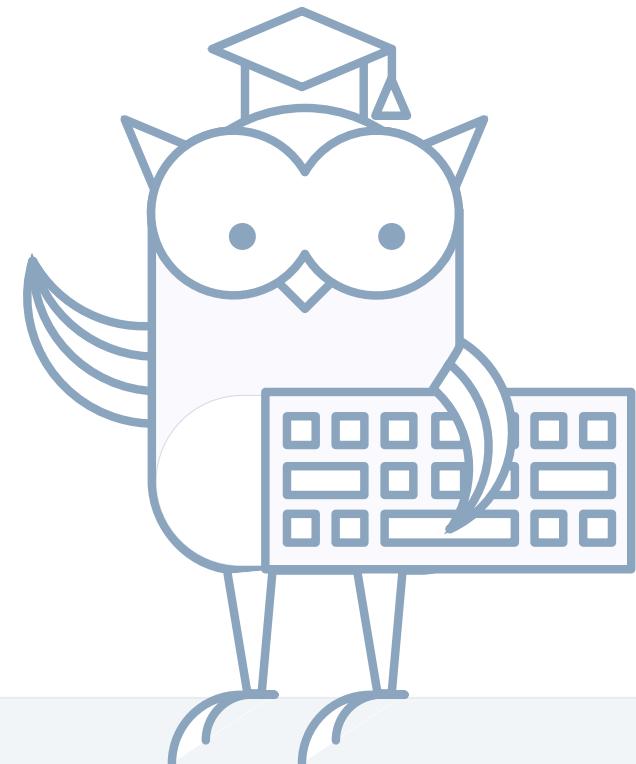
ОНЛАЙН-ОБРАЗОВАНИЕ



Анализ текстовых данных

Тематическое моделирование ВК

О чём говорят любители рэпа?



- Познакомиться с задачей тематического моделирования
- Узнать об основных этапах работы с текстовыми данными
- Получить представление о самой популярной модели – LDA и её реализации в Python
- Посмотреть на расширенные возможности тематического моделирования

- Зачем это нужно?
- Как это делать?
- Что из этого получается?

01

Зачем нужно тематическое моделирование?



Donald J. Trump

@realDonaldTrump

Follow



why would Kim Jong-un insult me by calling me "old," when I would NEVER call him short and fat? Oh well, I try so hard to be his friend - and maybe someday that will happen!

7:48 PM - 11 Nov 2017 from Vietnam

96,010 Retweets **215,716** Likes



68K

96K



216K



<http://maketrump tweetseightagain.com>

Зачем?

OTUS



- Быстро понять, какие темы поднимаются в тексте
- Выделить наиболее обсуждаемые топики в комментариях
- Лучше узнать свою аудиторию для таргетирования
- Находить сообщества, которые интересуются нужными темами
- Как это сделать?

02

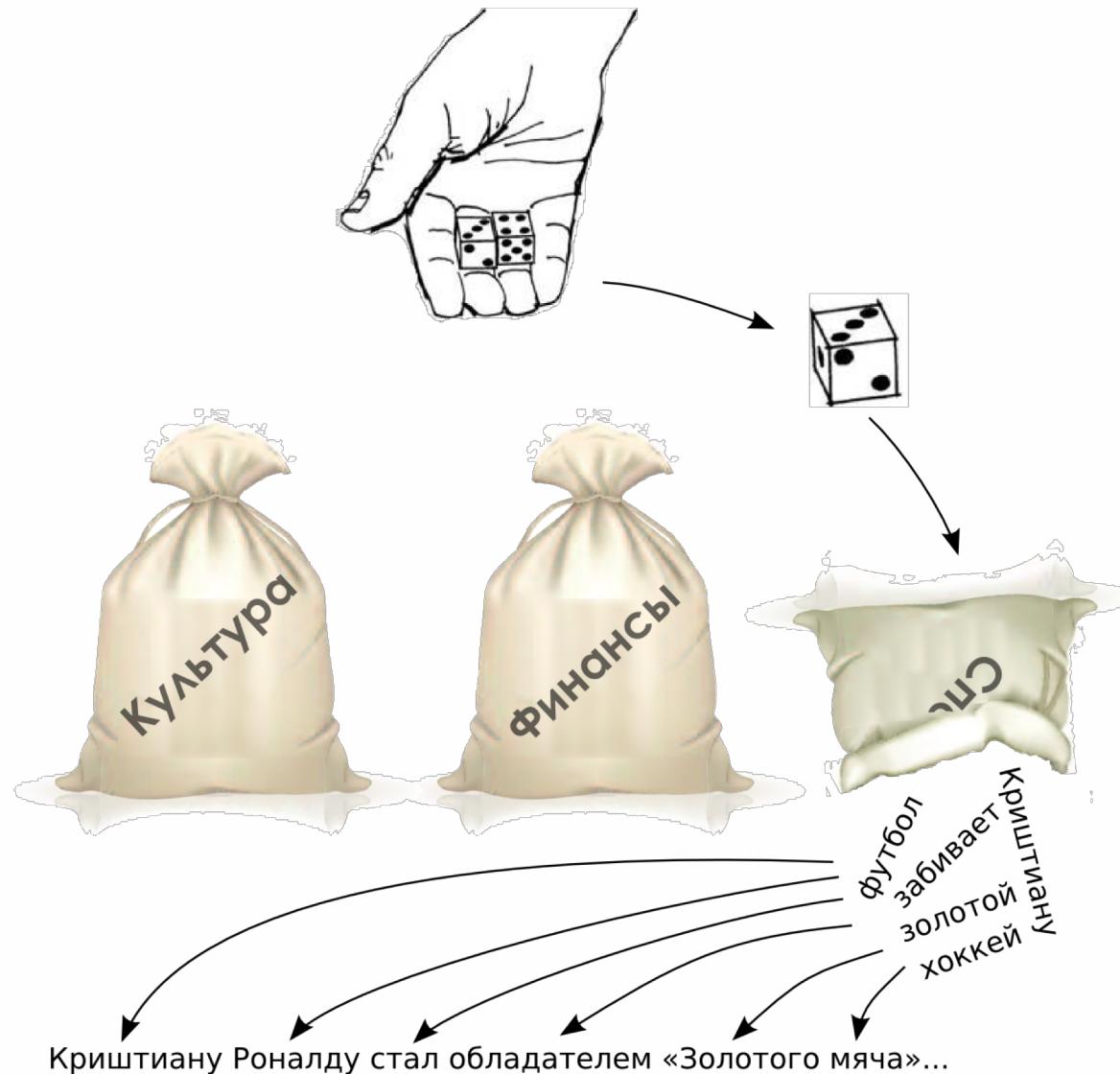
Тематическое моделирование на пальцах



<https://habr.com/ru/company/surfingbird/blog/228249/>

Происхождение текста

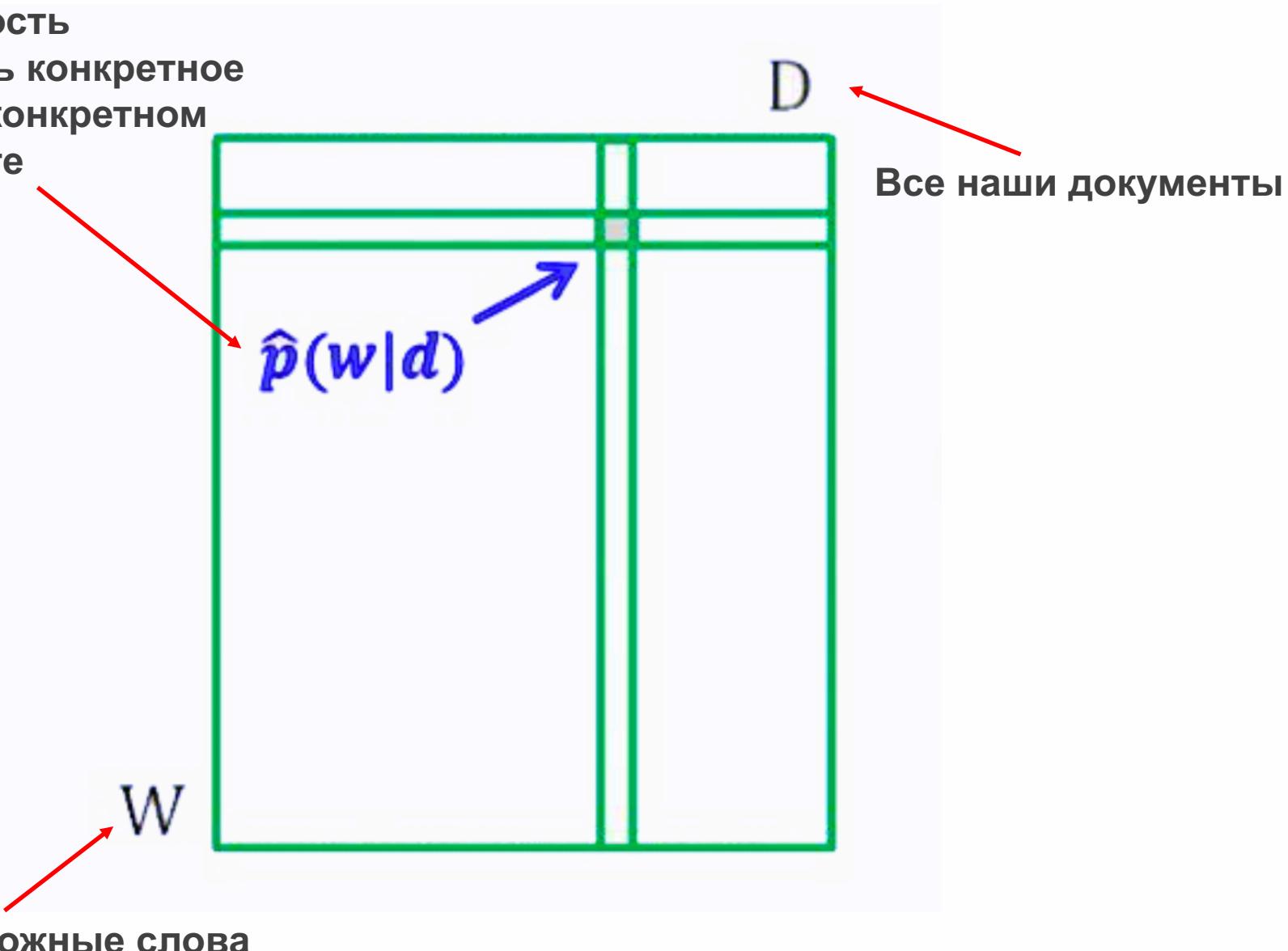
O T U S



Немножко определений

- Будем называть каждый текст **документом, d**
- Каждый документ состоит из нескольких **тем, t**
- Каждая тема описывается некоторым набором **слов, w**

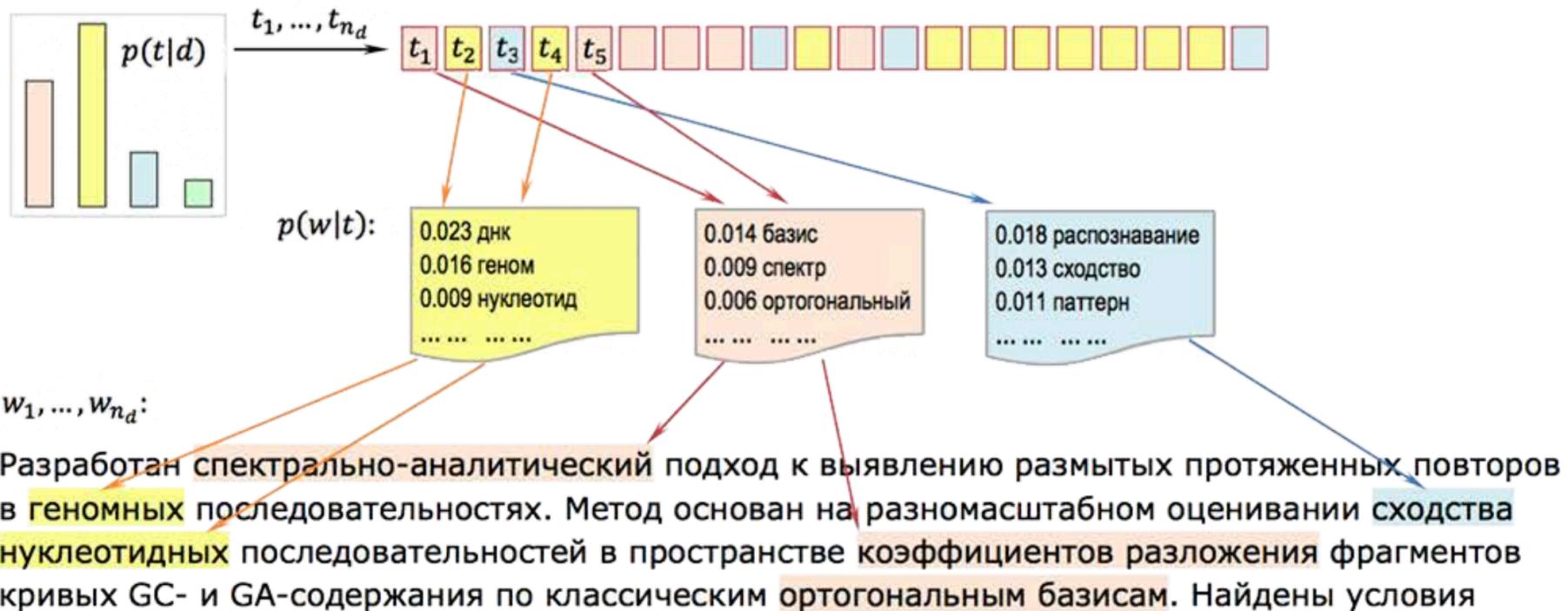
Вероятность
встретить конкретное
слово в конкретном
документе



Все возможные слова

Мы хотели бы научиться выделять в каждом документе темы

- Пусть изначально тем было T штук: t_1, \dots, t_n
- Для каждого документа есть распределение тем: $p(t|d) = \theta_{td}$
- А для каждой темы – распределение слов в ней: $p(w|t) = \phi_{td}$



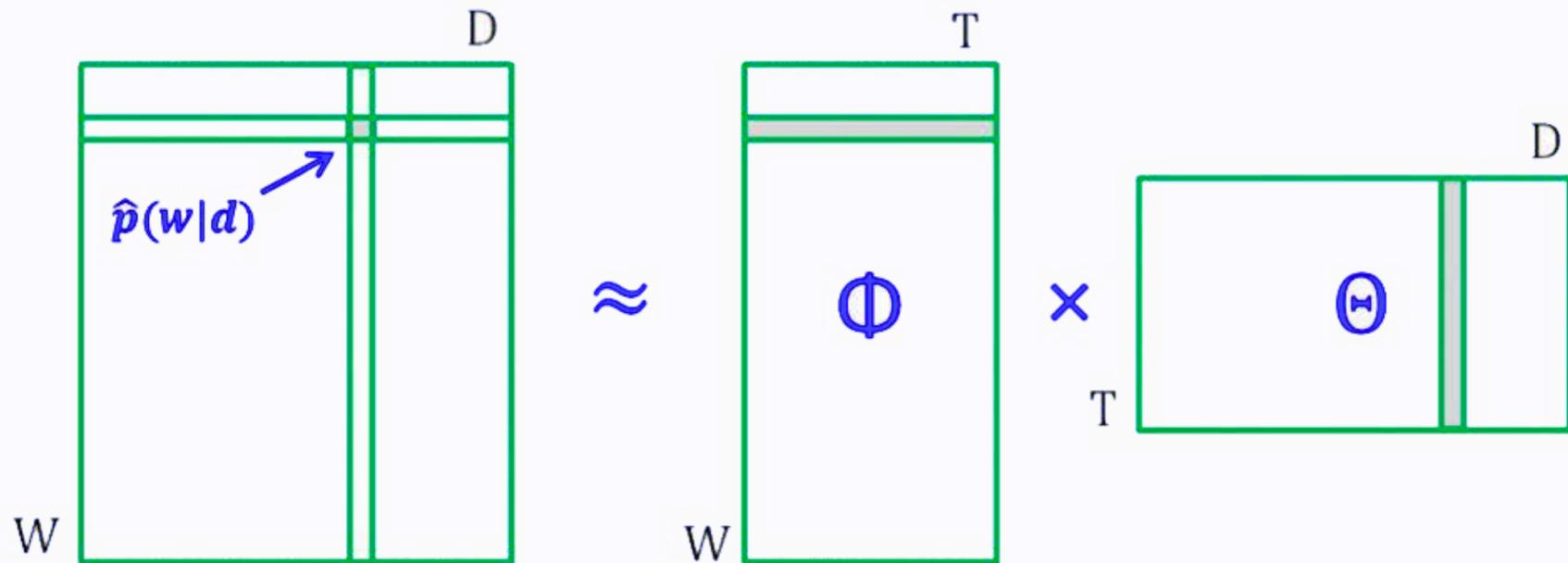
<http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>

Итого:

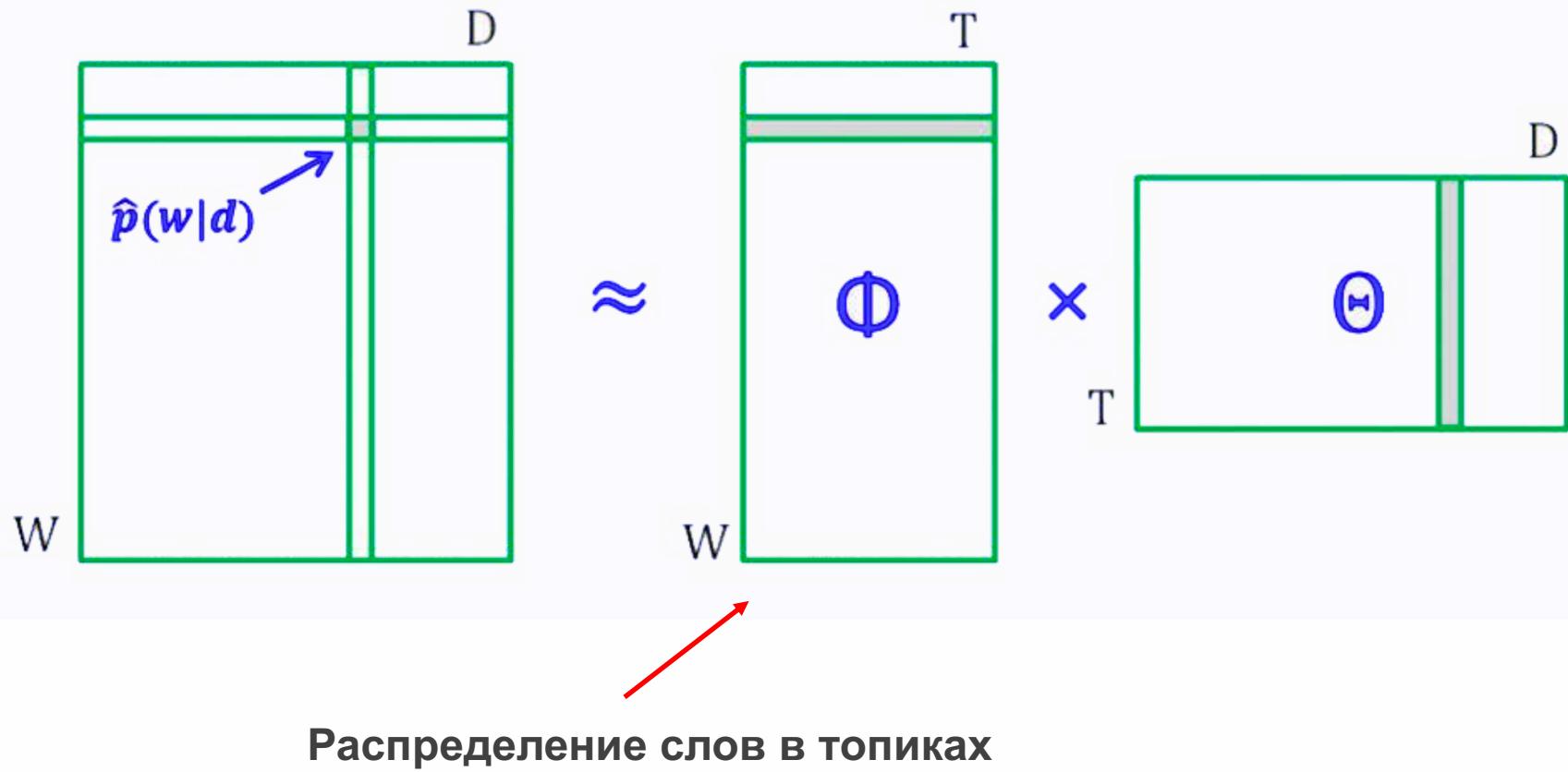
- Слово может встретиться в документе, если оно пришло из темы t_1 , или t_2 , или ..., t_n
- Всего тем T , значит, есть T возможных вариантов/гипотез возникновения слова
- Тогда какой будет полная вероятность встретить слово в документе?

$$p(w|d) = \sum_{t \in T} p(w|t) \cdot p(t|d) = \sum_{t \in T} \phi_{wt} \cdot \theta_{td}$$

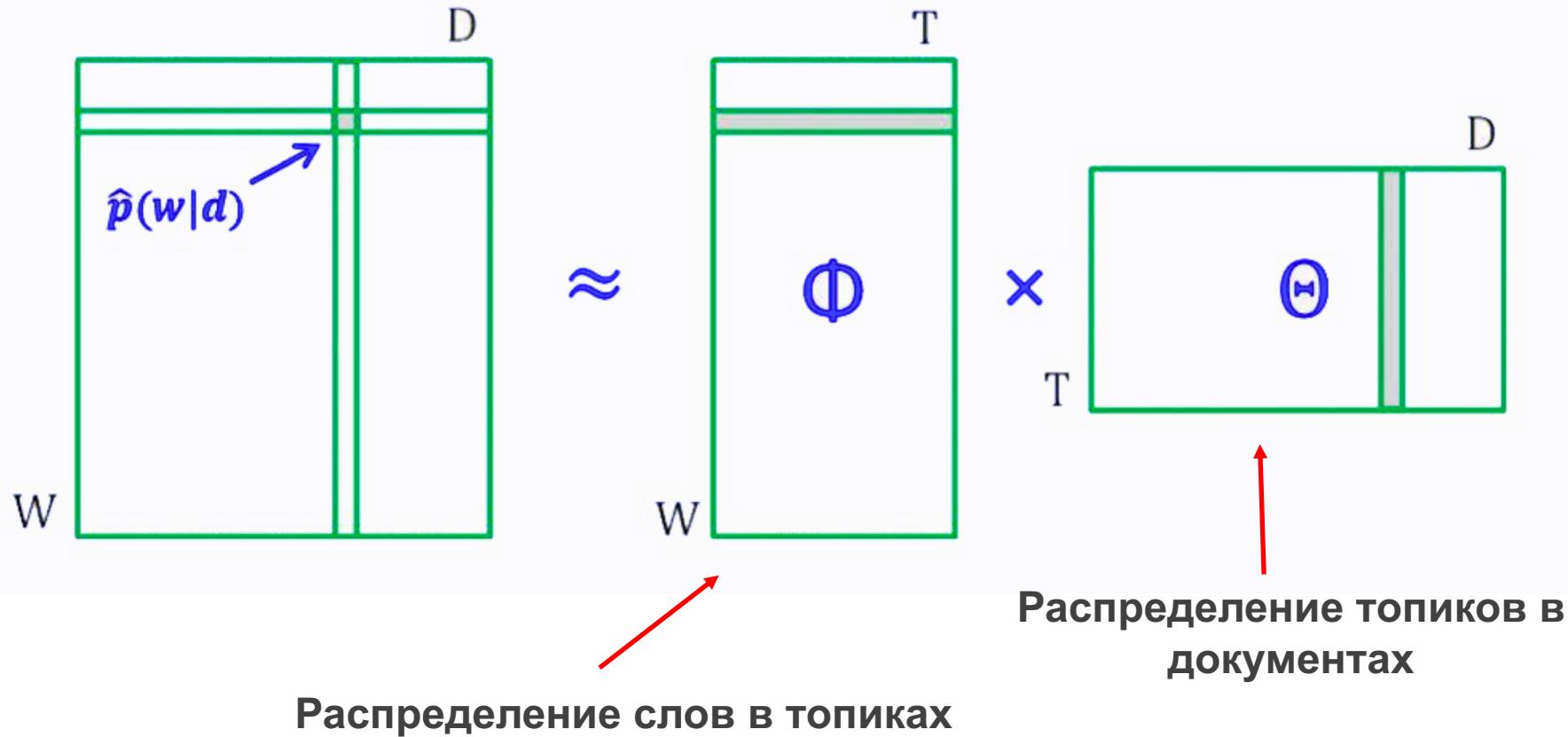
$$p(w|d) = \sum_{t \in T} p(w|t) \cdot p(t|d) = \sum_{t \in T} \phi_{wt} \cdot \theta_{td}$$



$$p(w|d) = \sum_{t \in T} p(w|t) \cdot p(t|d) = \sum_{t \in T} \phi_{wt} \cdot \theta_{td}$$



$$p(w|d) = \sum_{t \in T} p(w|t) \cdot p(t|d) = \sum_{t \in T} \phi_{wt} \cdot \theta_{td}$$



Итак, нужно оценить неизвестные вероятности в матрицах

- Имеем задачу **матричного разложения**
- Множество методов решения – от аналитических, до градиентного спуска
- Однако найденное решение не будет единственным!

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi^1\Theta^1$$

- Поэтому вводятся дополнительные ограничения - **регуляризация**

03

Латентное Размещение Дирихле

Самая знаменитая регуляризация - **регуляризация Дирихле**

- Придумана в 2003 году и с тех пор является самой популярной
- Модель, использующая эту регуляризацию – **Latent Dirichlet Allocation**, a.k.a. LDA
- Для Python отличная реализация в библиотеке **gensim**, также есть версия в `sklearn`

<https://radimrehurek.com/gensim/models/ldamodel.html>

[https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.
LatentDirichletAllocation.html](https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html)

04

Тематическое моделирование VK

- Выяснить, различаются ли темы для общения у слушателей разных жанров
- Понять, можно ли с выделенными темами сделать что-то еще

Зачем?

O T U S



- Потенциально: исследование аудитории ваших подписчиков
- Поиск похожих по тематикам обсуждения сообществ
- Таргетирование рекламы
- Проведение социологических исследований

- Найти жанры

×

Весь список

Под настроение

Для занятий

Поп

Инди

Рок

Метал

Альтернатива

Электроника

Танцевальная

Рэп и хип-хоп

R&B

Джаз

Блюз

Регги

Ска

Панк

Классика

Эстрада

Шансон

Кантри

Авторская песня

Лёгкая музыка

Саундтреки

Музыка мира

Детская

- Найти исполнителей по жанрам

The screenshot shows a dark-themed interface for a music platform. At the top left is the genre name 'Шансон'. To the right are two buttons: a yellow one labeled 'Радио жанра' with a radio icon, and a white one with a heart icon. Below these are five navigation tabs: 'ОБЗОР', 'ТРЕКИ', 'АЛЬБОМЫ', 'ИСПОЛНИТЕЛИ' (which is underlined in yellow), and 'КОНЦЕРТЫ'. A section titled 'Популярные исполнители' follows, featuring four circular profile pictures of artists: Mikhail Krug, Irina Krug, Sergey Trofimov, and Butyrka. Each artist's name and genre ('шансон') are listed below their respective photos.

Исполнитель	Жанр
Михаил Круг	шансон
Ирина Круг	шансон
Сергей Трофимов	шансон
Бутырка	шансон

- Найти группы исполнителей в VK и скачать все комментарии

ГРУППА "БУТЫРКА" • Русский Шансон
Поклонники творчества группы "БУТЫРКА", ПОДПИСЫВАЙТЕСЬ 

 ГРУППА "БУТЫРКА" • Русский Шансон запись закреплена
30 апр в 17:49

Нас 95.000!
Уважаемые подписчики, спасибо, что Вы с нами!
ЛЕГЕНДАРНЫЕ ХИТЫ "ЗОЛОТОГО" СОСТАВА ГРУППЫ "БУТЫРКА" ДЛЯ ВАС!



- Подробно о парсинге – на занятии «Сбор данных» курса Machine Learning

[Ссылка на код с парсером](#)

- 11 жанров
- 134 исполнителя
- 10 миллионов комментариев



Предобработка

ОУС

```
import pandas as pd

# посмотрим на получившийся датасет
data = pd.read_csv('data/example.csv', index_col=0)
data.head()
```

Out [1]:

	comment_id	date	dirty_text	emoji	likes	link	music_style	performer	stickers
	52228413_282387_282388	2018-12-31 21:40:35	Шикарный подарок , песня замечательная !!!! Сп...		NaN	21	NaN	estrada	лепс
	52228413_282387_282390	2018-12-31 21:53:36	Самый лучший подарок - закончить год под песни...		NaN	15	NaN	estrada	лепс
	52228413_282387_282399	2018-12-31 22:40:49	Песни хит. Уверен, что получит много наград в ...		NaN	12	NaN	estrada	лепс

Во время предобработки нам нужно:

- Привести слова к начальной форме (лемматизация)
- Токенизировать предложения (мешок слов/bag-of-words)
- Убрать слишком короткие комментарии, не несущие смысла
- Добавить биграммы
- Выкинуть слишком частые/редкие слова и словосочетания

Очистка от пунктуации

ОТУС

```
import re

def strip_punctuation(string):
    return re.sub(r'[^w\s]', '', string)

data['text'] = data['dirty_text'].apply(strip_punctuation)
data.head(1)
```

Out [1]:

comment_id	date	dirty_text	emoji	likes	link	music_style	performer	stickers	text
52228413_282387_282388	2018-12-31 21:40:35	Шикарный подарок , песня замечательная !!!! Сп...		NaN	21	NaN	estrada	лепс	шикарный подарок песня замечательный спасибооо...

```
from pymystem3 import Mystem

m = Mystem()

# применяем лемматизацию
data['text_bow'] = data['text'].apply(lambda x: m.lemmatize(x))
data.head(1)

# заодно посчитаем длину комментариев
data['comment_len'] = data['text_bow'].apply(len)
```

Out [1]:

dirty_text	emoji	likes	link	music_style	performer	stickers	text	video	text_bow	comment_len
Шикарный подарок , песня замечательная !!!! Сп...	NaN	21	NaN	estrada	лепс	NaN	шикарный подарок песня замечательный спасибооо...	NaN	['шикарный', 'подарок', 'песня', 'замечательны...']	8

<https://github.com/nlpub/pymystem3>

```
from gensim import corpora, models

# создаём объект для генерации биграм
# min_count и threshold нужны для обрезки слишком редких
bigram = models.Phrases(data.text_bow, min_count=3, threshold=5)

# оборачиваем объект для быстрой обработки текста
bigram_mod = models.phrases.Phraser(bigram)

def make_bigrams(texts):
    return [bigram_mod[doc] for doc in texts]

# и генерируем текст с биграммами
texts = make_bigrams(data.text_bow)
```

```
from gensim import corpora, models

# составляем словарь из терминов
dictionary = corpora.Dictionary(texts)

# и убираем слишком редкие (no_below) и слишком частые (no_above)
dictionary.filter_extremes(no_below=3, no_above=0.4)

# наконец, составляем финальный готовый корпус
corpus = [dictionary.doc2bow(text) for text in texts]

# и не забываем сохранить
import pickle
with open('lda_models/corpus', 'wb') as f:
    pickle.dump(corpus, f)
```

```
from gensim import corpora, models

# ждём примерно три часа
ldamodel = models.ldamodel.LdaModel(
    corpus=corpus,
    id2word=dictionary,
    num_topics=30,
    passes=5
)

# не забываем сохранить!
ldamodel.save("lda_models/ldamodel_30_bigrams_long_comments")

# для каждого топика берём наиболее вероятный топ слов
topics = ldamodel.show_topics(
    num_topics=30,
    num_words=100,
    formatted=False
)

# и тоже сохраняем
with open('lda_models/lda_30_topics_bigrams_long_comments', 'wb') as f:
    pickle.dump(topics, f)
```

- При сохранении модели нужно отдельно сохранять атрибут **expElogbeta**
- При подгрузке модели отдельно его подгружать и присваивать
- Иначе будет ругаться и придётся переучивать модель

```
with open('lda_models/ldamodel_30_bigrams_long_comments', 'rb') as f:  
    ldamodel_30 = pickle.load(f)
```

```
with open('lda_models/corpus', 'rb') as f:  
    corpus = pickle.load(f)
```

```
expElogbeta = np.load("lda_models/ldamodel_30_bigrams_long_comments.expElogbeta.npy")  
ldamodel_30.expElogbeta = expElogbeta
```

05

Результаты

<https://bit.ly/308k1bt>



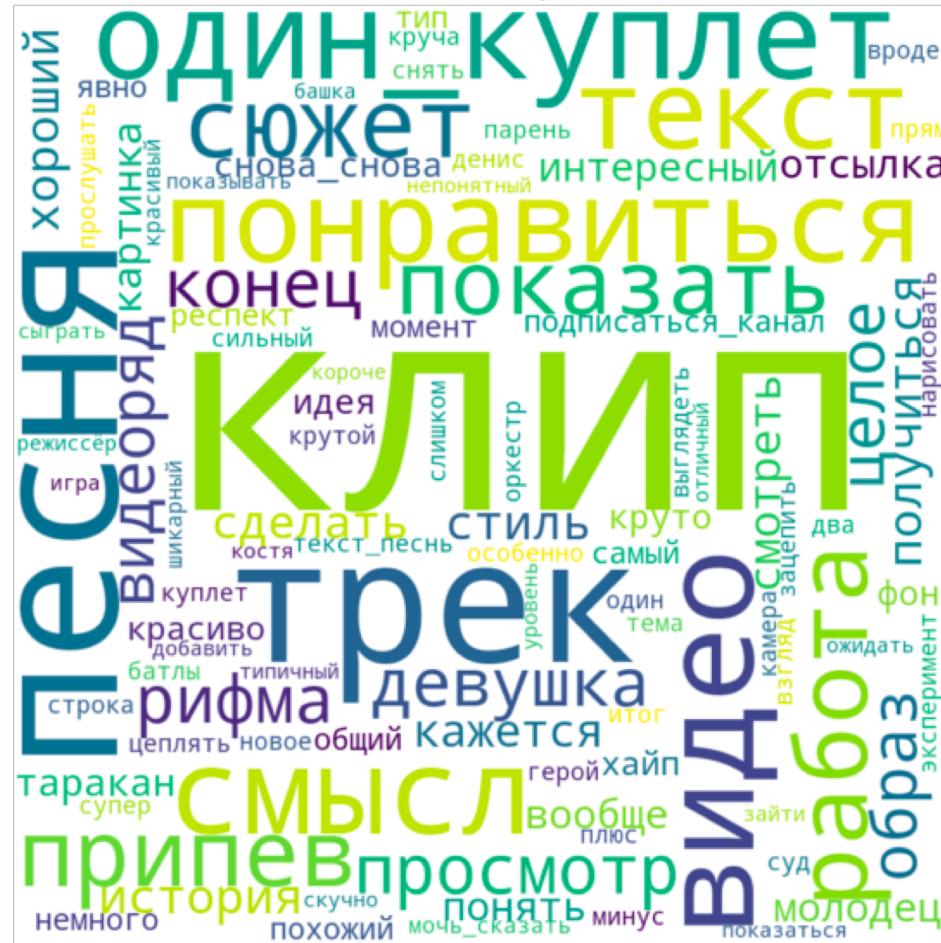
```
from wordcloud import WordCloud

def plotWordCloud(topic_number):
    """
        Строит визуализацию слов на основе текстов топиков
    """

    # получаем частоты и слова топика
    text = dict(lda_30_topics[topic_number][1])

    # строим облако слов
    wordcloud = WordCloud(
        background_color="white",
        max_words=100,
        width=900,
        height=900,
        collocations=False)
    wordcloud = wordcloud.generate_from_frequencies(text)
    plt.figure(figsize=(15, 10))
    plt.title("Топик номер {}".format(topic_number))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis("off");
```

Топик номер 0



Топик номер 1

Топик номер 6

Топик номер 19

Топик номер 27

```
themes_30 = {  
    0 : 'клипы на песни, видео',  
    2 : 'общество и человек',  
    4 : 'социальные темы в рэпе',  
    6 : 'понравившиеся альбомы песен',  
    8 : 'клубы и нюша',  
   10: 'неинтерпретируемая',  
   12: 'Тимати',  
   14: 'высокие темы: поэзия, творчество, книги',  
   16: 'новости и политика',  
   18: 'школа и её проблемы',  
   20: 'благодарности за концерт',  
   22: 'глубоко обсуждаем музыку',  
   24: 'неинтерпретируемая',  
   26: 'ссылки на YouTube',  
   28: 'неинтерпретируемая',  
    1 : 'политика, Украина-Россия',  
    3 : 'рэп и говно',  
    5 : 'песни о любви',  
    7 : 'человек, говорим, думаем и слушаем',  
    9 : 'тексты рок песен (Алиса)',  
   11: 'веб-ссылки и старые песни',  
   13: 'песни о жизни и любви',  
   15: 'Ария',  
   17: 'тексты попсовых песен',  
   19: 'покупка билетов на концерт',  
   21: 'денежные переводы',  
   23: 'фильмы и съемки кино',  
   25: 'Баста и Егор Крид',  
   27: 'Путин, царь и политика',  
   29: 'куча позитива и благодарностей'  
}
```

- Обычно выделение интерпретируемых топиков – конечная цель
- Но что если вероятностные тематические вектора использовать как алгебраические?
- То есть считать расстояния между ними, искать средние и так далее?



Тематические профили



- Посчитав средние тематические вектора по подписчикам, можно получить тематические профили по нужным группировкам
 - Например, по жанрам:

Тематический профиль для shanson:

Тематические профили



- Или по исполнителям:

Тематический профиль для киркоров:

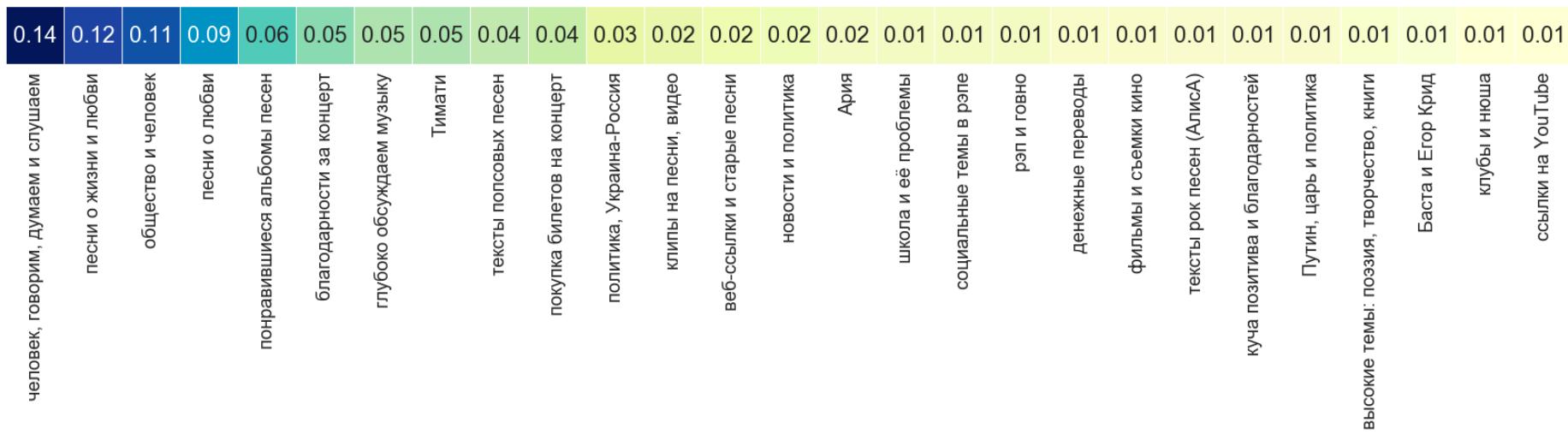
0.17	0.16	0.14	0.10	0.05	0.05	0.05	0.04	0.03	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00					
благодарности за концерт	человек, говорим, думаем и слушаем	песни о жизни и любви	общество и человек	глубоко обсуждаем музыку	тексты попсовых песен	покупка билетов на концерт	понравившиеся альбомы песен	песни о любви	политика, Украина-Россия	Баста и Егор Крид	клипы на песни, видео	веб-ссылки и старые песни	куча позитива и благодарностей	клубы и нюша	социальные темы в рэпе	Ария	высокие темы: поэзия, творчество, книги	новости и политика	фильмы и съемки кино	Тимати	рэп и говно	Путин, царь и политика	школа и её проблемы	ссылки на YouTube	денежные переводы	тексты рок песен (Алиса)
человек, говорим, думаем и слушаем	песни о жизни и любви	общество и человек	глубоко обсуждаем музыку	тексты попсовых песен	покупка билетов на концерт	понравившиеся альбомы песен	песни о любви	политика, Украина-Россия	Баста и Егор Крид	клипы на песни, видео	веб-ссылки и старые песни	куча позитива и благодарностей	клубы и нюша	социальные темы в рэпе	Ария	высокие темы: поэзия, творчество, книги	новости и политика	фильмы и съемки кино	Тимати	рэп и говно	Путин, царь и политика	школа и её проблемы	ссылки на YouTube	денежные переводы	тексты рок песен (Алиса)	
человек, говорим, думаем и слушаем	песни о жизни и любви	общество и человек	глубоко обсуждаем музыку	тексты попсовых песен	покупка билетов на концерт	понравившиеся альбомы песен	песни о любви	политика, Украина-Россия	Баста и Егор Крид	клипы на песни, видео	веб-ссылки и старые песни	куча позитива и благодарностей	клубы и нюша	социальные темы в рэпе	Ария	высокие темы: поэзия, творчество, книги	новости и политика	фильмы и съемки кино	Тимати	рэп и говно	Путин, царь и политика	школа и её проблемы	ссылки на YouTube	денежные переводы	тексты рок песен (Алиса)	
человек, говорим, думаем и слушаем	песни о жизни и любви	общество и человек	глубоко обсуждаем музыку	тексты попсовых песен	покупка билетов на концерт	понравившиеся альбомы песен	песни о любви	политика, Украина-Россия	Баста и Егор Крид	клипы на песни, видео	веб-ссылки и старые песни	куча позитива и благодарностей	клубы и нюша	социальные темы в рэпе	Ария	высокие темы: поэзия, творчество, книги	новости и политика	фильмы и съемки кино	Тимати	рэп и говно	Путин, царь и политика	школа и её проблемы	ссылки на YouTube	денежные переводы	тексты рок песен (Алиса)	

Тематические профили



- Или по исполнителям:

Тематический профиль для тимати:



- А имея средние вектора, можно считать расстояния между ними!
- Например, косинусное:

Расстояния между жанрами

O T U S

	Estrada	Popsa	Rock	Metal	Classic	Rap	Shanson	Jazz	Pank	Dance	Indi
Estrada	0.00	0.05	0.01	0.06	0.12	0.04	0.09	0.12	0.03	0.05	0.02
Popsa	0.05	0.00	0.06	0.12	0.20	0.04	0.08	0.20	0.08	0.05	0.04
Rock	0.01	0.06	0.00	0.06	0.09	0.04	0.08	0.12	0.01	0.04	0.01
Metal	0.06	0.12	0.06	0.00	0.15	0.09	0.16	0.09	0.07	0.07	0.06
Classic	0.12	0.20	0.09	0.15	0.00	0.14	0.23	0.04	0.09	0.14	0.11
Rap	0.04	0.04	0.04	0.09	0.14	0.00	0.11	0.15	0.04	0.01	0.02
Shanson	0.09	0.08	0.08	0.16	0.23	0.11	0.00	0.23	0.14	0.13	0.08
Jazz	0.12	0.20	0.12	0.09	0.04	0.15	0.23	0.00	0.12	0.14	0.12
Pank	0.03	0.08	0.01	0.07	0.09	0.04	0.14	0.12	0.00	0.03	0.02
Dance	0.05	0.05	0.04	0.07	0.14	0.01	0.13	0.14	0.03	0.00	0.02
Indi	0.02	0.04	0.01	0.06	0.11	0.02	0.08	0.12	0.02	0.02	0.00

Расстояния между профилями

O T S

- А также можно искать похожих (и непохожих) друг на друга исполнителей!

Киркоров: топ-10 похожих и непохожих

0.02	0.02	0.03	0.03	0.03	0.04	0.04	0.04	0.04	0.05	0.29	0.29	0.30	0.32	0.36	0.38	0.39	0.46	0.48	0.82
ёлка	апэ_Вера	лолита	агутин	леонтьев	serebro	лазарев	валерия	николаев	арбенина	сектор_газа	rsac	denderty	e_music_jazz	михайлов	антоха_mc	рыцарные_рыцари	classic_best	contemporary_classical	буйнов

Расстояния между профилями

O T U S

- А также можно искать похожих (и непохожих) друг на друга исполнителей!

Лсп: топ-10 похожих и непохожих

0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.04	0.04	0.29	0.30	0.31	0.31	0.32	0.40	0.41	0.43	0.77
------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------

эпджей

Максим

oxxxymiron

gazgolder

noizemc

барских

little_big

пошлая_мопли

майами

ГШ

антоха_mc

басков

хадн_дадн

ludvigvonbeethoven

шуфутинский

русскийшансон

contemporary_classical

рыцарные_рыцари

classic_best

буйнов

Король_и_шут: топ-10 похожих и непохожих

0.01	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.22	0.23	0.23	0.24	0.24	0.25	0.28	0.32	0.35	0.70
------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------

лспс

бм_2

руки_вверх

агутин

эмфира

ротару

мумий_тролль

звери

трофимов

электрофорез

jazz_джаз

miyagi&эндишиль

peterchaikovsky

inclassic

classic_best

classicstories

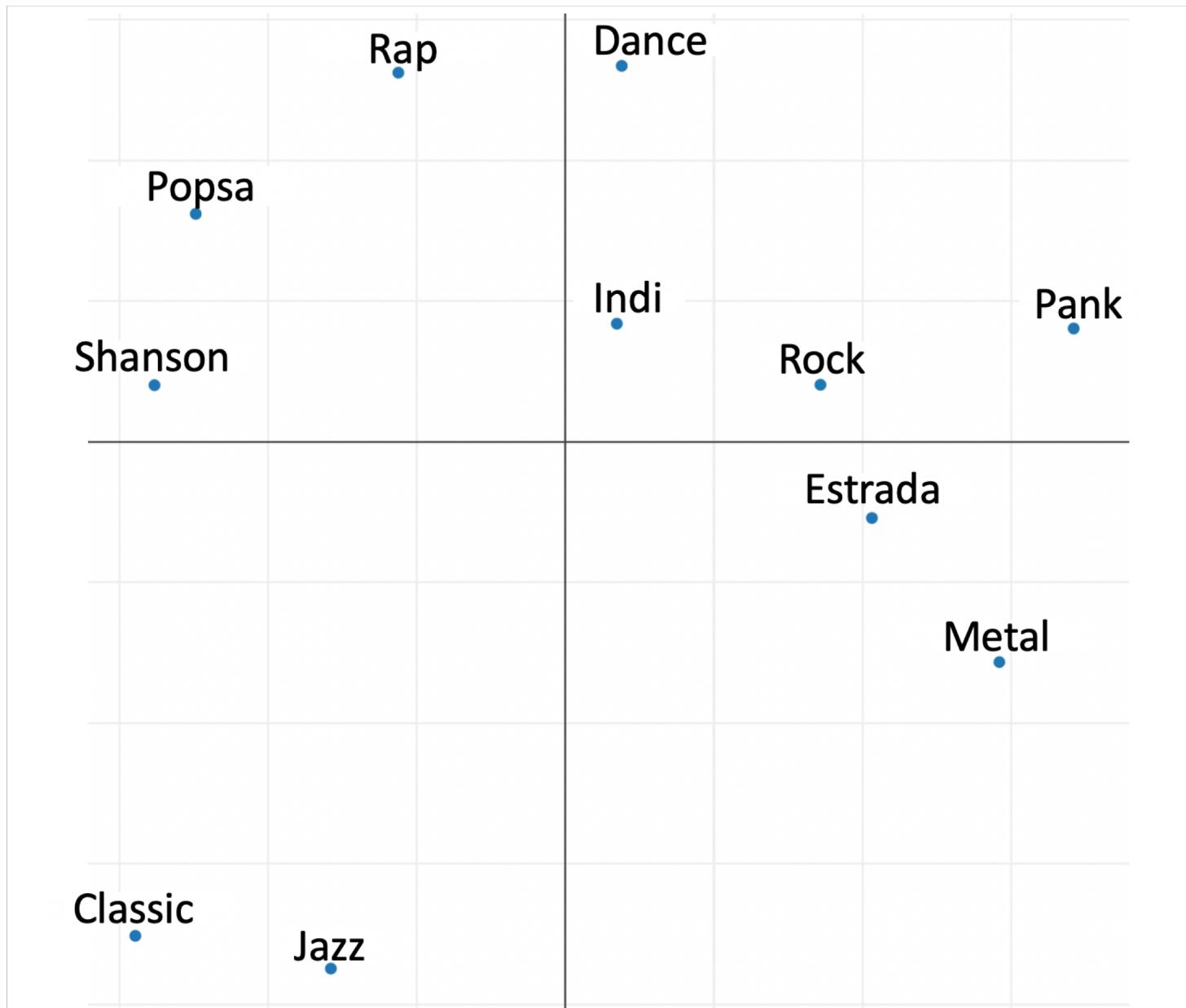
казускома

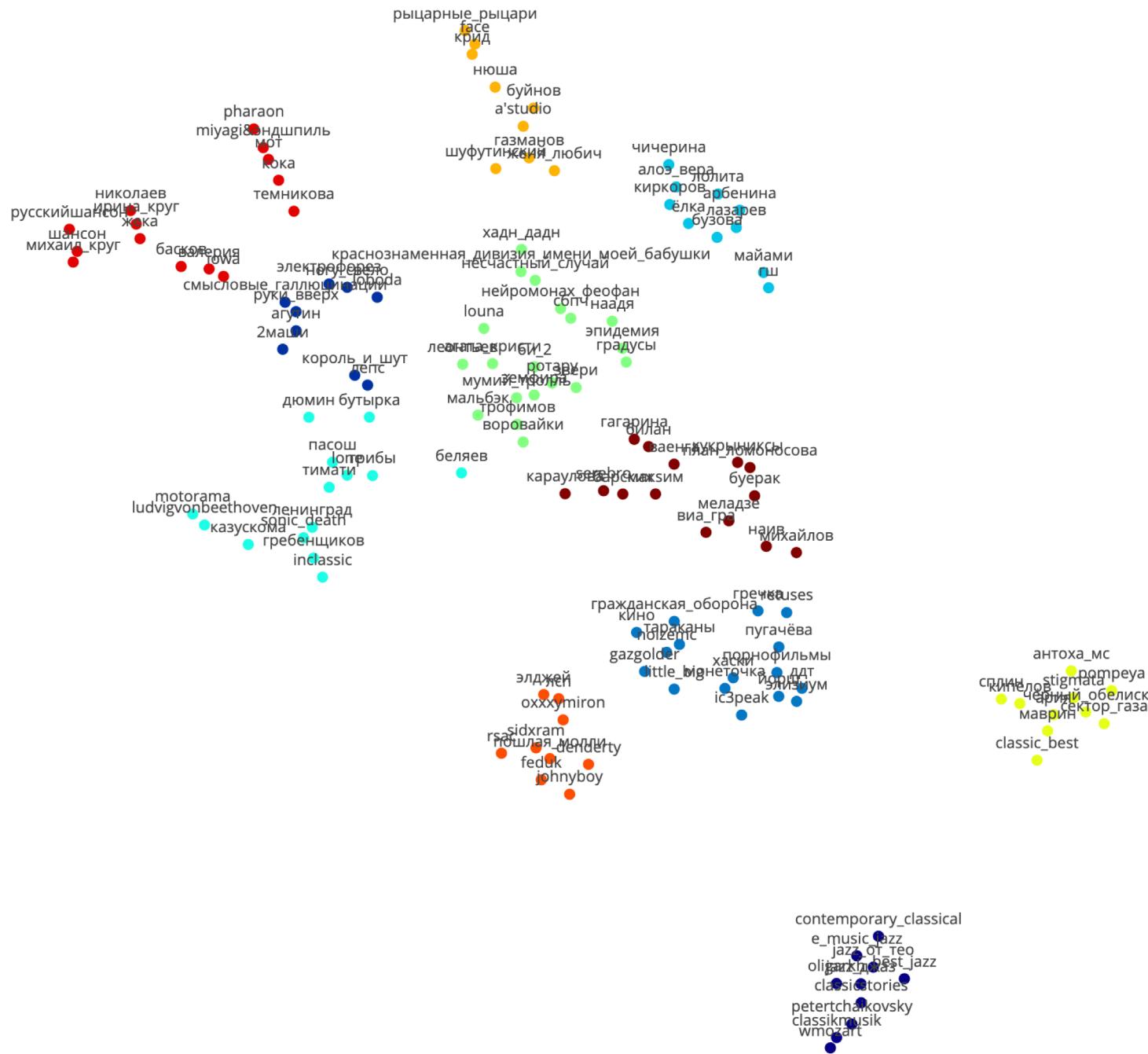
contemporary_classical

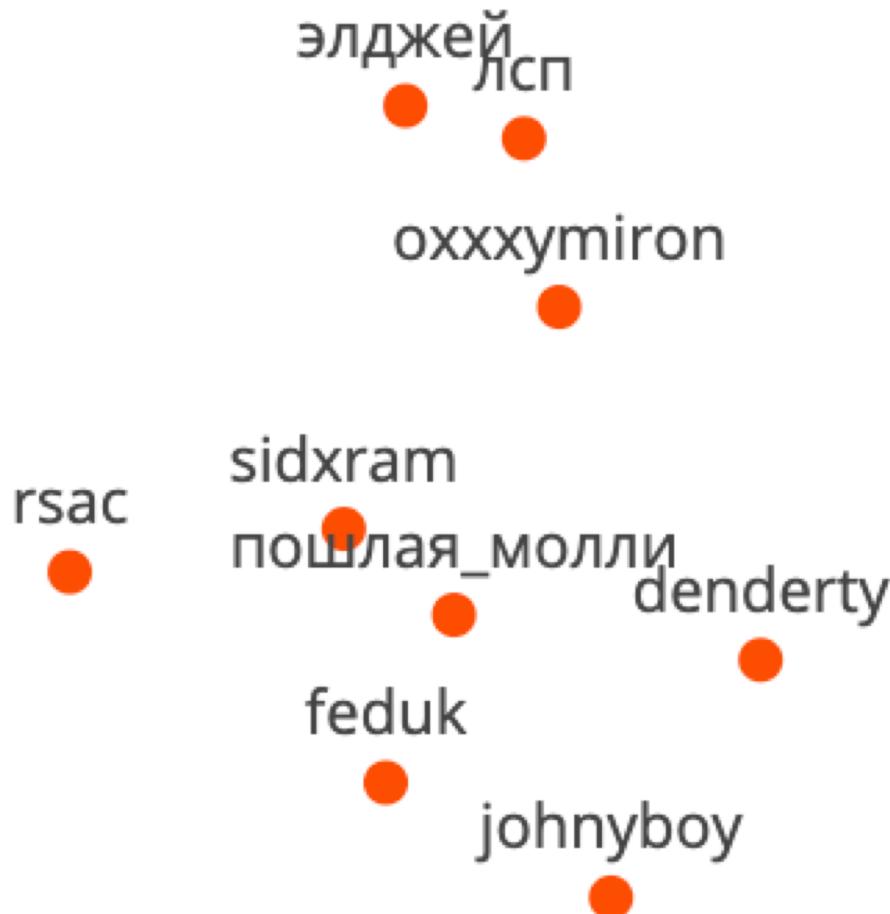
рыцарные_рыцари

буйнов

- Наконец, можно подсмотреть в многомерное пространство на плоскости!
- Сейчас у нас 30—мерные вектора средних вероятностей (по числу топиков)
- При помощи tSNE можно сжать их в двумерное представление
- Multicore-TSNE – как TSNE, только Multicore
<https://github.com/DmitryUlyanov/Multicore-TSNE>
- Подробнее – на занятии «Методы уменьшения размерности»







Карта русской музыки

О T U S







- Слушатели разных жанров действительно обсуждают разное
- Настолько разное, что по темам обсуждений можно заново выделять кластеры музыкальных жанров
- И это круто!

05

Итоги

- Познакомились с задачей тематического моделирования:
 - Для чего нужно
 - Что в основе

- Познакомились с задачей тематического моделирования
- Узнали об основных этапах предобработки текста:
 - Очистка
 - Стэмминг
 - Токенизация
 - Фильтрация

- Познакомились с задачей тематического моделирования
- Узнали об основных этапах предобработки текста
- Посмотрели на возможности тематического моделирования:
 - Получение интерпретируемых топиков
 - Построение тематических профилей
 - Поиск схожих сообществ

- Ссылка на полный проект:

<https://github.com/DmitrySerg/top-russian-music>

- Сбор данных из VK через API
- Настройка параметров LDA
- Вычисление количественных метрик качества модели: перплексия, когерентность и т.д.
- Работа с библиотекой pyLDAvis

<https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/>



Сергеев Дмитрий

Sergeyev.D.A@yandex.ru

tg: @dmitryserg