ПОРТРЕТ ПОЛИТЕХНИЧЕСКОГО ВУЗА ЧЕРЕЗ ПРИЗМУ НОВОСТНЫХ ТЕКСТОВ

Татьяна Евтушенко
Санкт-Петербургский политехнический университет

Цель

Выявление тематик, обсуждаемых в новостной ленте современного политехнического ВУЗа

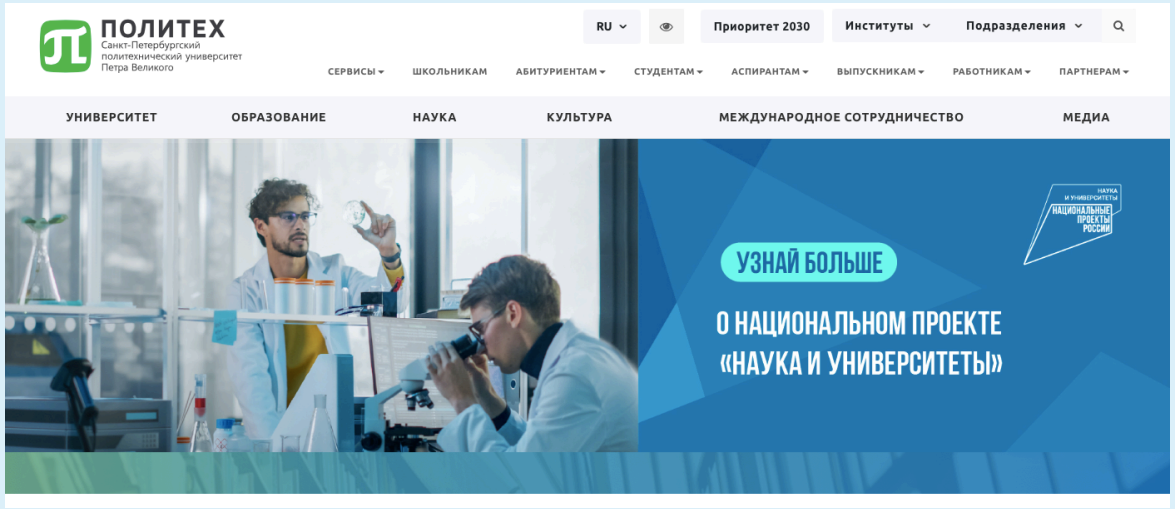
Алгоритм работы

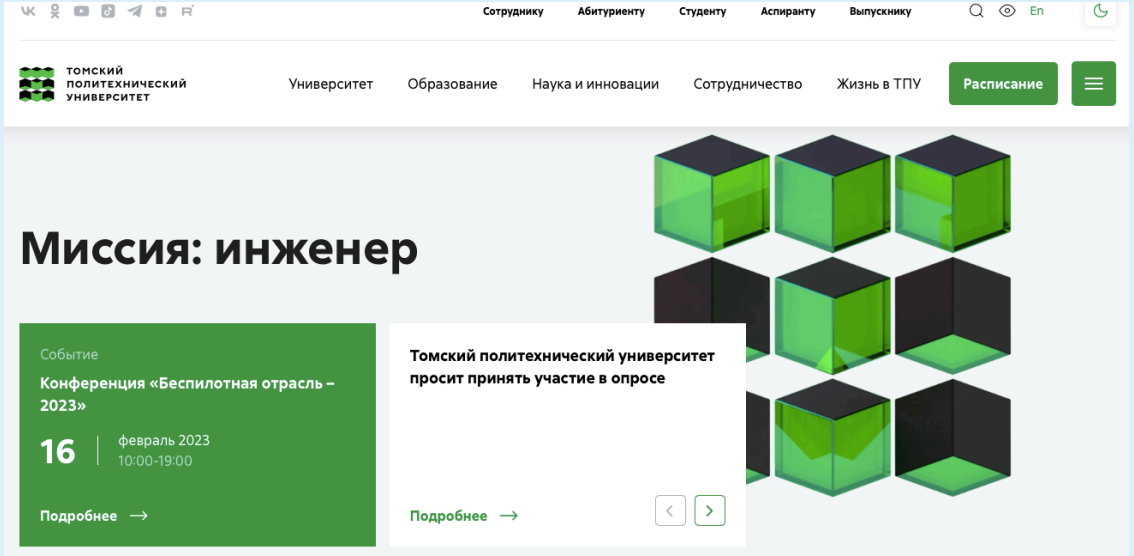
Материал исследования

- Новостные тексты с сайта СПбПУ (txt)
- Новостные тексты с сайта ТПУ (txt)

- 300 текстов

  Период март-июнь 2022г





Выявление темы текста

студент, 6
праздник, 6
программа, 3
спбпу, 3
блюдо, 3
страна, 3
иностранный, 2
Политех, 2
polyunion, 2
организатор, 2
выступить, 2
молодёжь, 2

```
from collections import Counter
Counter(words_lemmatized).most_common(15)
```

```
[('контроль', 11),
 ('метод', 7),
 ('неразрушающий', 6),
 ('тепловой', 5),
 ('материал', 5),
 ('лаборатория', 5),
 ('учёный', 4),
 ('комплекс', 4),
 ('водородный', 4),
 ('энергетика', 4),
 ('который', 4),
 ('тпу', 3),
 ('композит', 3),
 ('исследование', 3),
 ('разрабатывать', 2)]
```

Студенческая жизнь

Облако слов


polytech
местности
команда
первый
проект
студенты
полезной
спирита
марта
корпуса
прошли
spirit
студенческая
успешно
разработки
испытания
ходовые
себя
первые
команды
платформа
модули
политехнического
уже
для
это
нагрузки
voltage machine
приехали
собственной
робототехнической


волонтерыспасатель
санктпетербург
важный
институт
курс
политех
цвп
студент
гуманитарный
навык
проект
организация
осмотр
преподаватель
такой
рассмотреть
исотрудник
инструктор
травма
обморожение
действие
участник
вцентр
дыхание
помощь
пострадать
ожог
сотрудник
общественный
слушатель
вуз
опасный
первый
спбп
восьмичасовой
гармония
наслучай
волонтёрский
остановка
стать
состояние

Тексты ТПУ (пример)

Ключевые слова		Биграммы	Тема
материал графен лазерный соль диазония композит электропроводящий	технология основа прочный учёный обработка воздействие полимерный использование	солями диазония графена модифицированного модифицированного солями лазерной обработки исследователи томского	Научные разработки
контроль метод неразрушающий тепловой материал лаборатория учёный	комплекс водородный энергетика тпу композит исследование разрабатывать	неразрушающего контроля ученые тпу тпу разрабатывают универсальный комплекс водородной атомной атомной энергетике ультразвуковой тепловой	Научные разработки

Тексты СПбПУ (пример)

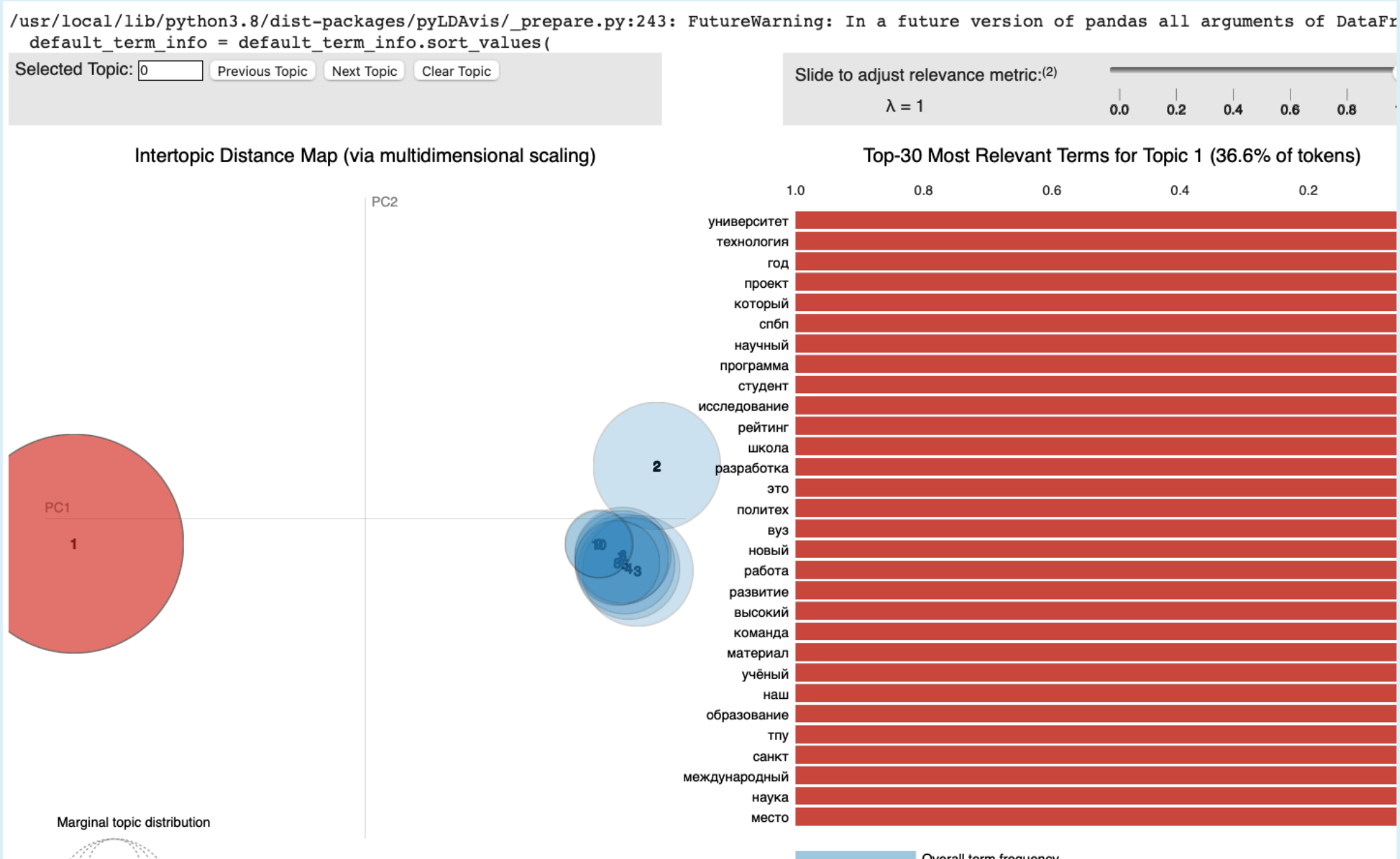
Ключевые слова		Биграммы	Тема
спорт светлана вуз спортивный журов создать	занятие студент молодёжь политех гуманитарный помощь	светлана журова олимпийская чемпионка физкультурой испортом патриотического воспитания гуманитарной помощи спортивных объектов	Спорт
производственный технология передовой спбпу олимпиада центр	профиль технологический национальный задача участник санкт-петербург москва	производственные технологии передовые производственные национальной технологической технологической олимпиады проектной деятельности	Технологии / инновации
курс помощь спбпу первый преподаватель навык организация волонтер	спасатель санкт-петербург гармония политех сотрудник вуз студент	волонтеры спасатели первой помощи общественной организации волонтеры провели	студенческая жизнь

TF-IDF

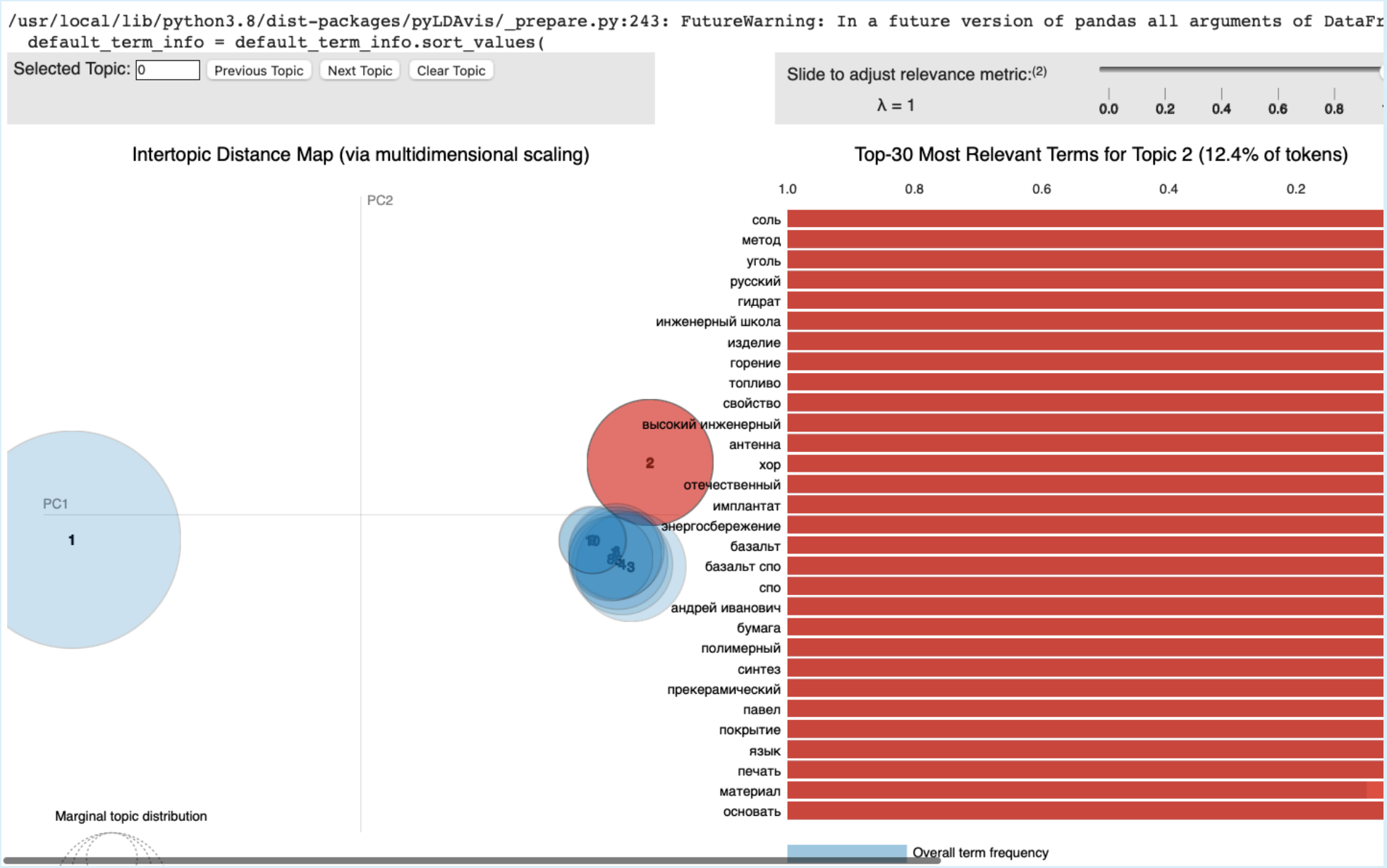
```
feature_names = np.array(tfidf_vectorizer.get_feature_names())

for i, article in enumerate(df.text.head()):
        article_vector = tfidf[i, :]
        words = get_top_tf_idf_words(article_vector, feature_names, 10)
        print(article)
        print(words, '\n')
```

```
российский научный фонд подвести итог два региональный конкурс проект научный группа малый научный группа результат поддержка получить шесть проект учёный
['научный группа' 'научный' 'малый научный' 'группа' 'конкурс'
 'региональный конкурс' 'исследовательский школа' 'малый' 'химический'
 'биомедицинский технология']

учёный томский политехнический университет сколтех предложить эффективный экономичный метод синтез сверхтвёрдый материал пентаборид вольфрам применяться ра
['вольфрам' 'борид вольфрам' 'борид' 'метод' 'синтез' 'синтезировать'
 'катод' 'объём' 'фаза' 'отсутствие необходимость']

проект реализовать лаборатория фаблабнуть политех занять первый второе место первый чемпионат мейкер brics maker competition который проводиться китайский
['тренажёр брайль' 'тренажёр' 'брайль' 'жюри' 'электролонгборд'
 'доработка' 'проект тренажёр' 'студент политех' 'проект' 'чемпионат']

даниил снетковый первокурсник кафедра механика процесс управление физико механический институт спбп достигнуть небывалый высота необычный вид спорт судомод
['судомоделизм' 'спорт' 'модель' 'первенство' 'фаблаба' 'юношеский'
 'секция' 'корабль' 'ты' 'даниил']

учёный тпу разрабатывать универсальный комплекс неразрушающий контроль композит использовать водородный атомный энергетика объединить метод неразрушающий к
['контроль' 'неразрушающий' 'неразрушающий контроль' 'тепловой' 'метод'
 'водородный' 'лаборатория' 'композит' 'комплекс' 'водородный атомный']
```

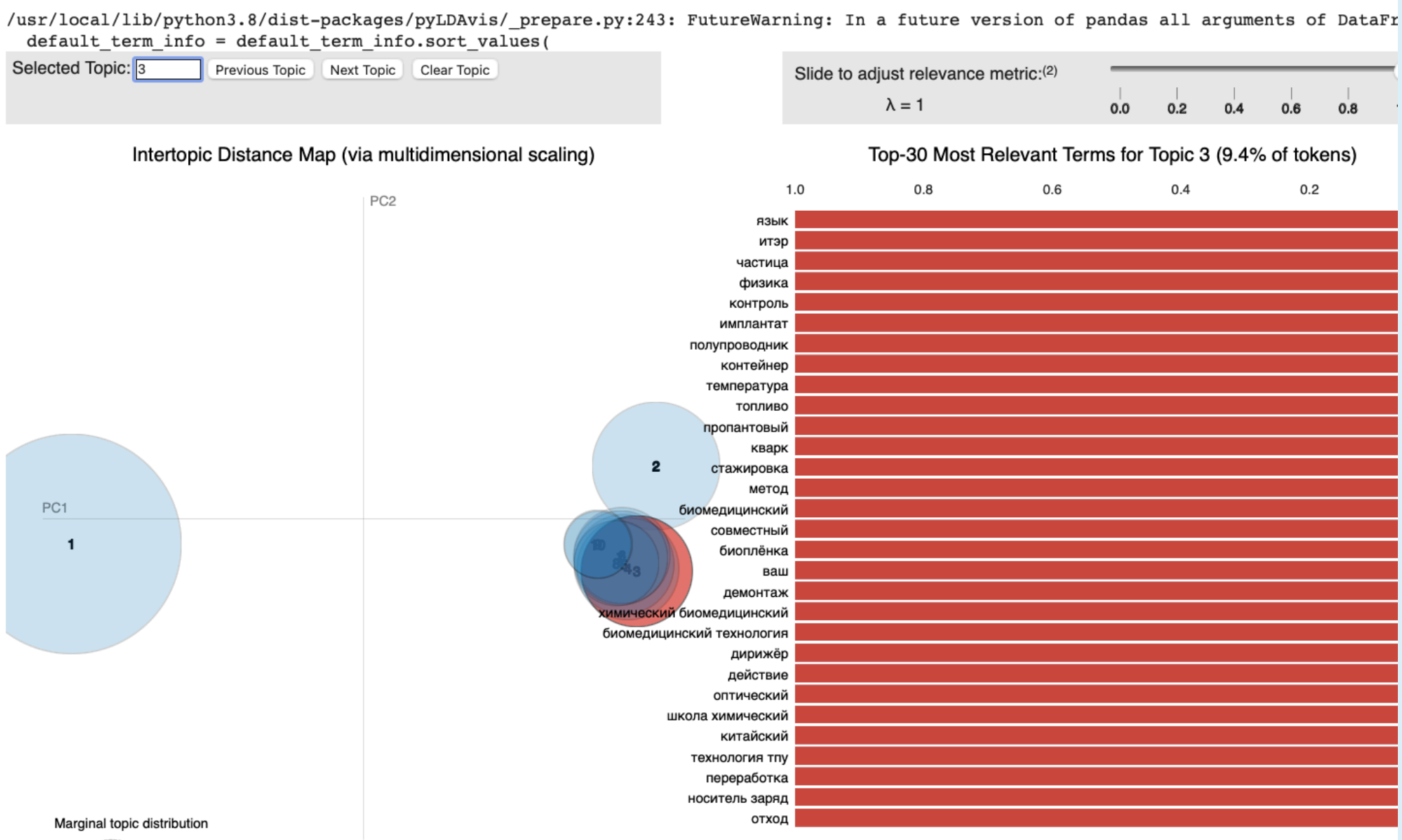
LDA_Тема 1


/usr/local/lib/python3.8/dist-packages/pyLDAvis/_prepare.py:243: FutureWarning: In a future version of pandas all arguments of DataFr
default_term_info = default_term_info.sort_values(
Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 1 (36.6% of tokens)
1.0
0.8
0.6
0.4
0.2
университет
технология
год
проект
который
спбп
научный
программа
студент
исследование
рейтинг
школа
разработка
это
политех
вуз
новый
работа
развитие
высокий
команда
материал
учёный
наш
образование
тпу
санкт
международный
наука
место
Overall term frequency

LDA_Тема 2


/usr/local/lib/python3.8/dist-packages/pyLDAvis/_prepare.py:243: FutureWarning: In a future version of pandas all arguments of DataFr
default_term_info = default_term_info.sort_values(
Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 2 (12.4% of tokens)
1.0
0.8
0.6
0.4
0.2
соль
метод
уголь
русский
гидрат
инженерный школа
изделие
горение
топливо
свойство
высокий инженерный
антенна
хор
отечественный
имплантат
энергосбережение
базальт
базальт спо
спо
андрей иванович
бумага
полимерный
синтез
прекерамический
павел
покрытие
язык
печать
материал
основать
Overall term frequency

LDA_Тема 3


/usr/local/lib/python3.8/dist-packages/pyLDAvis/_prepare.py:243: FutureWarning: In a future version of pandas all arguments of DataFr
default_term_info = default_term_info.sort_values(
Selected Topic: 3
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 3 (9.4% of tokens)
1.0
0.8
0.6
0.4
0.2
язык
итэр
частица
физика
контроль
имплантат
полупроводник
контейнер
температура
топливо
пропантовый
кварк
стажировка
метод
биомедицинский
совместный
биоплёнка
ваш
демонтаж
химический биомедицинский
биомедицинский технология
дирижёр
действие
оптический
школа химический
китайский
технология тпу
переработка
носитель заряд
отход

Портрет современного политехнического ВУЗа

Библиотеки

```
import os
import pandas as pd
from nltk import word_tokenize

import nltk
nltk.download('punkt')

from nltk.tokenize import WordPunctTokenizer

nltk.download('stopwords')
from nltk.corpus import stopwords
stopwords = stopwords.words('russian')

import string

!pip install pymorphy2
from pymorphy2 import MorphAnalyzer
morph = MorphAnalyzer()

from collections import Counter

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer

import numpy as np

from sklearn.decomposition import LatentDirichletAllocation

!pip install pyldavis

import pyLDAvis
import pyLDAvis.sklearn
```

Схожие проекты

- Смирнова В.Д. Автоматическое определение тем, ассоциированных с пандемией covid-19, в русскоязычном корпусе социальных медиа (ВКР)
- Tatiana Sherstinova et al. Topic modeling of the Russian short stories of 1900–1930s: the most frequent topics and their dynamics
- А.В. Чижик Исследование динамики общественного настроения в социальных сетях с использованием методов тематического моделирования International Journal of Open Information Technologies ISSN: 2307-8162 vol. 9, no. 12, 2021
- М.Апишев Анализ текстов Лекция Предобработка и выделение признаков.
- Чечнева Н.С. Исследование оценочной лексики потребительских отзывов в системе Яндекс.Маркет

Перспективы работы

- Посмотреть TfIdf
- Посмотреть LDA
- NER
- Изучить и сопоставить разные подходы для получения более точного результата
- Sentiment анализ
- …

СПАСИБО ЗА ВНИМАНИЕ