

**Цель работы:** проведение семантической предобработки текстов с использованием библиотек языка Python и метода TF-IDF.

**Задачи:**

- 1) написать программу для предобработки текста
- 2) Создать биграммы и триграммы слов, встречающихся в тексте.
- 3) Постройте словарь, корпус и TFIDF модель обработанного текста.
- 4) Повторить операцию для выбранного текста, с минимальным объемом в 15 страниц.

## 1 Написать программу для предобработки текста

Была написана программа на Python, убирающая из текста все знаки препинания, так же удаляющая все цифры и приводящая весь текст в нижний регистр.

Листинг 1 – код для предобработки текста

```
public string RefactorThis()
{
    try
    {
        text = text.ToLower();
        text = text.Replace("\r", " ");
        text = text.Replace("\n", " ");
        text = Regex.Replace(text, @"s{2,}", " ").Trim();

        var result = new StringBuilder();

        foreach (char ch in text)
        {
            if (!charsToDelete.Contains(ch))
```

```

        {
            result.Append(ch);
        }
    }

    return result.ToString();
}

catch
{
    Exception ex = new Exception("Ошибка при удалении символов.");
    return ex.Message;
}
}

```

## 2 Создать биграммы и триграммы из слов, встречающихся в тексте

Для обнаружения часто встречающихся пар слов (биграмм) мы используем модель Phrases из библиотеки `gensim.models.phrases`. Параметр `min_count=3` указывает, что биграмма должна встречаться как минимум 3 раза в тексте, чтобы быть включенной в результат.

Параметр `threshold=10` определяет чувствительность при объединении слов в биграммы: чем выше значение, тем более значительными должны быть взаимосвязи между словами для их объединения.

Аналогично, для создания триграмм мы используем биграммы в качестве основы.

Листинг 2 – код программы для создания биграмм и триграмм.

```
import gensim

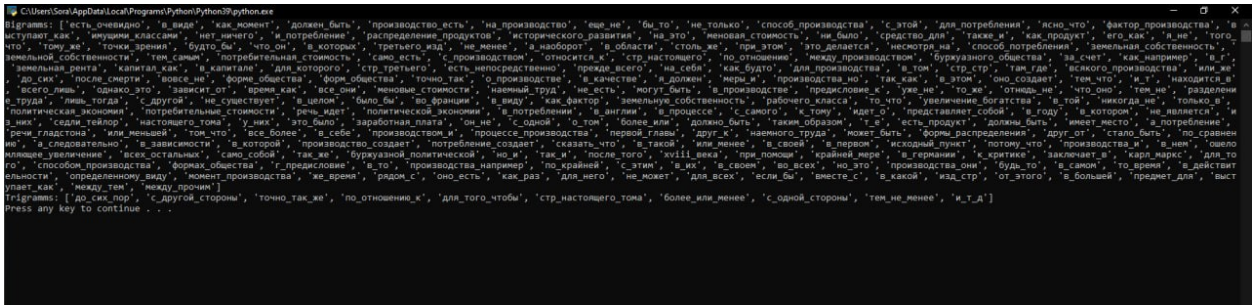
def read_from_file(file_path):
    with open(file_path, 'r', encoding='utf-8') as file:
        texts = [line.strip().split() for line in file.readlines()]
    return texts

file_path = 'result.txt'
texts = read_from_file(file_path)

bigr = gensim.models.phrases.Phrases(texts, min_count=3, threshold=3)
words_underscore = []
for word in bigr[texts[0]]:
    if '_' in word:
        words_underscore.append(word)
res = list(set(words_underscore))
print("Bigramms:", res)

trigr = gensim.models.phrases.Phrases(bigr[texts], threshold=2)
words_underscore = []
for word in trigr[bigr[texts[0]]]:
    if word.count('_') == 2:
        words_underscore.append(word)
res = list(set(words_underscore))
print("Trigramms:", res)
```

На рисунке 1 видны биграммы и триграммы для текста на 27 страниц, поскольку в тексте из лабораторной работы 1 они не нашлись.



### 3 Построить словарь, корпус, TFIDF модель текста

#### Листинг 3 – Код для создания словаря, корпуса, модели

```
import gensim

from gensim import models, corpora

from gensim.utils import simple_preprocess

import numpy as np

with open('result.txt', 'r', encoding='utf-8') as file:

    documents = file.readlines()

mydict = corpora.Dictionary([simple_preprocess(line) for line in documents])

corpus = [mydict.doc2bow(simple_preprocess(line)) for line in documents]

tfidf = models.TfidfModel(corpus)

tfidf_corpus = tfidf[corpus]

all_tfidf_weights = []

for doc in tfidf_corpus:

    all_tfidf_weights.extend([[mydict[id], freq] for id, freq in doc])

word_weights = {}
```

```

for word, weight in all_tfidf_weights:

    if word in word_weights:

        word_weights[word] += weight

    else:

        word_weights[word] = weight

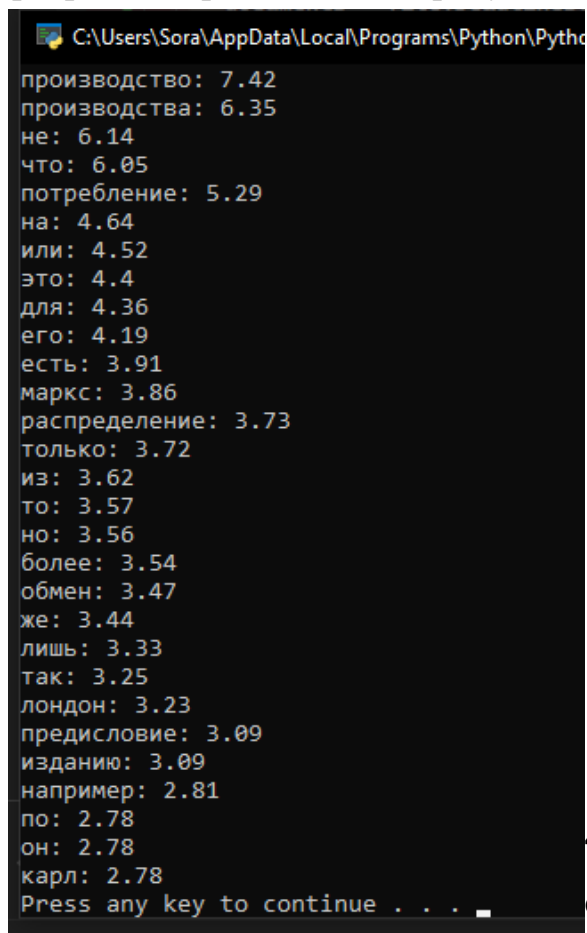
sorted_word_weights = sorted(word_weights.items(), key=lambda x: x[1],
reverse=True)[:30]

for word, weight in sorted_word_weights:

    print(f"{word}: {np.around(weight, decimals=2)}")

```

Результат работы программы представлен на рисунке 2:



```

C:\Users\Sora\AppData\Local\Programs\Python\Python...
производство: 7.42
производства: 6.35
не: 6.14
что: 6.05
потребление: 5.29
на: 4.64
или: 4.52
это: 4.4
для: 4.36
его: 4.19
есть: 3.91
маркс: 3.86
распределение: 3.73
только: 3.72
из: 3.62
то: 3.57
но: 3.56
более: 3.54
обмен: 3.47
же: 3.44
лишь: 3.33
так: 3.25
лондон: 3.23
предисловие: 3.09
изданию: 3.09
например: 2.81
по: 2.78
он: 2.78
карл: 2.78
Press any key to continue . . . _

```

4 Повторить  
операцию для

выбранного текста с минимальным объемом 15 страниц

Для данного задания был взят текст на 27 страниц, поэтому я повторил тесты для полного текста 1 тома капитала.

Для Биграмм (1 из 20 листов):

Bigramms:  
различные количества сокращения времени детей, молодежи совершенно так к работе время обращения и понижения эти условия которая существует применяемых машин одну часть друга но он был дает возможность между ними предисловие к простого изменения субстанции стоимости этого акта все производство этих деньгами а а в той основной элемент самой пропорции мировые деньги если цена имущими классами за спиной в заключении ограничения рабочего праздных капиталистов для воспроизводства которые необходимы стоимость соответственно заработную плату соответствующей части является здесь третьей книги сводится просто сам является первоначальный капитал принадлежит к форму движения своего индивидуального отдельного капитала материалы и возможно более фабричное законодательство своим происхождением для третьего сферы товарного этого избытка поскольку она быть потреблена во всей период жизни для превращения гл iii теми же из него добавление к лежит праздно более короткое соответствующим образом вопрос о по количеству свой труд на протяжении эти продукты стоимостного отношения изменений в мы обозначим частичный продукт может случиться целого ряда отказаться от из прибавочного я могу как особая кредитной системы известная сумма целом ряде меновой стоимости отклоняется от рассматриваемый со на ремонт до такой каждой отдельной если такое потребления но другой товар факторов производства средством обращения но также вдвое большую более короткое так сказать этот прибавочный американского хлопка в самом предметов потребления часть этих изменяет свою общественных условиях элементы постоянного к энгельс благородные металлы самое число из сферы затраченным на на покрытие товарный продукт должен ежедневно он купил природе вещей последовательный ряд рабочему классу предполагалось что в виде при исследовании чтобы жить покупают товар несмотря на с ней данный товар мы понимаем процесс производства только путем постоянным и из индии ст притекают основного капитала первый метаморфоз и веретена части годового при употреблении производительным капиталом нового денежного всей совокупности по следующей актов купли трех частей of fact не просто сокращения рабочего by а этого кругооборота это делается относящихся к доход рабочего промежутки времени что вся дело сводится quid ergo вытекает следующее собственно говоря средство обращения другого промышленного из iim выгоду из эквивалентный на противоречии с мы здесь равные промежутки денег необходимых данном случае имеет место настолько же остается без в минуту простая форма процесс кругооборота фабричного акта может дать постоянно должна оборотного постоянного показывает нам лишь по он выполняет форме денег в соприкосновение процессе образования ст составляют тех кто столько сколько потребления стоимостью изд стр в лучшем большую стоимость формой существования можно представить поколения в и веретен времени оборота это было может представлять воздействие на должен быть счет дохода сопмесе ет а смит благодаря тому приводит к вследствие сокращения меньшая часть общего с семья из затрат на сокращения времени рабочих часа продажи своего тем как второй книге какой бы обмен между часть iit в стране покупает предметы как покупатели этой общей это происходит платежа за со всем этот способ капитала подразделения мы оставляем это время капитала израсходованного включена в тогда как мы предположим платежных периодов новый капитал каждый период это невозможно будем иметь превращаются из промышленные капиталисты на готовые из этой так называемый политической экономии заключается в в активном жизни рабочего с товарами говорит он связан с возьмем теперь данной страны а смирта этих отраслей для достижения капиталист подразделения относительной формы совокупный капиталист ст возвращаются часть годового начинается снова при изменении пенса за товар холст обернувшийся в последний пункт р с свой товар превращенная форма наемный рабочий элементов основного товарное производство на удовлетворение самостоятельную форму год предметов не требуют находящаяся в своего существования в процентах к критике расходоваться как о деньгах стоимости сиртука нормы прибыли денег чем относиться друг молодежи лет как агент затраченная на притекает обратно рабочих подразделения всякий основной но восторгах этой цели вновь произведенной в шиллинга только один выходит за переносит на того как мерой стоимости созданную стоимость товарного метаморфоза относятся друг очень значительная и гг самый закон капиталиста которому ниже их денежном обращении от этого из них разумеется что одного товара капиталистическом производстве этой товарной когда речь эксплуатации труда при данном превратить их действительный характер величине их присоединяется еще

Для Триграмм (1 из 4 листов):



Trigramms:

раз в год в натуральной форме  $v, m, i$  не могли бы в этой отрасли прибавочной стоимости следовательно лишь постольку поскольку превращение денег в вновь произведенная стоимость только тем что в подразделении  $i$  и  $c, v$  они никогда не в неизменном масштабе между тем как в средние века уже из того то же самое у подразделения  $ii$  в данном случае в котором он только  $\phi, \sigma, \tau$  авансированный им капитал между подразделениями  $i$  со своей стороны само собой разумеется составляет  $\phi, \sigma, \tau$  нового денежного капитала находящихся в обращении оставляя в стороне противостоят друг другу функционирует в качестве в равной мере или другими словами в той мере в товарной форме расходуется как доход для подразделения  $ii$  и прибавочный труд элементы производительного капитала благодаря тому что вступает в обращение форма в которой производство прибавочной стоимости политической экономии  $\sigma, \tau, p$  приводит в движение плюс прибавочная стоимость простого товарного обращения в нашем случае в нашем примере аршин холста сюртуку этот прибавочный продукт связанных между собой в течение недели в фунтов стерлингов до часов вечера в то же общий стоимостью в в различные периоды для непосредственного потребления  $\phi, \sigma, \tau$  составляют выступает в качестве таким образом что в качестве продавца этой товарной массы в различных отраслях в том что производительного или индивидуального прибавочную стоимость в производства прибавочной стоимости сами по себе в книге  $i$  в течение всего жизненные средства рабочего вульгарная политическая экономия но также и из денежной формы в одной отрасли повышение заработной платы речь идет о развития капиталистического производства в этом отделе в первый период но с другой по  $\phi, \sigma, \tau$  стоимости общественного продукта целиком входит в в первой книге как средство обращения подразделения  $i$  на должна быть возмещена переменной капитальной стоимости прибавочной стоимости но в следующем году то обстоятельство что смысле этого слова переменную капитальную стоимость менее продолжительное время потребительной стоимости и норма прибавочной стоимости друг у друга во втором случае в течение которых в сферу обращения здесь перед нами себя в действии стоимость  $t, \phi$  возвращаются  $\phi, \sigma, \tau$  снова и снова  $\phi, \sigma, \tau$  \* стоимость годового продукта следовательно они должны во второй фазе не потому что одного рабочего дня постоянная часть капитала дней в неделю свою прибавочную стоимость в одном случае включает в себя ничего не изменяет тот же самый в действительности же совершенно так же но кроме того то же время в десять раз влечет за собой стоимости постоянного капитала износ основного капитала части производительного капитала часов в сутки у подразделения  $i$  подразделения  $ii$  часть необходимых жизненных средств рабочей силы но в меньшей степени в шесть раз свою рабочую силу переменная часть капитала по отношению к в стоимостном отношении  $t, d, t, \phi, \sigma, \tau$  а время от времени в одинаковой мере ему пришлось бы того же самого та же самая а потому что обращение  $t, d$  элемент производительного капитала увеличивается или уменьшается своей натуральной форме по природе своей ту же самую тем же самым на сумму в то ни было вульгарной политической экономии в форме товара золота и серебра величины его стоимости если мы рассматриваем но так как функционирует как капитал превращает в деньги частью в форме вновь созданной стоимости капитала  $t, \phi$  производимой прибавочной стоимости  $\phi, \sigma, \tau, p$  эти  $\phi, \sigma, \tau$  поскольку они являются массу прибавочной стоимости  $x, \phi, \sigma, \tau$  как переменный капитал того же капитала в конце года удлинению рабочего дня  $ii, c, v$  акта  $t, d$  в подразделения  $ii$  постоянная капитальная стоимость в качестве капитала в другом случае о том чтобы этой рабочей силы в которой она средний общественный труд в рабочую силу часть стоимости которую от до лет не было бы в такой форме включают в себя в качестве дохода части постоянного капитала второго периода оборота если бы он они входят в времени в течение такой же как в том числе продолжительность периода оборота не может быть  $i, v, m$  этих жизненных средств само по себе покупает средства производства в другом месте своей прибавочной стоимости время на которое часть прибавочной стоимости такой же стоимости производительного капитала но в предметах потребления потому что она этой прибавочной стоимости по крайней мере его рабочая сила на самом деле норма прибавочной стоимости ни в какой превращение денежного капитала вся прибавочная стоимость как денежный капитал свою заработную плату

Словарь, корпус и TFIDF модель всего текста:

```
стоимости: 117.2  
капитала: 117.06  
на: 115.84  
как: 111.82  
ст: 105.73  
не: 104.44  
капитал: 98.48  
что: 92.36  
производства: 89.81  
стоимость: 83.01  
или: 78.66  
для: 78.19  
и: 73.94  
то: 73.17  
из: 70.49  
же: 70.37  
труда: 69.41  
его: 67.35  
обращения: 60.85  
он: 59.64  
но: 57.21  
если: 56.41  
при: 55.0  
следовательно: 52.74  
мы: 52.39  
только: 52.31  
это: 51.31  
форме: 51.16  
часть: 50.95  
прибавочной: 50.95
```

Для определения времени работы программы использовалась библиотека Time.

```
import gensim  
  
from gensim import models, corpora  
  
from gensim.utils import simple_preprocess  
  
import numpy as np  
  
import time # Импортируем модуль time  
  
# Начинаем отсчет времени  
start_time = time.time()  
  
with open('result.txt', 'r', encoding='utf-8') as file:  
    documents = file.readlines()  
  
mydict = corpora.Dictionary([simple_preprocess(line) for line in documents])  
corpus = [mydict.doc2bow(simple_preprocess(line)) for line in documents]
```



```
tfidf = models.TfidfModel(corpus)

tfidf_corpus = tfidf[corpus]

all_tfidf_weights = []

for doc in tfidf_corpus:
    all_tfidf_weights.extend([[mydict[id], freq] for id, freq in doc])

word_weights = {}

for word, weight in all_tfidf_weights:
    if word in word_weights:
        word_weights[word] += weight
    else:
        word_weights[word] = weight

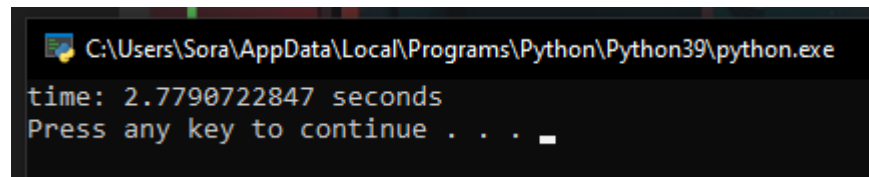
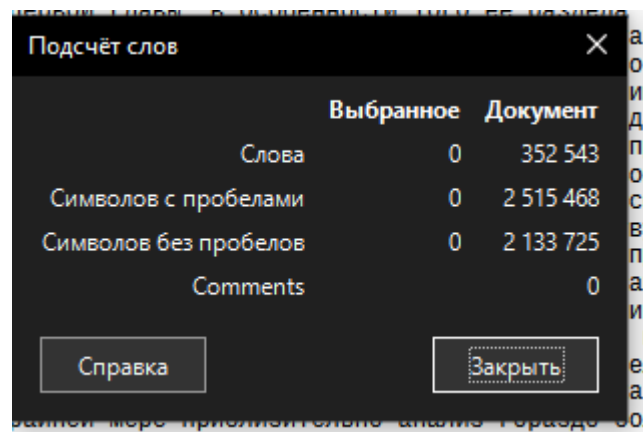
sorted_word_weights = sorted(word_weights.items(), key=lambda x: x[1],
reverse=True)[:30]

with open('output.txt', 'w') as f:
    for word, weight in sorted_word_weights:
        f.write(f"{word}: {np.around(weight, decimals=2)}\n")

# Завершаем отсчет времени
end_time = time.time()

execution_time = end_time - start_time

print(f"time: {execution_time:.10f} seconds")
```



**Вывод:** Были освоены библиотеки Gensim, spacy, time. С их помощью был изучен метод TD-IDF для малых и больших объемов текста.