

Министерство науки и высшего образования Российской Федерации  
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Филологический факультет

УТВЕРЖДЕНО:  
Декан  
И. В. Тубалова

Оценочные материалы по дисциплине

Искусственный интеллект в NLP

по направлению подготовки

**45.04.03 Фундаментальная и прикладная лингвистика**

Направленность (профиль) подготовки:  
**Компьютерная и когнитивная лингвистика**

Форма обучения  
**Очная**

Квалификация  
**Магистр**

Год приема  
**2024**

СОГЛАСОВАНО:  
Руководитель ОП  
З.И. Резанова

Председатель УМК  
Ю.А. Тихомирова

Томск – 2025

## **1. Компетенции и индикаторы их достижения, проверяемые данными оценочными материалами**

Целью освоения дисциплины является формирование следующих компетенций:

ПК-4 Способен разрабатывать проекты прикладной направленности в области когнитивной и компьютерной лингвистики с применением современных технических средств и информационных технологий, в том числе в области искусственного интеллекта.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИПК-4.1 Формулирует цель проекта прикладной направленности в области когнитивной и компьютерной лингвистики, обосновывает необходимость применения современных технических средств и информационных технологий, в том числе в области искусственного интеллекта

ИПК-4.2 Разрабатывает программу действий по решению задач проекта в области когнитивной и компьютерной лингвистики с учетом имеющихся технических средств и информационных технологий, в том числе в области искусственного интеллекта

ИПК-4.3 Обеспечивает выполнение проекта в области когнитивной и компьютерной лингвистики с применением современных технических средств и информационных технологий, в том числе в области искусственного интеллекта, в соответствии с установленными целями, сроками и затратами

## **2. Оценочные материалы текущего контроля и критерии оценивания**

Элементы текущего контроля:

К числу форм контроля, оценивающих уровень достижения компетенций по текущей дисциплине, относятся тесты, устные вопросы, выполнение практических заданий ( ПК-4, ИПК-4.1, ИПК-4.2 ), контрольная и самостоятельная работы (ПК-4, ИПК-4.1).

Итоговый проект включает блок вопросов по теоретической части изучаемой дисциплины ( ПК-4, ИПК-4.1, ИПК-4.2):

### **Пример практических задач:**

**Примерные задания:** I. Зарегистрируйтесь на сайте Национального корпуса русского языка по адресу: <https://ruscorpora.ru/> и выполните следующие задания:

1. Как выбрать нужный подкорпус или оформить поисковый запрос в зависимости от задач исследования?

• Изучаем концепт СЧАСТЬЕ в русской языковой картине мира. Наши действия?

• Хотим отследить частотность употребления репрезентантов концепта СЧАСТЬЕ за последние два столетия.

• Социолингвистическое исследование – изучение специфики женской и мужской речи и особенности концептуализации. Концепт СЧАСТЬЕ.

• Репрезентация концепта СЧАСТЬЕ в прозе Виктории Токаревой.

• Особенности репрезентации концепта СЧАСТЬЕ в медиа картине мира за последние 10 лет. В художественной картине мира за последние 100 лет?

II. Разработайте собственный корпус, скачав массив текстов информационного новостного агентства, с учетом метарзметки и классов (дата, рубрика, автор и т.п.) , с учетом лингвистической и экстралингвистической раметок. Аргументируйте свой выбор корпуса. Придумайте гипотезу и решите, основываясь на методах квантитативной лингвистики, следующие задачи: частота лексем в корпусе, вхождение и сравнение лексемы или коллокатов. Результаты исследования отразите в презентации.

Пример:

```
import re
import requests
from bs4 import BeautifulSoup as bs
import pandas as pd
```

```

URL = []
with open('links_ria_kultura.txt', 'r+') as f:
    f = [line.rstrip() for line in f]
    for line in f:
        URL.append(line)

extr_title = []
extr_lead = []
extr_text = []
result = []
n = 0

def clean(extr, cleaned):
    for elm in extr:
        elm = str(elm)
        elm = re.sub('<[^>]*>', "", elm)
        elm = re.sub('\xa0', "", elm)
        elm = re.sub('Подписывайтесь на наш канал в Яндекс.Дзен и присоединяйтесь к нашему Telegram-каналу.', "", elm)
        elm = re.sub('Ваш браузер не поддерживает данный формат видео.', "", elm)
        cleaned.append(elm)
    x = ''.join(cleaned)
    return x

for el in URL[:50]:
    r = requests.get(el)
    page_content = r.text
    soup = bs(page_content, "html.parser")
    extr_head = soup.findAll('h1', {'class': 'article__title'})
    cleanhead = []
    head = clean(extr_head, cleanhead)

    extr_lead = soup.findAll('div', {'class': 'article__announce-text'})
    cleanlead = []
    lead = clean(extr_lead, cleanlead)

    extr_text = soup.findAll('div', {'class': 'article__text'})
    if len(extr_text) <= 0:
        extr_text = soup.find_all('p')

    cleantext = []
    news = clean(extr_text, cleantext)
    result.append([head, lead, news])
    n += 1
    print(n)
    print(news)

df = pd.DataFrame(result)
df.columns = ['Head', 'Lead', 'Text']

```

```

df['Rubric'] = 'Культура'
df.to_csv('RIA_KULTURA_1.csv')
Продумайте и разработайте web-интерфейс корпуса, интегрируйте back и фронт в
контейнер Docker.

Пример:
services:
  db:
    image: mcr.microsoft.com/mssql/server:2019-latest
    environment:
      - ACCEPT_EULA=Y
      - SA_PASSWORD=YourStrong!Pass123
      - MSSQL_PID=Express
    ports:
      - "1433:1433"
    volumes:
      - sql_data:/var/opt/mssql
      - ./init.sql:/docker-entrypoint-initdb.d/init.sql
    healthcheck:
      test: ["CMD-SHELL", "/opt/mssql-tools/bin/sqlcmd -S localhost -U sa -P
YourStrong!Pass123 -Q 'SELECT 1' || exit 1"]
      interval: 10s
      timeout: 5s
      retries: 20

  web:
    build: .
    environment:
      - DB_SERVER=db
      - DB_NAME=Cookbook
      - DB_USER=sa
      - DB_PASSWORD=YourStrong!Pass123
    ports:
      - "5000:5000"
    depends_on:
      db:
        condition: service_healthy
    volumes:
      - ./app:/app

  volumes:
    sql_data:

```

#### Критерии оценивания ответа на теоретический вопрос

Оценка	Критерии
	<ol style="list-style-type: none"> <li>Понимание и логика высказывания изученного материала</li> <li>Представление взаимосвязей процесса и взаимосвязи теоретических модулей изучаемого предмета</li> <li>Полнота данных ответов;</li> <li>Аргументированность данных ответов;</li> <li>Правильность ответов на вопросы;</li> </ol>
«зачетно»	Полно и аргументировано даны ответы по содержанию

	задания. Обнаружено понимание материала, может обосновать свои суждения, применить знания на практике, привести необходимые примеры не только по учебнику, но и самостоятельно составленные. Изложение материала последовательно и правильно Ответы обучающегося удовлетворяют тем же требованиям, что и для оценки «отлично», но допускается 1-2 ошибки, которые сам же исправляет.
«не зачтено»	Демонстрация незнания ответа на соответствующее задание, допускаются ошибки в формулировке определений и правил, искажающие их смысл, беспорядочно и неуверенно излагается материал; отмечаются такие недостатки в подготовке, которые являются серьезным препятствием к успешному овладению последующим материалом.

Критерии оценивания практической работы:

Оценка	Критерии
	<ol style="list-style-type: none"> <li>1. Понимание и логика алгоритма работы</li> <li>2. Наличие или отсутствие ошибок в коде</li> <li>3. Полнота решения практических задач</li> <li>4. Своевременность выполнения;</li> <li>5. Умения связать практический материал с теоретическим;</li> <li>6. Понимание базовых формул обработки естественного языка и программирования;</li> </ol>
«зачтено»	Основные требования к решению практических задач выполнены. Продемонстрированы умение анализировать алгоритмы и находить оптимальное количество решений, умение работать с информацией, в том числе умение требовать дополнительную информацию, необходимую для уточнения реализации алгоритма, навыки разработки программного кода; Основные требования к решению практических задач выполнены, но при этом допущены недочеты. В частности, недостаточно раскрыты навыки стиля, недостаточно комментариев
«не зачтено»	Задача не решена, обнаруживается существенное непонимание проблемы

#### **Итоговый проект ( ПК-4, ИПК-4.1, ИПК-4.2 )**

Итоговый проект включает создание датасета текстов, в котором предполагается применение изученных методов с выводом цели, задач, гипотезы и результатов в виде кода и презентации

Выбор тем текстов предусматривает возможность выбора индивидуальной образовательной траектории, связывающую проектную деятельность со смежными гуманитарными дисциплинами. Однако существуют ограничительные критерии реализации проекта, направленные на формирование вышеуказанных компетенций: 1. Сбор и структуризация данных. 2. Наличие двух классов в обучающем датасете. 3. Объем материала не менее 50 тыс текстов на каждый класс. 4. Наличие формальных метрик оценивания качества результата машинного обучения

Сбор и структуризация данных предусматривает свободу выбора темы. В случае, если выбор не был осуществлен, то предлагаются следующие рубрики:

1. Новостные сайты и пресс релизы (классификация и генерация текстов по рубрике и/или информационному агентству)
2. Отзывы к товарам или фильмам (Наличие 3 класса: позитивный, негативный, нейтральный отзывы)
3. Комментарии в социальных сетях
4. Стены сообществ «ВКонтакте» (классификация по рубриками)

Критерии оценивания практической работы:

Оценка	Критерии
	<ol style="list-style-type: none"><li>1. Сформулирована гипотеза проекта</li><li>2. Собран и структурирован датасет</li><li>3. Четкая логика реализации алгоритмов обработки естественного языка в коде;</li><li>4. Правильность ответов на вопросы;</li><li>5. Наличие структурированной презентации</li><li>6. Полнота проекта</li></ol>
«зачтено»	<p>Выполнены все требования к проекту: сформулирована гипотеза, создан прасер сайтов для сбора и структуризации информации, написан код для обработки естественного языка</p> <p>Выполнены все требования к составлению презентаций: дизайн слайдов, логика изложения материала, текст хорошо написан и сформированные идеи ясно изложены и структурированы</p> <p>Существуют незначительные ошибки в проекте, не влияющие на конечный результат. В частности, может быть низкий уровень формальных метрик, неточности в визуализации данных</p> <p>Основные требования к презентациям выполнены, но при этом допущены недочеты. В частности, имеются неточности в изложении материала; отсутствует логическая последовательность в суждениях; не выдержан объем презентации</p>
«не зачтено»	<p>Критические ошибки в коде, гипотеза не подвержена, недостаточное количество или отсутствие обучающей выборки</p> <p>Тема презентации не раскрыта, обнаруживается существенное непонимание проблемы</p>

### **3. Оценочные материалы итогового контроля (промежуточной аттестации) и критерии оценивания**

**Зачет в третьем семестре** состоит из трех частей.

**Первая часть** представляет собой тест из 20 вопросов, проверяющих ПК-4, ИПК-4.1

Ответы на вопросы первой части даются путем выбора из списка предложенных, установления соответствия между объектами или формулирования однословного ответа на вопрос открытого типа.

### Примерный тест

1. Что такое токенизация?

- A. - Процесс преобразования текста в последовательность токенов
- B. - Процесс преобразования токенов в текст
- C. - Процесс преобразования слов в символы
- D. - + Процесс преобразования текста в последовательность токенов, где токеном может быть слово или символ

2. Что такое стемминг?

- A. - Процесс преобразования слов в их базовую форму
- B. - Процесс преобразования слов в их производные формы
- C. - + Процесс преобразования слов в их базовую форму с отбрасыванием окончаний
- D. - Процесс преобразования слов в символы

3. Какие алгоритмы используются для стемминга?

- A. - + Алгоритм Портера
- B. - Алгоритм Леммы
- C. - Алгоритм Байеса
- D. - Алгоритм Дамерау-Левенштейна

4. Что такое лемматизация?

- A. - Процесс преобразования слов в их базовую форму
- B. - + Процесс преобразования слов в их базовую форму с учетом контекста
- C. - Процесс преобразования слов в их производные формы
- D. - Процесс преобразования слов в символы

5. Какие алгоритмы используются для лемматизации?

- A. - Алгоритм Портера
- B. - Алгоритм Леммы
- C. - + Алгоритм Mystem
- D. - Алгоритм Дамерау-Левенштейна

6. Что такое частеречная разметка?

- A. - Процесс преобразования слов в их базовую форму
- B. - Процесс преобразования слов в их производные формы
- C. - + Процесс определения частей речи слов в тексте
- D. - Процесс преобразования слов в символы

7. Какие методы используются для частеречной разметки?

- A. - Алгоритм Портера
- B. - Алгоритм Леммы
- C. - + Методы машинного обучения
- D. - Алгоритм Дамерау-Левенштейна

8. Что такое морфологический анализ?

- A. - Процесс преобразования слов в их базовую форму
  - B. - Процесс преобразования слов в их производные формы
  - C. - Процесс определения частей речи слов в тексте
  - D. - + Процесс анализа грамматических и морфологических характеристик слов
9. Какие методы используются для морфологического анализа?
- A. - Алгоритм Портера
  - B. - Алгоритм Леммы
  - C. - Методы машинного обучения
  - D. - + Морфологический анализаторы, основанные на словарях и правилах
10. Что такое векторное представление слов?
- + Математическое представление слов в виде векторов
  - A. - Процесс преобразования слов в их базовую форму
  - B. - Процесс преобразования слов в их производные формы
  - C. - Процесс определения частей речи слов в тексте
11. Какие алгоритмы используются для получения векторного представления слов?
- A. - + Word2Vec
  - B. - Алгоритм Портера
  - C. - Алгоритм Леммы
  - D. - Алгоритм Дамерау-Левенштейна
12. Что такое модель языка?
- A. - Модель, используемая для обучения морфологического анализа
  - B. - Модель, используемая для обучения стемминга
  - C. - + Математическая модель, описывающая вероятность последовательности слов в языке
  - D. - Модель, используемая для обучения частеречной разметки
13. Какие методы используются для моделирования языка?
- A. - Алгоритм Портера
  - B. - Алгоритм Леммы
  - C. - + N-граммные модели
  - D. - Алгоритм Дамерау-Левенштейна
14. Что такое синтаксический анализ?
- A. - Процесс преобразования слов в их базовую форму
  - B. - Процесс преобразования слов в их производные формы
  - C. - Процесс определения частей речи слов в тексте
  - D. - + Процесс анализа синтаксической структуры предложения
15. Какие методы используются для синтаксического анализа?
- A. - Алгоритм Портера
  - B. - Алгоритм Леммы
  - C. - + Методы статистического анализа
  - D. - Алгоритм Дамерау-Левенштейна
16. Что такое машинный перевод?
- A. - Процесс преобразования слов в их базовую форму

- B. - Процесс преобразования слов в их производные формы  
 C. - Процесс определения частей речи слов в тексте  
 D. - + Процесс автоматического перевода текста с одного языка на другой
17. Какие методы используются для машинного перевода?  
 A. - Алгоритм Портера  
 B. - Алгоритм Леммы  
 C. - + Статистические модели  
 D. - Алгоритм Дамерау-Левенштейна
18. Что такое информационный поиск?  
 A. - Процесс преобразования слов в их базовую форму  
 B. - Процесс преобразования слов в их производные формы  
 C. - Процесс определения частей речи слов в тексте  
 D. - + Процесс поиска и извлечения информации из больших текстовых коллекций
19. Какие методы используются для информационного поиска?  
 A. - Алгоритм Портера  
 B. - Алгоритм Леммы  
 C. - + Модели вероятностного поиска  
 D. - Алгоритм Дамерау-Левенштейна
20. Что такое извлечение информации?  
 A. - Процесс преобразования слов в их базовую форму  
 B. - Процесс преобразования слов в их производные формы  
 C. - Процесс определения частей речи слов в тексте  
 D. - + Процесс автоматического извлечения структурированной информации из текста

**Критерии оценивания тестирования:**

Оценка	Критерии
	<ol style="list-style-type: none"> <li>1. Полнота выполнения тестовых заданий;</li> <li>2. Своевременность выполнения;</li> <li>3. Правильность ответов на вопросы;</li> <li>4. Самостоятельность тестирования</li> </ol>
«зачтено»	<p>Выполнено более 85 % заданий предложенного теста, в заданиях открытого типа дан полный, развернутый ответ на поставленный вопрос</p> <p>Выполнено более 70 % заданий предложенного теста, в заданиях открытого типа дан полный, развернутый ответ на поставленный вопрос; однако были допущены неточности в определении понятий, терминов и др.</p>
«не зачтено»	<p>Выполнено не более 53 % заданий предложенного теста, на поставленные вопросы ответ отсутствует или неполный, допущены существенные ошибки в теоретическом материале</p>

(терминах, понятиях).

**Вторая часть** представляет решение практических задач ( ПК-4, ИПК-4.1, ИПК-4.2):

1. Дан корпус размеченных текстов с омографами «зАмок» и «замОк». Обучите алгоритмы word2vec, FastText, модели Transformer. С помощью формальных метрик оценивания определите лучший результат

2. На основе собранного корпуса из задания 1. На основе корпуса, созданного в задании 1, обучите генеративные языковые модели для заголовков новостей.

3. Дан корпус текстов результатов футбольных матчей российской премьер-лиги. С помощью контекстно-свободных грамматик извлеките следующие сущности: команда 1 – команда 2 – результат встречи – счет.

**Третья часть** ( ПК-4, ИПК-4.1, ИПК-4.2): предполагает выполнение итоговой проектной работы в команде с распределением ролей. Задача: дан корпус художественных текстов и комментариев пользователей к ним. Опишите корпус и создайте приложения оценки художественного произведения. Извлеките персонажей (NER), постройте между ними связи (SNA). Изучите комментарии к книгам и разработайте классификатор отзывов (Sentiment analysis)

Критерии оценивания второй и третьей практических работ идентичны и оцениваются по следующим критериям:

Оценка	Критерии
	<ol style="list-style-type: none"><li>1. Применение и понимание методов NLP</li><li>2. Наличие или отсутствие ошибок в коде</li><li>3. Полнота решения практических задач</li><li>4. Своевременность выполнения</li><li>5. Умения связать практический материал с теоретическим/</li><li>6. Умение презентации проектного материала</li></ol>
«зачтено»	Основные требования к решению практических задач выполнены. Продемонстрированы умение анализировать алгоритмы и находить оптимальное количество решений, умение работать с информацией, в том числе умение требовать дополнительную информацию, необходимую для уточнения реализации алгоритма, навыки разработки программного кода; Основные требования к решению практических задач выполнены, но при этом допущены недочеты. В частности, недостаточно раскрыты навыки стиля, недостаточно комментариев
«не зачтено»	Задача не решена, обнаруживается существенное непонимание проблемы

#### **4. Оценочные материалы для проверки остаточных знаний (сформированности компетенций)**

Первая часть. Решение тестовых задач:

##### **Тест №1**

1. Какие методы используются для извлечения информации?

- A. - Алгоритм Портера
- B. - Алгоритм Леммы
- C. - Методы статистического анализа
- D. - + Методы машинного обучения

2. Что такое автоматическая классификация текстов?
- A. - Процесс преобразования слов в их базовую форму
  - B. - Процесс преобразования слов в их производные формы
  - C. - Процесс определения частей речи слов в тексте
  - D. - + Процесс автоматического присвоения текстам определенных категорий
3. Какие методы используются для автоматической классификации текстов?
- A. - Алгоритм Портера
  - B. - Алгоритм Леммы
  - C. - Методы статистического анализа
  - D. - + Методы машинного обучения
4. Что такое тематическое моделирование?
- A. - Процесс преобразования слов в их базовую форму
  - B. - Процесс преобразования слов в их производные формы
  - C. - Процесс определения частей речи слов в тексте
  - D. - + Процесс выявления скрытых тематик в текстовой коллекции
5. Какие методы используются для тематического моделирования?
- A. - Алгоритм Портера
  - B. - Алгоритм Леммы
  - C. - + Модель LDA (Latent Dirichlet Allocation)
  - D. - Алгоритм Дамерау-Левенштейна
6. Что такое семантический анализ текста?
- A. - Процесс преобразования слов в их базовую форму
  - B. - Процесс преобразования слов в их производные формы
  - C. - Процесс определения частей речи слов в тексте
  - D. - + Процесс анализа смысловых связей и значений слов и выражений
7. Какие методы используются для семантического анализа текста?
- A. - Алгоритм Портера
  - B. - Алгоритм Леммы
  - C. - + Методы семантической векторной близости
  - D. - Алгоритм Дамерау-Левенштейна
8. Что такое определение тональности текста?
- A. - Процесс преобразования слов в их базовую форму
  - B. - Процесс преобразования слов в их производные формы
  - C. - Процесс определения частей речи слов в тексте
  - D. - + Процесс определения позитивной, негативной или нейтральной окраски текста
9. Какие методы используются для определения тональности текста?
- A. - Алгоритм Портера
  - B. - Алгоритм Леммы
  - C. - + Методы машинного обучения
  - D. - Алгоритм Дамерау-Левенштейна
10. Что такое генерация текста?
- A. - Процесс преобразования слов в их базовую форму
  - B. - Процесс преобразования слов в их производные формы

- C. - Процесс определения частей речи слов в тексте
- D. - + Процесс автоматического создания текста компьютерной программой

11. Какие методы используются для генерации текста?

- A. - Алгоритм Портера
- B. - Алгоритм Леммы
- C. - + Рекуррентные нейронные сети
- D. - Алгоритм Дамерау-Левенштейна

12. Что такое извлечение именованных сущностей?

- A. - Процесс преобразования слов в их базовую форму
- B. - Процесс преобразования слов в их производные формы
- C. - Процесс определения частей речи слов в тексте
- D. - + Процесс выделения и классификации именованных сущностей, таких как имена людей, организаций, мест и дат

13. Какие методы используются для извлечения именованных сущностей?

- A. - Алгоритм Портера
- B. - Алгоритм Леммы
- C. - + Методы машинного обучения
- D. - Алгоритм Дамерау-Левенштейна

14. Что такое автоматическая генерация резюме?

- A. - Процесс преобразования слов в их базовую форму
- B. - Процесс преобразования слов в их производные формы
- C. - Процесс определения частей речи слов в тексте
- D. - + Процесс автоматического создания краткого описания навыков и опыта из текста резюме

15. Какие методы используются для автоматической генерации резюме?

- A. - Алгоритм Портера
- B. - Алгоритм Леммы
- C. - + Методы обработки естественного языка
- D. - Алгоритм Дамерау-Левенштейна

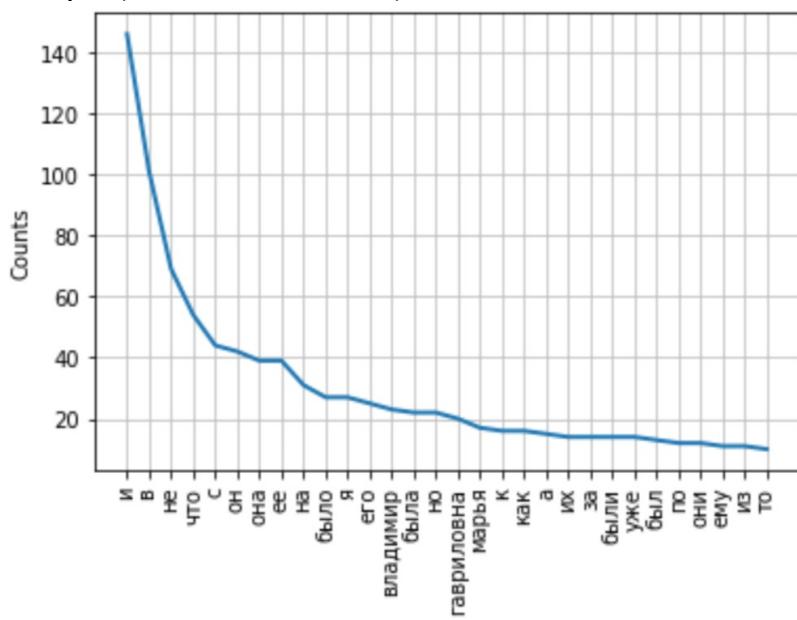
16. Что такое автоматическое рефериование текста?

- A. - Процесс преобразования слов в их базовую форму
- B. - Процесс преобразования слов в их производные формы
- C. - Процесс определения частей речи слов в тексте
- D. - + Процесс автоматического создания краткого изложения текста

Задача 1. Разработайте программу для определения частоты встречаемости слов в тексте.

Ответ

```
f = open('pushkin-metel.txt', "r", encoding="utf-8")
text = f.read()
type(text)
str
len(text)
22968
text[:300]
# перевод в единый регистр (например, нижний)
```



Задача 2. Создайте алгоритм для удаления стоп-слов из предложения.

Ответ:

```
from nltk.corpus import stopwords  
russian_stopwords = stopwords.words("russian")  
russian_stopwords.extend(['Это', 'Но'])
```

```

print(len(russian_stopwords))
# russian_stopwords
153
%%time
text_tokens = [token.strip() for token in text_tokens if token not in russian_stopwords]
Wall time: 6.98 ms
print(len(text_tokens))
2158
text = nltk.Text(text_tokens)
fdist_sw = FreqDist(text)
fdist_sw.most_common(10)
Задача 3. Реализуйте функцию для токенизации слова.
from nltk import word_tokenize
text_tokens = word_tokenize(text)
print(type(text_tokens), len(text_tokens))
text_tokens[:10]
<class 'list'> 3402
['метель',
'кони',
'мчатся',
'по',
'буграм',
'топчут',
'снег',
'глубокой',
'вот',
'в']
import nltk
text = nltk.Text(text_tokens)
print(type(text))
text[:10]
Задача 4. Напишите программу для определения тональности текста
(положительная/отрицательная).

```

Ответ:

```

class TextBlobWrapper():

    def __init__(self):
        self.log = logging.getLogger()
        self.is_model_trained = False
        self.classifier = None

    def init_app(self):
        self.log.info('>>>> TextBlob initialization started')
        self.ensure_model_is_trained()
        self.log.info('>>>> TextBlob initialization completed')

    def ensure_model_is_trained(self):
        if not self.is_model_trained:

            ds = SentimentLabelledDataset()
            ds.load_data()

```

```

        # train the classifier and test the accuracy
        self.classifier = NaiveBayesClassifier(ds.train)
        acr = self.classifier.accuracy(ds.test)
        self.log.info(str.format('>>> NaiveBayesClassifier trained with accuracy {}',
        acr))

        self.is_model_trained = True

        return self.classifier
    
```

Задача 5. Разработайте алгоритм для извлечения именованных сущностей (имена, организаций, места) из текста.

```

import spacy
nlp = spacy.load('en_core_web_sm')
doc = nlp('Dogecoin is a parody cryptocurrency created by software engineer Billy
Markus and Jackson Palmer in 2013.')
for word in doc.ents:
    print(word.text, word.label_)
from spacy.tokens import Span
    
```

```

new_ent = Span(doc, 0, 1, label = "MONEY")
doc.set_ents([new_ent], default = 'unmodified')
for word in doc.ents:
    print(word.text, word.label_)
    
```

Задача 6. Создайте модель классификации текстов на заданное количество категорий.

```
documents = []
```

```
from nltk.stem import WordNetLemmatizer
```

```
stemmer = WordNetLemmatizer()
```

```
for sen in range(0, len(X)):
    # Remove all the special characters
    document = re.sub(r'\W', ' ', str(X[sen]))
```

```
# remove all single characters
document = re.sub(r'\s+[a-zA-Z]\s+', ' ', document)
```

```
# Remove single characters from the start
document = re.sub(r'^[a-zA-Z]\s+', ' ', document)
```

```
# Substituting multiple spaces with single space
document = re.sub(r'\s+', ' ', document, flags=re.I)
```

```
# Removing prefixed 'b'
document = re.sub(r'^b\s+', ' ', document)
```

```
# Converting to Lowercase
document = document.lower()
```

```
# Lemmatization
```

```

document = document.split()

document = [stemmer.lemmatize(word) for word in document]
document = ' '.join(document)

documents.append(document) from sklearn.feature_extraction.text import
CountVectorizer
vectorizer = CountVectorizer(max_features=1500, min_df=5, max_df=0.7,
stop_words=stopwords.words('english'))
X = vectorizer.fit_transform(documents).toarray()
Term frequency = (Number of Occurrences of a word)/(Total words in the document)
IDF(word) = Log((Total number of documents)/(Number of documents containing the
word))
from sklearn.feature_extraction.text import TfidfTransformer
tfidfconverter = TfidfTransformer()
X = tfidfconverter.fit_transform(X).toarray()
from sklearn.feature_extraction.text import TfidfVectorizer
tfidfconverter = TfidfVectorizer(max_features=1500, min_df=5, max_df=0.7,
stop_words=stopwords.words('english'))
X = tfidfconverter.fit_transform(documents).toarray()
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
Задача 7. Напишите программу для определения языка текста.

```

Ответ:

```
from polyglot.detect import Detector
```

```

mixed_text = u"""
China (simplified Chinese: 中国; traditional Chinese: 中國),
officially the People's Republic of China (PRC), is a sovereign state
located in East Asia.
"""

for language in Detector(mixed_text).languages:
    print(language)

```

```

# name: English  code: en    confidence: 87.0 read bytes: 1154
# name: Chinese  code: zh_Hant confidence: 5.0 read bytes: 1755
# name: un       code: un    confidence: 0.0 read bytes: 0

```

Задача 8. Разработайте алгоритм для извлечения ключевых слов из текста.

Ответ:

```
from multi_rake import Rake
```

```

text_en = (
    текст.'
)

```

```
rake = Rake()
```

```
keywords = rake.apply(text_en)
```

```
print(keywords[:10])
```

Задача 9. Создайте модель генерации текста на основе заданного контекста.

```
pip3 install tensorflow==2.0.1 numpy requests tqdm
import tensorflow as tf
import numpy as np
import os
import pickle
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, LSTM, Dropout
from string import punctuation
sequence_length = 100
BATCH_SIZE = 128
EPOCHS = 30
# dataset file path
FILE_PATH = "data/wonderland.txt"
BASENAME = os.path.basename(FILE_PATH)
# read the data
text = open(FILE_PATH, encoding="utf-8").read()
# remove caps, comment this code if you want uppercase characters as well
text = text.lower()
# remove punctuation
text = text.translate(str.maketrans("", "", punctuation))
```

```
# print some stats
n_chars = len(text)
vocab = ''.join(sorted(set(text)))
print("unique_chars:", vocab)
n_unique_chars = len(vocab)
print("Number of characters:", n_chars)
print("Number of unique characters:", n_unique_chars)
```

### Информация о разработчиках

Степаненко Андрей Александрович, старший преподаватель кафедры общей, компьютерной и когнитивной лингвистики