

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

Технологии автоматической обработки текста

по направлению подготовки

45.03.03 Фундаментальная и прикладная лингвистика

Направленность (профиль) подготовки:

Фундаментальная и прикладная лингвистика

Форма обучения

Очная

Квалификация

Бакалавр

Год приема

2025

СОГЛАСОВАНО:

Руководитель ОП

А.В. Васильева

Председатель УМК

Ю.А. Тихомирова

Томск – 2025

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

Целью дисциплины является изучение основных принципов и методов автоматической обработки текстов на естественном языке (ЕЯ)

Целью освоения дисциплины является формирование следующих компетенций:

ПК-4. Способен разрабатывать программный код при решении задач автоматической обработки текстов

ОПК-2. Способен к ведению профессиональной деятельности с опорой на основы математических дисциплин, необходимых для формализации лингвистических знаний и процедур анализа и синтеза лингвистических структур

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИПК-4.1. Применяет способы формализации и алгоритмизации поставленных задач в сфере автоматической обработки текстов

ИПК-4.2. Создает программный код с использованием языков программирования и манипулирования данными в сфере автоматической обработки текстов

ИОПК-2.3 Применяет математические методы в процедурах анализа и синтеза лингвистических структур

2. Оценочные материалы текущего контроля и критерии оценивания

К числу форм контроля, оценивающих уровень достижения компетенций по текущей дисциплине, относятся тесты, устные вопросы, выполнение практических заданий (ПК-4, ИПК-4.1), итоговый проект (продукт (ПК-4, ОПК-2, ИПК-4.1, ИПК-4.2, ИОПК-2.3), код (ИПК-4.2), контрольная и самостоятельная работы (ИПК-4.1).

Итоговый проект включает блок вопросов по теоретической части изучаемой дисциплины (ПК-4, ОПК-2, ИПК-4.1, ИПК-4.2, ИОПК-2.3):

Пример практических задач:

1. Реализуйте программу для анализа тональности отзывов на продукты/услуги.
2. Напишите алгоритм для определения синонимов и антонимов слова.
3. Разработайте модель для автоматического перевода текста с одного языка на другой
4. Создайте алгоритм для определения частей речи слова в предложении.
5. Реализуйте программу для извлечения информации из текста (например, даты, адреса, номера телефонов).
6. Напишите алгоритм для определения семантической близости между двумя предложениями.
7. Разработайте модель для генерации автоматических ответов на основе заданных вопросов.
8. Создайте программу для определения настроения пользователя на основе его сообщений.
9. Реализуйте алгоритм для автоматического суммирования текста.
10. Напишите модель для автоматического распознавания и классификации именованных сущностей в тексте.

Ответы:

Теоретические вопросы (ПК-4, ОПК-2, ИПК-4.1):

Критерии оценивания ответа на теоретический вопрос

Оценка	Критерии
	1. Понимание и логика высказывания изученного материала

	2. Представление взаимосвязей процесса и взаимосвязи теоретических модулей изучаемого предмета 3. Полнота данных ответов; 4. Аргументированность данных ответов; 5. Правильность ответов на вопросы;
«зачтено»	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. Обучающийся демонстрирует знание и понимание основных положений данного задания, но: 1) излагает материал неполно и допускает неточности в определении понятий или формулировке правил; 2) не умеет достаточно глубоко и доказательно обосновать свои суждения и привести свои примеры; 3) излагает материал непоследовательно и допускает ошибки.
«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
	1. Понимание и логика алгоритма работы 2. Наличие или отсудив ошибок в коде 3. Полнота решения практических задач 4. Своевременность выполнения; 5. Умения связать практический материал с теоретическим; 6. Понимание базовых формул обработки естественного языка и программирования;
«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 затребовать 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 Имеются существенные отступления от решения практических задач. В частности отсутствуют навык и умения моделировать решения в соответствии с заданием, представлять различные подходы к разработке алгоритмов,

	ориентированных на конечный результат
«не зачтено»	Задача не решена, обнаруживается существенное непонимание проблемы

Итоговый проект (ПК-4, ОПК-2, ИПК-4.1, ИПК-4.2, ИОПК-2.3)

Итоговый проект включает создание корпуса текстов. К числу тем могут относиться тексты социальных сетей, отзывы, литературных произведений и т.п..

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, наличие числовых переменных.

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
	1. Сформулирована гипотеза проекта 2. Собран и структурирован датасет 3. Четкая логика реализации алгоритмов обработки естественного языка в коде; 4. Правильность ответов на вопросы; 5. Наличие структурированной презентации 6. Полнота проекта
«зачтено»	Выполнены все требования к проекту: сформулирована гипотеза, создан прасер сайтов для сбора и структуризации информации, написан код для обработки естественного языка Выполнены все требования к составлению презентаций: дизайн слайдов, логика изложения материала, текст хорошо написан и сформированные идеи ясно изложены и структурированы Существуют незначительные ошибки в проекте, не влияющие на конечный результат. В частности, может быть низкий уровень формальных метрик, неточности в визуализации данных 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 Существенные ошибки в коде, отсекает логика представления проекта, недостаточный объем датасета Имеются существенные отступления от требований к презентациям. В частности: тема освещена лишь частично; допущены фактические ошибки в содержании презентаций

	или при ответе на дополнительные вопросы.
«не зачтено»	Критические ошибки в коде, гипотеза не подвержена, недостаточное количество или отсутствие обучающей выборки Тема презентации не раскрыта, обнаруживается существенное непонимание проблемы

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

Зачет в пятом семестре состоит из трех частей.

Первая часть представляет собой тест из 20 вопросов, проверяющих ИПК-4.2, ИОПК-2.3

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
	1. Полнота выполнения тестовых заданий; 2. Своевременность выполнения; 3. Правильность ответов на вопросы; 4. Самостоятельность тестирования

«зачтено»	Выполнено более 85 % заданий предложенного теста, в заданиях открытого типа дан полный, развернутый ответ на поставленный вопрос 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. Выполнено более 54 % заданий предложенного теста, в заданиях открытого типа дан неполный ответ на поставленный вопрос, в ответе не присутствуют доказательные примеры, текст со стилистическими и орфографическими ошибками.
«не зачтено»	Выполнено не более 53 % заданий предложенного теста, на поставленные вопросы ответ отсутствует или неполный, допущены существенные ошибки в теоретическом материале (терминах, понятиях).

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

Первая часть. Решение тестовых задач:

Тест №1

- Какие методы используются для извлечения информации?
 - Алгоритм Портера
 - Алгоритм Леммы
 - Методы статистического анализа
 - + Методы машинного обучения
- Что такое автоматическая классификация текстов?
 - Процесс преобразования слов в их базовую форму
 - Процесс преобразования слов в их производные формы
 - Процесс определения частей речи слов в тексте
 - + Процесс автоматического присвоения текстам определенных категорий
- Какие методы используются для автоматической классификации текстов?
 - Алгоритм Портера
 - Алгоритм Леммы
 - Методы статистического анализа
 - + Методы машинного обучения
- Что такое тематическое моделирование?
 - Процесс преобразования слов в их базовую форму
 - Процесс преобразования слов в их производные формы
 - Процесс определения частей речи слов в тексте
 - + Процесс выявления скрытых тематик в текстовой коллекции
- Какие методы используются для тематического моделирования?
 - Алгоритм Портера
 - Алгоритм Леммы
 - + 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 реферирование 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.

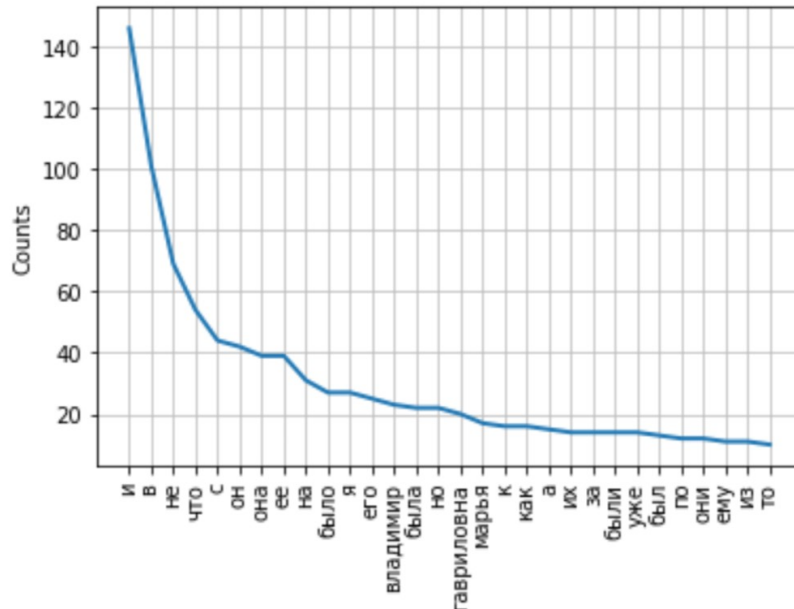
Ответ

```
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)
text = text.lower()
import string
string.punctuation
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
type(string.punctuation)
str
spec_chars = string.punctuation + '\n\xa0«»\t—...'
%%time
text = "".join([ch for ch in text if ch not in spec_chars])
Wall time: 4 ms
import re
text = re.sub("\n", "", text)
def remove_chars_from_text(text, chars):
    return "".join([ch for ch in text if ch not in chars])
%%time
text = remove_chars_from_text(text, spec_chars)
Wall time: 4.02 ms
%%time
text = remove_chars_from_text(text, string.digits)
%%time
from nltk.probability import FreqDist
fdist = FreqDist(text)
```

```

fdist
Wall time: 6.98 ms
FreqDist({'и': 146, 'в': 101, 'не': 69, 'что': 54, 'с': 44, 'он': 42, 'она': 39, 'ее': 39, 'на': 31,
'было': 27, ...})
fdist.most_common(5)
[('и', 146), ('в', 101), ('не', 69), ('что', 54), ('с', 44)]
fdist.plot(30, cumulative=False)

```



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', 'нею'])
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, организации, 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