МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное автономное образовательное учреждение высшего образования «Балтийский федеральный университет имени Иммануила Канта»
Высшая школа компьютерных наук и прикладной математики

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

«Основы обработки текстов»

Шифр: 01.03.02
Направление подготовки: «Прикладная математика и информатика»
Профиль: «Искусственный интеллект и анализ данных»

Квалификация (степень) выпускника: бакалавр

Калининград
2023

Лист согласования

Составители:

1. Верещагин Сергей Дмитриевич, к. ф.-м.н., доцент
2. Верещагин Михаил Дмитриевич, к. ф.-м.н., доцент
3. Мищук Богдан Ростиславович, к. ф.-м.н., доцент

Рабочая программа утверждена на заседании
Ученого совета ОНК «Институт высоких технологий»

Протокол № 4 от «24» января 2023 г.

Председатель Ученого совета ОНК
«Институт высоких технологий»

Профессор, д.ф.-м.н. А.В. Юров

Руководитель ОПОП ВО Е.П. Ставицкая

Содержание

1.Наименование дисциплины: «Основы обработки текстов».

Целью курса «Основы обработки текстов» - дать обзор современных подходов к применению искусственных нейронных сетей в задачах анализа текстов на естественном языке.

Изучаются основные возможности библиотеки Tensorflow для проектирования и обучения нейронных сетей. Формируется владение подходами к разработке приложений и модулей обработки текстов на естественном языке, навыки проектирования и обучения искусственных нейронных сетей для решения задач обработки текстов.

2. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

Код компетенции	Результаты освоения образовательной программы (ИДК)	Результаты обучения по дисциплине
ПК-9. Способен создавать и внедрять одну или несколько сквозных цифровых субтехнологий искусственного интеллекта	ПК-9.2. Участвует в реализации проектов в области сквозной цифровой субтехнологии «Обработка естественного языка»	Знать: 1. фундаментальные понятия и идеи в области компьютерной обработки текстов 2. современные направления исследований в данной области 3. основные проблемы, возникающие при обработке текстов. Уметь: 1. решать задачи из области обработки текстов 2. проектировать системы для анализа отдельных текстовых документов и коллекций текстовых документов 3. применять методы статистического анализа и машинного обучения для решения прикладных задач области. Владеть: 1. современными технологиями и программными инструментами для обработки текстов.

3. Место дисциплины в структуре образовательной программы

«Основы обработки текстов» представляет собой дисциплину Части, формируемой участниками образовательных отношений (Б1.В.ДВ.02.01), дисциплина по выбору, направления подготовки бакалавриата 01.03.02 «Прикладная математика и информатика», профиль «Искусственный интеллект и анализ данных».

4. Виды учебной работы по дисциплине.

Виды учебной работы по дисциплине зафиксированы учебным планом основной профессиональной образовательной программы по указанному направлению и профилю, выражаются в академических часах. Часы контактной работы и самостоятельной работы обучающегося и часы, отводимые на процедуры контроля, могут различаться в учебных планах ОПОП по формам обучения. Объем контактной работы включает часы контактной аудиторной работы (лекции/практические занятия/ лабораторные работы), контактной внеаудиторной работы (контроль самостоятельной работы), часы контактной работы в период аттестации. Контактная работа, в том числе может проводиться посредством электронной информационно-образовательной среды университета с использованием ресурсов сети Интернет и дистанционных технологий

5. Содержание дисциплины, структурированное по темам (разделам)

Исходя из рамок, установленных учебным планом по трудоемкости и видам учебной работы по дисциплине, преподаватель самостоятельно выбирает тематику занятий по формам и количеству часов проведения контактной работы: лекции и иные учебные занятия, предусматривающие преимущественную передачу учебной информации преподавателем и (или) занятия семинарского типа (семинары, практические занятия, практикумы, лабораторные работы, коллоквиумы и иные аналогичные занятия), и (или) групповые консультации, и (или) индивидуальную работу обучающихся с преподавателем, в том числе индивидуальные консультации (по курсовым работам/проектам – при наличии курсовой работы/проекта по данной дисциплине в учебном плане). Рекомендуемая тематика занятий максимально полно реализуется в контактной работе с обучающимися очной формы обучения. В случае реализации образовательной программы в заочной / очно-заочной форме трудоемкость дисциплины сохраняется, однако объем учебного материала в значительной части осваивается обучающимися в форме самостоятельной работы. При этом требования к ожидаемым образовательным результатам обучающихся по данной дисциплине не зависят от формы реализации образовательной программы.

№ п/п	Наименование разделов (тем) дисциплины	Содержание разделов (тем) дисциплин
1.	Введение. Задачи обработки текста.	Многозначность при обработке текста. Проблема понимания. Тест Тьюринга. Китайская комната
2.	Регулярные выражения и конечные автоматы.	Распознавание языка с помощью КА. Построение КА для регулярных выражений
3.	Методы поиска словосочетаний.	Общая схема. Методы поиска кандидатов. Проверка статистических гипотез.
4.	Языковые модели и задача определения частей речи.	Модель N-грамм. Оценка вероятности высказывания. Методы сглаживания. Оценка качества. Тренировочный и проверочный корпуса. Задача определения частей речи. Существующие подходы. Алгоритмы, основанные на правилах. Алгоритмы, основанные на трансформации..
5.	Скрытые марковские модели.	Вероятность последовательности. Прямой алгоритм. Наиболее правдоподобное объяснение. Использование скрытой марковской модели для определения частей речи. Алгоритм Витерби. Методы классификации документов. Наивный

		байесовский классификатор. Логистическая регрессия. Модель максимальной энтропии
6.	Контекстно-свободные грамматики и синтаксический анализ.)	Типы грамматик. Грамматика составляющих. Грамматика зависимостей. Категориальная грамматика. Контекстно-свободные грамматики. КС грамматики и регулярные языки. Банк деревьев. синтаксический разбор. Разбор сверху вниз и снизу вверх. Алгоритм Кока-Янгера-Касами (CKY parsing). Эквивалентность КС грамматик. Группировка (chunking
7.	Статистические методы синтаксического анализа.	Стохастические контекстно-свободные грамматики. Разрешение синтаксической многозначности. Моделирование языка. Обучение стохастических КС грамматик. Вероятностная версия алгоритма Кока-Янгера-Касами. Оценка качества. Проблемы стохастический КС грамматик. Алгоритм Коллинза.
8.	Лексическая семантика. WordNet.	Значения слов. Разрешение лексической многозначности. Алгоритмы классификации. Самонастройка. Методы основанные на словарях и тезаурусах. Варианты алгоритма Леска. Методы оценки качества. Семантическая близость слов. Подходы на основе тезаурусов. Подходы на основе статистик. Методы оценки качества.
9.	Информационный поиск.	Ранжирование документов. Векторная модель. Взвешивание терминов. Индексирование. Инвертированный индекс. Запросы с джокером. Исправление опечаток
10.	Вопросно-ответные системы.	Общая архитектура. Обработка запроса. Извлечение фрагментов текста. Обработка ответа. Автоматическое реферирование. Общая архитектура.
11.	Машинный перевод.	Классические подходы. Статистический машинный перевод. Модель зашумленного канала. Модель перевода на основе фраз. Выравнивание фраз. Декодирование. Выравнивание слов. Модель IBM Model 1. Тренировка моделей выравнивания.
12.	Тематическое моделирование.	Вероятностная латентная семантическая модель. Латентное размещение Дирихле. Робастные модели.

6. Рекомендуемая тематика учебных занятий в форме контактной работы

Рекомендуемая тематика учебных занятий *лекционного* типа (предусматривающих преимущественную передачу учебной информации преподавателями):

№ п/п	Наименование разделов (тем) дисциплины	Содержание разделов (тем) дисциплин
1	Введение. Задачи обработки текста.	Лекция 1. Многозначность при обработке текста. Проблема понимания.
2	Регулярные выражения и конечные автоматы.	Лекция 2. Распознавание языка с помощью КА.
3	Методы поиска словосочетаний.	Лекция 3. Методы поиска кандидатов. Проверка статистических гипотез.
4	Языковые модели и задача определения частей речи.	Лекция 4. Модель N-грамм. Лекция 5. Задача определения частей речи.
5	Скрытые марковские модели.	Лекция 6. Вероятность последовательности. Прямой алгоритм. Лекция 7. Методы классификации документов. Наивный байесовский классификатор.
6	Контекстно-свободные грамматики и синтаксический анализ.)	Лекция 8. Типы грамматик. Лекция 9. Банк деревьев. синтаксический разбор.
7	Статистические методы синтаксического анализа.	Лекция 10. Стохастические контекстно-свободные грамматики. Лекция 11. Вероятностная версия алгоритма Кока-Янгера-Касами.
8	Лексическая семантика. WordNet.	Лекция 12. Алгоритмы классификации. Самонастройка. Лекция 13. Семантическая близость слов. Подходы на основе тезаурусов. Подходы на основе статистик. Методы оценки качества.
9	Информационный поиск.	Лекция 14. Ранжирование документов. Векторная модель.
10	Вопросно-ответные системы.	Лекция 15. Общая архитектура. Обработка запроса.
11	Машинный перевод.	Лекция 16. Классические подходы. Статистический машинный перевод. Лекция 17. Выравнивание фраз. Декодирование.
12	Тематическое моделирование.	Лекция 18. Вероятностная латентная семантическая модель.

Рекомендуемая тематика *практических* занятий:

№ п/п	Наименование разделов (тем) дисциплины	Содержание разделов (тем) дисциплин
1	Методы поиска словосочетаний.	Методы поиска кандидатов. Проверка статистических гипотез.
2	Языковые модели и задача определения частей речи.	Модель N-грамм. Задача определения частей речи.

3	Контекстно-свободные грамматики и синтаксический анализ.)	Типы грамматик. Банк деревьев. синтаксический разбор.
4	Статистические методы синтаксического анализа.	Стохастические контекстно-свободные грамматики. Вероятностная версия алгоритма Кока-Янгера-Касами.
5	Лексическая семантика. WordNet.	Алгоритмы классификации. Самонастройка. Семантическая близость слов. Подходы на основе тезаурусов. Подходы на основе статистик. Методы оценки качества.
6	Информационный поиск.	Ранжирование документов. Векторная модель.
7	Вопросно-ответные системы.	Общая архитектура. Обработка запроса.
8	Машинный перевод.	Классические подходы. Статистический машинный перевод. Выравнивание фраз. Декодирование.
9	Тематическое моделирование.	Вероятностная латентная семантическая модель.

Требования к самостоятельной работе обучающихся

1. Работа с лекционным материалом, предусматривающая проработку конспекта лекций и учебной литературы, по всем темам из п. 6 настоящей рабочей программы.
2. Выполнение домашнего задания, предусматривающего решение задач, выполнение упражнений, выдаваемых на практических занятиях, по всем темам из п. 6 настоящей рабочей программы.

Руководствуясь положениями статьи 47 и статьи 48 Федерального закона от 29 декабря 2012 г. N 273-ФЗ «Об образовании в Российской Федерации» научно-педагогические работники и иные лица, привлекаемые университетом к реализации данной образовательной программы, пользуются предоставленными академическими правами и свободами в части свободы преподавания, свободы от вмешательства в профессиональную деятельность; свободы выбора и использования педагогически обоснованных форм, средств, методов обучения и воспитания; права на творческую инициативу, разработку и применение авторских программ и методов обучения и воспитания в пределах реализуемой образовательной программы и отдельной дисциплины.

Исходя из рамок, установленных учебным планом по трудоемкости и видам учебной работы по дисциплине, преподаватель самостоятельно выбирает тематику занятий по формам и количеству часов проведения контактной работы: лекции и иные учебные занятия, предусматривающие преимущественную передачу учебной информации преподавателем и (или) занятия семинарского типа (семинары, практические занятия, практикумы, лабораторные работы, коллоквиумы и иные аналогичные занятия), и (или) групповые консультации, и (или) индивидуальную работу обучающихся с преподавателем, в том числе индивидуальные консультации (по курсовым работам/проектам – при наличии курсовой работы/проекта по данной дисциплине в учебном плане).

Рекомендуемая тематика занятий максимально полно реализуется в контактной работе с обучающимися очной формы обучения. В случае реализации образовательной программы в заочной / очно-заочной форме трудоемкость дисциплины сохраняется, однако объем учебного материала в значительной части осваивается обучающимися в форме

самостоятельной работы. При этом требования к ожидаемым образовательным результатам обучающихся по данной дисциплине не зависят от формы реализации образовательной программы.

7. Методические рекомендации по видам занятий

Лекционные занятия.

В ходе лекционных занятий обучающимся рекомендуется выполнять следующие действия. Вести конспектирование учебного материала. Обращать внимание на категории, формулировки, раскрывающие содержание тех или иных явлений и процессов, научные выводы и практические рекомендации по их применению. Задавать преподавателю уточняющие вопросы с целью уяснения теоретических положений, разрешения спорных ситуаций.

Желательно оставить в рабочих конспектах поля, на которых во внеаудиторное время можно сделать пометки из рекомендованной литературы, дополняющие материал прослушанной лекции, а также подчеркивающие особую важность тех или иных теоретических положений.

Практические и семинарские занятия.

На практических и семинарских занятиях в зависимости от темы занятия выполняется поиск информации по решению проблем, практические упражнения, контрольные работы, выработка индивидуальных или групповых решений, итоговое обсуждение с обменом знаниями, участие в круглых столах, разбор конкретных ситуаций, командная работа, представление портфолио и т.п.

Самостоятельная работа.

Самостоятельная работа осуществляется в виде изучения литературы, эмпирических данных по публикациям и конкретных ситуаций из практики, подготовке индивидуальных работ, работа с лекционным материалом, самостоятельное изучение отдельных тем дисциплины; поиск и обзор литературы и электронных источников; чтение и изучение учебника и учебных пособий.

8. Фонд оценочных средств

8.1. Перечень компетенций с указанием этапов их формирования в процессе освоения образовательной программы в рамках учебной дисциплины

Основными этапами формирования указанных компетенций при изучении обучающимися дисциплины являются последовательное изучение содержательно связанных между собой тем учебных занятий. Изучение каждой темы предполагает овладение обучающимися необходимыми компетенциями. Результат аттестации обучающихся на различных этапах формирования компетенций показывает уровень освоения компетенций.

Контролируемые разделы (темы) дисциплины	Индекс контролируемой компетенции (или её части)	Оценочные средства по этапам формирования компетенций
		текущий контроль по дисциплине
Введение. Задачи обработки текста.	ПК-9	Решение индивидуальных заданий
Регулярные выражения и конечные автоматы.	ПК-9	Решение индивидуальных заданий

Контролируемые разделы (темы) дисциплины	Индекс контролируемой компетенции (или её части)	Оценочные средства по этапам формирования компетенций
		текущий контроль по дисциплине
Методы поиска словосочетаний.	ПК-9	Решение индивидуальных заданий
Языковые модели и задача определения частей речи.	ПК-9	Решение индивидуальных заданий
Скрытые марковские модели.	ПК-9	Решение индивидуальных заданий
Контекстно-свободные грамматики и синтаксический анализ.)	ПК-9	Решение индивидуальных заданий
Статистические методы синтаксического анализа.	ПК-9	Решение индивидуальных заданий
Лексическая семантика. WordNet.	ПК-9	Решение индивидуальных заданий
Информационный поиск.	ПК-9	Решение индивидуальных заданий
Вопросно-ответные системы.	ПК-9	Решение индивидуальных заданий
Машинный перевод.	ПК-9	Решение индивидуальных заданий
Тематическое моделирование.	ПК-9	Решение индивидуальных заданий

8.2. Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений и навыков и (или) опыта деятельности процессе текущего контроля

Текущий контроль успеваемости осуществляется путем оценки результатов выполнения заданий практических (семинарских) занятий, самостоятельной работы, предусмотренных учебным планом и посещения занятий/активность на занятиях.

В качестве оценочных средств текущего контроля успеваемости предусмотрены:

Решение индивидуальных заданий

Примеры индивидуальных заданий

Постановка задачи

Целью работы является создание системы, позволяющей оценивать эмоциональную окраску сообщений микроблога Twitter. Задача анализа эмоциональной окраски (или тональности) текста (англ. sentiment analysis или opinion mining) состоит в автоматическом выявлении в текстах эмоциональной оценки автора по отношению к некоторому объекту.

Примеры:

- "Начинается новый день" - нейтральная оценка
- "Какой прекрасный день" - позитивная
- "Ужасный день" - негативная

Предлагается разработать систему, которая на вход получает короткий текст (твит), а на выходе отдает одну из трех меток:

- neutral - нейтральная
- positive - позитивная
- negative - негативная

Решение задачи

Практические аспекты

Решения должны быть написаны на языке Python. Можно использовать все стандартные библиотеки, а также

- NLTK - инструменты для обработки текстов
- scikit-learn - алгоритмы машинного обучения
- numpy - работа с многомерными массивами

Доступ в Интернет на проверяющей машине закрыт.

Теоретические аспекты

Предполагается использование алгоритмов машинного обучения. Для обучения алгоритма требуется придумать признаки и дать ему на вход правильные примеры - обучающий корпус. Считается, что чем больше обучающий корпус, тем лучше работает алгоритм. Однако создание большого обучающего корпуса - довольно трудоемкая задача, непосильная одному человеку. Поэтому предлагается создать его с помощью коллективной работы. Чтобы облегчить эту работу, был сделан сайт: http://markup.at.ispras.ru.

Разметка обучающего корпуса

Для разметки корпуса необходимо зарегистрироваться на сайте http://markup.at.ispras.ru. Пожалуйста, вводите правильные данные, так как они будут использоваться при выставлении зачетов. Вне рамок практикума эти данные использоваться не будут.

После регистрации появится окно с тремя полями (Рис 1).

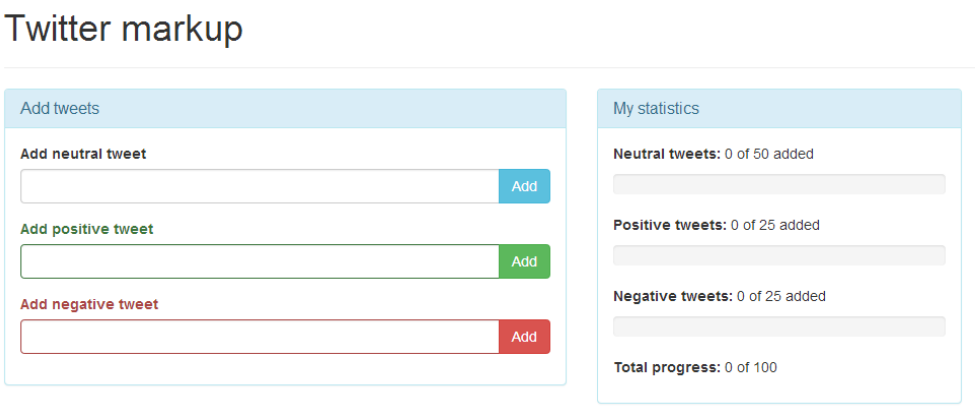

Рис 1. Форма для ввода твитов

Далее вы идете на сайт twitter.com выбираете любых пользователей, которые пишут **на русском языке**. Ищете у них сообщения, которые содержат эмоциональную окраску (либо однозначно позитивную, либо однозначно негативную) или не содержат никаких эмоций (например, констатация факта). Копируете эти сообщения в соответствующие поля формы и нажимаете add (enter тоже работает). Если есть сомнение в передаваемых эмоциях, лучше пропустить твит (см. следующий раздел). После того, как будет размечено не менее 25 позитивных, 25 негативных и 50 нейтральных сообщений, появится кнопка, позволяющая скачать размеченные твиты (см. раздел "тренировочный корпус").

Под формой ввода твитов будут появляться все твиты, размеченные и вашими коллегами. Эмоциональная окраска будет отмечена разными цветами (зеленый - позитивная, красный - негативная, серый - нейтральная). Вы можете оценить правильность разметки, нажав соответствующую кнопку рядом сообщением. Эта информация будет доступна в файле с тренировочным корпусом и ее можно использовать при обучении.

Какие твиты стоит размечать

Для разметки следует использовать твиты на русском языке. То есть твиты могут содержать иностранные названия, но основная часть текста должна быть представлена на русском языке кириллическими буквами.

ни дня без хорошей новости - recent studies suggest that sexual activity causes neurogenesis in the hippocampus.	Этот твит не считается твитом на русском языке. Такие твиты добавлять не стоит.
Мне понравилось видео "Anime | THE SIMPSONS | ANIMATION on FOX" (http://youtu.be/R94Q6NhuS3A?a) на @YouTube.	Этот твит нельзя считать твитом на русском языке, поскольку слов на английском языке значительно больше, чем на русском.
Жители Сан-Франциско атаковали автобус Google http://bit.ly/JZ9iZF	Этот твит можно считать твитом на русском языке, поскольку он содержит небольшое количество иностранных слов, которые являются именами собственными

При оценке эмоциональной окраски твита следует учитывать только субъективное мнение **автора текста** (твита) по отношению к описываемому объекту/явлению. Твиты могут содержать как явную эмоциональную окраску, так и не явную. Не следует путать печальные сообщения (по смыслу) с негативной эмоциональной окраской сообщения.

Фотосессии на ВМК всегда весёлые! http://instagram.com/p/jwm9GKx3Q5/	Это сообщение имеет явную положительную эмоциональную окраску. Такие твиты стоит добавлять.
Кочкин становится свидетелем того, как кортеж олигарха сбивает женщину, переходящую дорогу. #photo	Это эмоционально нейтральное сообщение. Несмотря на то, что описывает оно печальное событие.
Никогда ничего не покупайте в магазине Pleer.ru http://j.mp/KKiqSe	Это сообщение имеет неявную отрицательную эмоциональную окраску.

Не стоит добавлять твиты, содержащие одновременно и негативную и позитивную оценку каких-либо объектов

С одной стороны, я бы не хотела жить одна - скучно и одиноко, но с другой стороны - чистота в доме, свобода, не нужно готовить - рай.	Это сообщение содержит в себе две эмоционально окрашенные части: первая – отрицательная, вторая – положительная. Однозначно трактовать эмоциональную окраску всего предложения не возможно. Такие твиты добавлять не стоит.

Не стоит добавлять твиты, содержащие сарказм:

Члены партии единой России обладают великим искусством правильно подобрать варианты ответа к опросу http://er.ru/poll/5.html/	Сарказм. Без дополнительных знаний о контексте этого сообщения невозможно определить положительное оно или

	отрицательное. Такие сообщения добавлять не стоит.

Рекомендуется размечать максимально честно, так как от этого будет зависеть качество всех классификаторов. Если есть сомнения, к какому классу лучше отнести сообщение, то его стоит пропустить.

Тренировочный корпус

Тренировочный корпус будет доступен для скачивания в формате json. Для извлечение информации из этого файла рекомендуется использовать стандартную библиотеку Python с одноименным названием.

Для синхронизации обучения и тестирования в течении недели, корпус будет состоять из твитов, размеченных автором классификатора, плюс все твиты, размеченные в течении предшествующей недели.

Тестирование

Вместе с кнопкой скачивания тренировочного корпуса появится ссылка на форму для загрузки файла и личную страницу со статистикой. На личной странице находится статистика со всеми результатами в т.ч. результатами последнего тестирования (дата, описание, достоверность).

Загрузка решения. Загружаемый файл должен представлять собой zip архив с любым именем. Архив должен обязательно содержать:

- классификатор в файле SentimentAnalyzer.py. В файле должен содержаться класс SentimentAnalyzer. В классе должны присутствовать методы
    - train(self, training_corpus), где training_corpus - это список пар (text, class). Внимание: метод train будет вызываться отдельно, так что не стоит вызывать его в конструкторе класса.
    - getClasses (self,texts), который получает на вход список текстовых сообщений и возвращает список ответов классификатора. (Пример: [neutral, positive, positive, ...])
- описание применяемых алгоритмов в файле description.txt
- все используемые внешние библиотеки, кроме библиотек пакетов NLTK, scikit-learn и numpy (они доступны автоматически).

Результаты тестирования появятся на личной странице, как только закончится обучение и тестирование. При загрузке нового классификатора обучение будет производится на корпусе из твитов, размеченных автором классификатора, плюс все твиты, размеченные в течении предшествующей загрузке недели.

В течении недели студенты не видят прогресс своих коллег и могут посмотреть только свой результат. В конце каждой недели (каждый вторник в 23.59.59) будет производится переобучение последнего присланного решения от каждого студента на новом корпусе, а результаты тестирования будут показаны в сводной таблице.

Ограничения

1. каждую неделю можно послать только 10 версий программы (внимание! Итоговое тестирование будет проводится на последнем загруженном решении)
2. размер архива не может превышать 15Мб

В связи с первым ограничением, для тестирования на локальной машине рекомендуется использовать метод перекрестной проверки (http://en.wikipedia.org/wiki/Cross-

validation_(statistics)). В библиотеке scikit-learn есть функции, которые могут помочь в использовании этого метода. Рекомендуется использовать метод StratifiedKFold().

Оценка качества

Для оценки качества используются метра достоверности (accuracy), которая равна отношению количества правильных ответов к общему количеству примеров в тестовой выборке.

accuracy $= \frac{correct\ answers}{total\ questions}$

Описание в документации к библиотеке scikit-learn: http://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score

Baseline

Baseline 1. В качестве нижней границы используется классификатор, который дает всегда ответ "neutral". Достоверность этого метода равна 0.5.

Baseline 2. В качестве второй, более сложной нижней границы используется один из стандартных алгоритмов классификации с N-граммами в качестве признаков. Этот классификатор будет тренироваться на том же корпусе, что и присланные алгоритмы, и его достоверность будет меняться соответственно.

8.3. Перечень вопросов и заданий для промежуточной аттестации по дисциплине

Вопросы к экзамену.

1. Задачи обработки текста. Многозначность при обработке текста. Проблема понимания. Тест Тьюринга. Китайская комната
2. Регулярные выражения
3. Конечные автоматы, распознавание языка с помощью КА
4. Регулярные языки и конечные автоматы. Построение КА для регулярных выражений
5. Методы поиска словосочетаний. Общая схема. Методы поиска кандидатов
6. Методы поиска словосочетаний. Проверка статистических гипотез
7. Модель N-грамм. Оценка вероятности высказывания
8. Модель N-грамм. Сглаживание (Лапласа и Откат)
9. Модель N-грамм. Оценка качества. Тренировочный и проверочный корпуса
10. Задача определения частей речи. Существующие подходы. Алгоритмы, основанные на правилах. Алгоритмы, основанные на трансформации.
11. Использование скрытой марковской модели для определения частей речи.
12. Скрытые марковские модели. Вероятность последовательности. Прямой алгоритм
13. Скрытые марковские модели. Наиболее правдоподобное объяснение. Алгоритм Витерби
14. Модели классификации. Наивный байесовский классификатор
15. Модели классификации. Логистическая регрессия
16. Модели классификации. Модель максимальной энтропии
17. Модели классификации. Марковская модель максимальной энтропии
18. Типы грамматик. Грамматика составляющих. Грамматика зависимостей. Категориальная грамматика

19. Контекстно-свободные грамматики. КС грамматики и регулярные языки. Банк деревьев.
20. Синтаксический разбор. Разбор сверху вниз и снизу вверх
21. Синтаксический разбор. Алгоритм Кока-Янгера-Касами (CKY parsing). Эквивалентность КС грамматик
22. Синтаксический разбор. Группировка (chunking)
23. Стохастические контекстно-свободные грамматики. Разрешение синтаксической многозначности
24. Моделирование языка. Обучение стохастических КС грамматик
25. Вероятностная версия алгоритма Кока-Янгера-Касами. Оценка качества
26. Проблемы стохастический КС грамматик. Алгоритм Коллинза. Оценка качества
27. Лексическая семантика. WordNet. Значения слов
28. Разрешение лексической многозначности. Алгоритмы классификации. Самонастройка. Методы оценки качества
29. Разрешение лексической многозначности. Методы основанные на словарях и тезаурусах. Варианты алгоритма Леска. Методы оценки качества
30. Семантическая близость слов. Подходы на основе тезаурусов. Методы оценки качества
31. Семантическая близость слов. Подходы на основе статистик. Методы оценки качества
32. Информационный поиск. Ранжирование документов. Векторная модель. Взвешивание терминов. TF-IDF
33. Информационный поиск. Индексирование. Инвертированный индекс. Запросы с джокером. Исправление опечаток.
34. Вопросно-ответные системы. Общая архитектура. Обработка запроса
35. Вопросно-ответные системы. Общая архитектура. Извлечение фрагментов текста
36. Вопросно-ответные системы. Общая архитектура. Обработка ответа
37. Автоматическое реферирование. Общая архитектура
38. Машинный перевод. Классические подходы
39. Статистический машинный перевод. Модель зашумленного канала. Модель перевода на основе фраз. Выравнивание фраз. Декодирование
40. Статистический машинный перевод. Выравнивание слов. Модель IBM Model 1
41. Статистический машинный перевод. Выравнивание слов. Тренировка моделей выравнивания
42. Статистический машинный перевод. Методы оценки качества. BLUE

8.4. Планируемые уровни сформированности компетенций обучающихся и критерии оценивания

Уровни	Содержательное описание уровня	Основные признаки выделения уровня (этапы формирования компетенции, критерии оценки сформированности)	Пятибалльная шкала (академическая) оценка	Двухбалльная шакала, зачет	БРС, % освоения (рейтинговая оценка)
Повышенный	Творческая деятельность	*Включает нижестоящий уровень.* Умение самостоятельно принимать решение, решать проблему/задачу	отлично	зачтено	86-100

		теоретического и прикладного характера на основе изученных методов, приемов, технологий			
Базовый	Применение знаний и умений в более широких контекстах учебной и профессиональной деятельности, нежели по образцу с большей степени самостоятельности и инициативы	*Включает нижестоящий уровень.* Способность собирать, систематизировать, анализировать и грамотно использовать информацию из самостоятельно найденных теоретических источников и иллюстрировать ими теоретические положения или обосновывать практику применения	хорошо		71-85
Удовлетворительный (достаточный)	Репродуктивная деятельность	Изложение в пределах задач курса теоретически и практически контролируемого материала	удовлетворительно		55-70
Недостаточный	Отсутствие признаков удовлетворительного уровня		неудовлетворительно	не зачтено	Менее 55

9. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины.

Основная литература

1. Дадян, Э. Г. Данные: хранение и обработка : учебник / Э. Г. Дадян. — Москва : ИНФРА-М, 2021. — 205 с. — (Высшее образование: Бакалавриат). - ISBN 978-5-16-016447-2. - Текст : электронный. - URL: https://znanium.com/catalog/product/1149101 (дата обращения: 04.04.2023). – Режим доступа: по подписке.

Дополнительная литература

1. Целых, А. Н. Современные методы прикладной информатики в задачах анализа данных : учебное пособие по курсу "Методы интеллектуального анализа данных" / А. Н. Целых, А. А. Целых, Э. М. Котов ; Южный федеральный университет. - Ростов-на-Дону ; Таганрог : Издательство Южного федерального университета,2021. - 130 с. - ISBN 978-5-9275-3783-9. - Текст : электронный. - URL: https://znanium.com/catalog/product/1894428 (дата обращения: 04.04.2023). – Режим доступа: по подписке.

10. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины (модуля).

- НЭБ Национальная электронная библиотека, диссертации и прочие издания
- ЭБС Консультант студента
- ПРОСПЕКТ ЭБС
- ЭБС ZNANIUM.COM
- ЭБС IBOOKS.RU
- Электронно-библиотечная система (ЭБС) Кантитана (https://elib.kantiana.ru/)

11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине.

Программное обеспечение обучения включает в себя:

- система электронного образовательного контента БФУ им. И. Канта – www.lms3.kantiana.ru, обеспечивающую разработку и комплексное использование электронных образовательных ресурсов;
- серверное программное обеспечение, необходимое для функционирования сервера и связи с системой электронного обучения через Интернет;
- корпоративная платформа webinar.ru;
- установленное на рабочих местах обучающихся ПО: Microsoft Windows 7, Microsoft Office Standart 2010, антивирусное программное обеспечение Kaspersky Endpoint Security.
- СУБД PostgreSQL (Свободное ПО, лицензия - Freeware).
- MongoDB (Свободное ПО, лицензия - Freeware).
- Python 2.7.15 (Anaconda2 5.2.0 64-bit)
- Python 3.6.5 (Anaconda3 5.2.0 64-bit)

12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине.

Для проведения занятий лекционного типа, практических и семинарских занятий используются специальные помещения (учебные аудитории), оборудованные техническими средствами обучения – мультимедийной проекционной техникой. Для проведения занятий лекционного типа используются наборы демонстрационного оборудования.

Для проведения лабораторных работ, (практических занятий – при необходимости) используются специальные помещения (учебные аудитории), оснащенные специализированным лабораторным оборудованием: персональными компьютерами с возможностью выхода в интернет и с установленным программным обеспечением, заявленным в п.11.

Для проведения групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации используются специальные помещения (учебные аудитории), оборудованные специализированной мебелью (для обучающихся), меловой / маркерной доской.

Для организации самостоятельной работы обучающимся предоставляются помещения, оснащенные компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду университета.

Для обучения инвалидов и лиц с ограниченными возможностями здоровья университетом могут быть представлены специализированные средства обучения, в том числе технические средства коллективного и индивидуального пользования.