Программа курса «Компьютерная лингвистика»

для образовательной программы «Прикладная математика и информатика»

по направлению подготовки 01.03.02 «Прикладная математика и информатика»

уровень бакалавр

Утверждена
Академическим советом ООП
Протокол № 8.1.2.1-11/03 от «29» июня 2018 г.

Автор	Карпов Н.В.
Число кредитов	3
Контактная работа (час.)	44
Самостоятельная работа (час.)	108
Курс	4
Формат изучения дисциплины	Без использования онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целями освоения дисциплины "Компьютерная лингвистика" являются формирование у студентов четкого представления места и роли современных систем извлечения данных, освоение теоретических основ моделирования и обработки информации на естественном языке, понимание тенденций развития отрасли и направления перспективных исследований, изучение студентами принципов построения современных информационно-поисковых систем. Лекционный курс направлен на подготовку специалистов, способных проводить информационное моделирование предметной области и решать прикладные задачи обработки информации на высоком техническом уровне. Практические занятия служат для получения устойчивых навыков обработки естественного языка с использованием современных высокоуровневых языков программирования в качестве прикладного программиста. Для выполнения заданий используется скриптовый язык Python3, а также технологическая платформа Anaconda4.

В результате освоения дисциплины студент должен:

- Знать сложившуюся в отечественной и зарубежной практике терминологию компьютерной лингвистики, виды корпусных моделей и соответствующее языковое обеспечение, основные типы алгоритмов анализа естественного язык, их архитектуру, функции и принципы использования в системах анализа данных, математические методы, влияющие на принципы разработки лингвистических систем;
- Уметь анализировать сырые данные, методы и алгоритмы, применять соответствующие методы к поставленной задаче;
- Уметь анализировать поставленную проблему и формализовать ее, используя аппарат математических и компьютерных наук;
- Уметь применять полученные знания к решению вопросов проектирования лингвистических систем и систем, имеющих интерфейс на естественном языке;

- Владеть опытом решения технологической задачи, включающей в себя разработку математической модели и оценку ее параметров при помощи самостоятельно разработанной программы.

Настоящая дисциплина относится к вариативной части дисциплин цикла дисциплин профиля подготовки, обеспечивающих подготовку бакалавра. Изучается на 4-м курсе в 3 модуле.

Изучение данной дисциплины базируется на следующих дисциплинах:

- Геометрия и алгебра;
- Теория вероятностей и математическая статистика;
- Анализ и разработка данных;
- Теория и средства трансляции и компиляции;

Для освоения учебной дисциплины, студенты должны владеть следующими знаниями и компетенциями:

- современные методы проектирования и реализации информационных систем;
- основные алгоритмы и структуры данных для быстрого поиска информации;
- программирование на языках C, C++

Основные положения данного курса используются при написании выпускных квалификационных работ.

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Введение в компьютерную лингвистику.

Лекция. Задачи компьютерной лингвистики

Лекция. Предобработка текстов. Токинизация, разбиение на предложения, нормализация, стемминг, лемматизация.

Самостоятельная работа. Собрать коллекцию текстов, включающую две темы (например политика, спорт) и сохранить в виде текстовых документов. Срок выполнения 2 недели.

Основы обработки текста в Python.

Практика . Основы Python: синтаксис и основные типы данных. Работа со строками.

Практика . Чтение собранной коллекции из текстовых файлов. Создание XML (lxml).

Лабораторная работа 1. Разделить текст на предложения и сохранить коллекцию в XML без знаков препинания.

Нормализация текста и обратный индекс.

Лекция . Поиск подстроки в строке. Дистанция редактирования, расстояние Левенштейна.

Лекция . Обратный индекс

Самостоятельная работа 2. Программа построения обратного индекса слов в документах

коллекции. Срок выполнения 2 недели.

Нормализация слов.

Практика . Предобработка, регулярные выражения. Вычисление дистанции редактирования.

Практика . Лемматизация Pymorphy2. Применение обратного индекса для поиска в документах.

Лабораторная работа 2. Программа нахождения документов из коллекции удовлетворяющих запросу при помощи обратного индекса.

Представление текстов в векторном простанстве признаков.

Лекция . Модель мешка слов, частоты слов, стоп слова, TFIDF, тексты в векторном пространстве.

Лекция . Норма вектора и расстояние в метрическом пространстве. Косинусное расстояние.

Самостоятельная работа 3. Поиск в коллекции при помощи TFIDF и косинусного расстояния. Срок выполнения 2 недели.

Парсинг XML структуры, вывод в файл.

Практика. Лемматизация Mystem3. Парсинг XML структуры.

Практика. Разреженные матрицы. Вычисление TF и косинусных расстояний между векторами, сохранение в CSV

Лабораторная работа 3. Программа нахождения в коллекции топ 5 текстов с минимальным косинусным расстоянием от запроса.

Вероятностная модель извлечения информации.

Практика. Вероятностная модель извлечения информации (BIRM).

Практика. Вычисление значений коэффициетов и визуализация функции распределения для релевантных и не релевантных документов коллекции

Лабораторная работа 4. Программа вычисления модели BIRM

Самостоятельная работа 4. Программа применения модели BIRM для оценки релевантности запроса. Срок выполнения 1 неделя.

Анализ качесва работы системы извлечения информации

Практика . Анализ точности классификации. Ошибки первого и второго рода, точность, полнота, ф-мера., ROC, AUC (возможно они это уже знают)

Практика. Вычисление показателей эффективности классификации при помощи sklearn.

Лабораторная работа 5. Программа анализа эффективности BIRM

Самостоятельная работа 5. Тестовая коллекция для анализа эффективности BIRM. Срок выполнения 1 неделя.

Языковая модель

Практика. Марковский процесс и N-граммная языковая модель. Перплексия.

Практика. Вычисление перплексии текста, используя частоты юниграмм НКРЯ

Лабораторная работа 6. Программа вычисления перплексии текста, используя частоты юниграмм НКРЯ

Самостоятельная работа 6. Программа вычисления перплексии текста, используя частоты биграмм НКРЯ. Срок выполнения 2 недели.

Скрытая марковская модель.

Практика. Скрытая марковская модель. OpenCorpora

Практика. Применение ТриТэгера для тегирования

Лабораторная работа 7. Программа применяющая ТриТэгер для тегирования текста

Самостоятельная работа 7. Выполнение 50 заданий в OpenCorpora. Срок выполнения 2 неделя.

Задача тегирования текста

Практика. Задача тегирования и скрытая марковская модель.

Практика. Разбор домашнего задания, работа над ошибками.

III. ОЦЕНИВАНИЕ

Контроль знаний студентов включает формы текущего и итогового контроля.
Текущий контроль осуществляется в течение всего модуля. По курсу предусмотрены самостоятельные работы студентов и работы студентов на практических занятиях. Каждая форма текущего контроля оценивается 10-балльной оценкой, которая выставляется в рабочую ведомость преподавателя. По результатам текущего контроля организуются индивидуальные консультации в рамках второй половины рабочего дня преподавателя.
Самостоятельная работа студентов предполагает выполнение лабораторных работ (заданий к лабораторным работам по темам, указанным в тематическом плане программы).
Форма итогового контроля – письменный экзамен по окончании всего модуля курса, который включает в себя 2 вопроса по материалам курса и оценивается по 10-балльной шкале. Продолжительность экзамена – 45 мин.

Самостоятельная работа:

оценка в 10 баллов проставляется в исключительных случаях самостоятельно проведенной работы, результаты которой могут в дальнейшем использоваться в учебном процессе или в исследовательской работе студента;

оценка в 8-9 баллов проставляется при самостоятельно разработанном или удачно адаптированном и отлично представленном исследовании по выбранной тематике;

оценка в 6-7 баллов проставляется при своевременно выполненном и самостоятельно представленном исследовании по выбранной тематике;

оценка в 4-5 баллов проставляется при частичном, несамостоятельном участии в выполнении работ над заданием;

оценка в 2-3 балла проставляется, когда студент не может самостоятельно представить работу или когда работа носит явные признаки заимствований (работу предлагается переделать);

оценка в 1 балл проставляется при наличии каких-либо демонстративных проявлений безграмотности и неэтичного отношения к работе.

Контрольная работа:

высшая оценка в 9 баллов (10 баллов только в исключительных случаях) проставляется при полностью правильных ответах на вопросы и отличном выполнении заданий (правильном решении задачи, четком и исчерпывающем ее представлении);

почти отличная оценка в 8 баллов проставляется при полностью правильных ответах на вопросы и отличном выполнении заданий, но при отсутствии четкого и исчерпывающего представления решаемой задачи;

оценка в 7 баллов проставляется при правильных ответах на вопросы и правильном решении задачи, но при наличии отдельных неточностей в ответах на вопросы;

оценка в 6 баллов проставляется при наличии отдельных неточностей в ответах на вопросы (включая грамматические ошибки) или неточностях в решении задачи непринципиального характера (описки и случайные ошибки);

оценка в 5 баллов проставляется в случаях, когда в ответах на вопросы и в решении задачи имеются неточности и ошибки, свидетельствующие о недостаточном понимании изучаемой дисциплины и требующие дополнительного обращения к учебным материалам;

оценка в 4 балла проставляется при наличии серьезных ошибок в ответах на вопросы и в решении задачи, что свидетельствует о наличии пробелов в знании изучаемой дисциплины;

оценка в 3 балла проставляется при наличии лишь отдельных положительных моментов в ответах на вопросы и в решении задач, говорящих лишь о потенциальной возможности в последующем более успешного выполнения заданий; оценка в 3 балла, как правило, ведет к повторному решению дополнительной задачи;

оценка в 2 балла проставляется при полном отсутствии положительных моментов в ответах на вопросы и в решении задачи и, как правило, ведет к повторному написанию контрольной работы в целом;

оценка в 1 балл проставляется в тех случаях, когда наряду с неправильными ответами на вопросы и решением задачи имеют место какие-либо демонстративные проявления безграмотности или неэтичное отношение к изучаемой дисциплине.

Высший балл при оценивании видов работ, не допускающих контроля за личным выполнением (домашние расчетные задания), может быть увязан с результатами контрольной работы по текущей теме.

Экзамены (промежуточный и итоговый контроль):

На экзамене, представляющем собой письменные ответы на вопросы и решение задачи с последующим собеседованием, оценка проставляется следующим образом:

высшая оценка в 9 баллов (10 баллов только в исключительных случаях) проставляется при отличном выполнении заданий (полных, с примерами и возможными обобщениями ответах на вопросы, при правильном решении задачи и детальном ее представлении);

почти отличная оценка в 8 баллов проставляется при полностью правильных ответах на вопросы и решении задачи, но при отсутствии примеров и обобщений, а также детального представления решаемой задачи;

оценка в 7 баллов проставляется при правильных ответах на вопросы и правильном решении задачи, но при отсутствии пояснений и обобщений, а также детального представления решаемой задачи;

оценка в 6 баллов проставляется при наличии отдельных неточностей в ответах на вопросы или неточностях в решении задачи непринципиального характера (описки и случайные ошибки);

оценка в 4-5 баллов проставляется в случаях, когда в ответах на вопросы и в решении задачи имеются существенные неточности и ошибки, свидетельствующие о недостаточном понимании изучаемой дисциплины;

оценка в 2-3 балла проставляется при наличии лишь отдельных положительных моментов в ответах на вопросы и в решении задачи;

оценка в 1 балл проставляется в тех случаях, когда наряду с неправильными ответами на вопросы и решением задачи имеют место какие-либо демонстративные проявления безграмотности или неэтичное отношение к изучаемой дисциплине.

По результатам устного собеседования с преподавателем возможны корректировки оценки в ту или иную сторону.

Порядок формирования оценок по дисциплине

Промежуточная оценка $O_{лабораторная}$ рассчитывается как средняя оценка за все лабораторные работы. Накопленная оценка $O_{накопленная}$ расчитывается как средняя оценка за все самостоятельные работы. Результирующая оценка $O_{результ}$ за дисциплину рассчитывается следующим образом:

$$O_{результ} = 0{,}8 \cdot (O_{лабораторная} + O_{накопленная}) + 0{,}2 \cdot O_{экзамен}$$

Полученные после округления этих величин до целого значения ***выставляются в диплом*** как ***результирующие оценки по 10-балльной шкале.***

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Примерные задания *для самостоятельной работы:*

1. Собрать коллекцию текстов, включающую две темы (например политика, спорт) и сохранить в виде текстовых документов
2. Программа построения обратного индекса слов в документах коллекции.
3. Поиск в коллекции при помощи TFIDF и косинусного расстояния.
4. Программа применения модели BIRM для оценки релевантности запроса
5. Тестовая коллекция для анализа эффективности BIRM.
6. Программа вычисления перплексии текста, используя частоты биграмм НКРЯ
7. Выполнение 50 заданий в OpenCorpora

Тематика *контрольной работы*: применение аппарата реляционной алгебры и реляционного исчисления для составления запросов к базе данных клиентов торговой системы

Примерный перечень вопросов к экзамену (итоговый контроль):

1. Основные задачи компьютерной лингвистики
2. Предобработка текстов. Токинизация, разбиение на предложения, нормализация, стемминг, лемматизация.
3. Основы Python: синтаксис и основные типы данных.
4. Работа со строками в Python.
5. Поиск подстроки в строке. Дистанция редактирования, расстояние Левенштейна.
6. Обратный индекс.
7. Поиск при помощи регулярных выражений.
8. Алгоритм вычисление дистанции редактирования.
9. Лемматизация Pymorphy2. Применение обратного индекса для поиска в документах.
10. Модель мешка слов, частоты слов, стоп слова, TFIDF, тексты в векторном пространстве.
11. Норма вектора и расстояние в метрическом пространстве. Косинусное расстояние.
12. Разреженные матрицы. Вычисление TF и косинусных расстояний между векторами, сохранение в CSV
13. Вероятностная модель извлечения информации (BIRM).
14. Анализ точности классификации. Ошибки первого и второго рода, точность, полнота, Ф-мера.
15. Вычисление ROC, AUC.
16. Марковский процесс и N-граммная языковая модель.
17. Перплексия.
18. 3 задачи решаемые скрытой Марковской моделью.
19. Разрешение морфологической неоднозначности при помощи контекста

V. РЕСУРСЫ

5.1 Основная литература

1. Sarkar, Dipanjan. Text Analytics with Python [Электронный ресурс] / Dipanjan Sarkar, БД Springer . Springer Science+Business Media New York 2016. ISBN-13 (pbk): 978-1-4842-2387-1 ISBN-13 (electronic): 978-1-4842-2388-8. Режим доступа: https://proxylibrary.hse.ru:2184/book/10.1007/978-1-4842-2388-8 - загл. с экрана.

5.2 Дополнительная литература

1. Swamynathan, Manohar. Mastering Machine Learning with Python in Six Steps [Электронный ресурс] / Swamynathan, Manohar. БД Спрингер, Springer 2017. ISBN-13 (pbk): 978-1-4842-2865-4 ISBN-13 (electronic): 978-1-4842-2866-1. Режим доступа: https://proxylibrary.hse.ru:2184/book/10.1007/978-1-4842-2866-1 - Загл. с экрана.
2. Луне Педро Коэльо, Вилли Ричарт. Построение систем машинного обучения на языке Python. 2-е издание/ пер. с англ. Слинки А. А. - М.: ДМК Пресс, 2016. - 302 с. : ил. ISBN 978-5-97060-330-7

Дополнительная литература для самостоятельного изучения

1. Jurafsky, Daniel, and James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall. 2009
2. Большакова Е.И. и др. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие — М.: МИЭМ, 2011. — 272 с.
3. Бабкин Э.А., Козырев О.Р., Куркин А.А., Визгунов А.Н. Информационные системы поддержки принятия решений. Нижний Новгород: Н.Новгород: Литера, 2011. 306 с. Доступна электронная версия
4. П. Хоровиц, У. Хилл. Искусство схемотехники: В 2-х т. Пер. с англ. — М: Мир, 1984.
Фролов А., Фролов Г., Синтез и распознавание речи. Современные решения [Электронный ресурс] / Александр Фролов, Григорий Фролов. – Электрон. журн. – 2003. – http://www.frolov-lib.ru
5. Л.В. Бондарко. Звуковой строй современного русского языка. М.: Просвещение, 1997.

5.3 Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Python Software Foundation Python	свободное лицензионное соглашение

5.4 Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
	Профессиональные базы данных, информационно-справочные системы	
1.	NLPub каталог ресурсов для обработки естественного язык	URL: https://nlpub.ru/
2.	Проект создания нового открытого электрон-	URL: https://russianword.net/

	ного тезауруса русского языка. Разрабатывается усилиями представителей УрФУ, ВШЭ, ИММ УрО РАН и Kontur Labs.	
	Интернет-ресурсы (электронные образовательные ресурсы)	
1.	Открытое образование	URL: https://openedu.ru/
2.	Язык XML	URL: http://ru.wikipedia.org/wiki/XML
3.	Официальный сайт проекта GATE Developer	URL: https://gate.ac.uk/

5.5 Материально-техническое обеспечение дисциплины

Лекционные занятия проходят в аудиториях, оборудованных следующим мультимедийным оборудованием: преподавательским компьютером (или ноутбуком), экраном, проектором.

Практические занятия проходят в компьютерных классах, оснащенных преподавательским компьютером, персональными компьютерами, объединенными в локальную сеть с возможностью выхода в интернет.

Дистанционная поддержка дисциплины осуществляется путем использования системы управления версиями Git, веб-сервиса bitbucket.org, а также электронной почты для взаимодействия преподавателя и студентов.