Программа учебной дисциплины «Автоматическая обработка естественного языка**»**

Утверждена
Академическим советом ООП
Протокол № 18 от «23» августа_2019_ г.

Автор	К. А. Дроздова, Д. В. Литвинов, С.Ю. Толдова, Е. И. Мещерякова
Число кредитов	4
Контактная работа (час.)	40
Самостоятельная работа (час.)	112
Курс	4 курс ОП «Фундаментальная и компьютерная лингвистика»
Формат изучения дисциплины	без использования онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целями освоения дисциплины «Автоматическая обработка естественного языка» являются овладение студентами основными методами автоматического анализа текста и знакомство с современными задачами извлечения информации из текста.

В результате освоения дисциплины студент должен:

знать:

- основные типы задач автоматической обработки текста (АОТ), требующих выделения ключевых слов;
- основные теоретические подходы к выделению коллокаций, релевантные в задачах АОТ и основные методы автоматического выделения коллокаций;
- основные подходы к автоматическому разрешению семантической неоднозначности; основные ресурсы и методы использования лексикографических ресурсов при автоматической обработке текста;
- основные классы тональной лексики и основные методы автоматического создания тональных словарей;
- основные модели в тематическом моделировании;
- основные алгоритмы, используемые для построения векторных представлений слов;

уметь:

- уметь применять алгоритмы семантической обработки текста (алгоритмы выделения ключевых слов, выделения коллокаций, разрешения семантической неоднозначности, построения векторов слов);
- уметь реализовывать базовые алгоритмы семантической обработки текста;

владеть:

- владеть инструментами для семантической обработки текстов на русском языке.

Изучение дисциплины «Автоматическая обработка текста» базируется на следующих дисциплинах:

- математика в объеме средней школы;
- «Компьютерные инструменты лингвистического исследования»;
- «Программирование (язык Python)»;
- «Автоматическая обработка естественного языка» в объеме владения морфологической и синтаксической обработки текста.

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

- знать основные методы автоматической обработки текста (на морфологическом и синтаксическом уровне);
- основные понятия теории множеств; теории вероятности и статистики;
- обладать навыками программирования на языке Python.

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин:

- Онтологии и семантические технологии,
- Машинный перевод,
- работа над курсовыми и дипломными работами.

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема (раздел дисциплины)	**Объем в часах**	**Планируемые результаты обучения (ПРО), подлежащие контролю**	**Формы контроля**
Тема 1. Введение. Квантитативные характеристики слов и использование их в автоматической обработке.	лк2 см 2 ср8	владеет основными методами выделения ключевых слов к тексте	домашнее задание по применению изучаемых алгоритмов выделения ключевых слов
Тема 2. Выделение устойчивых словосочетаний.	лк2 см 4 ср12	владеет основными методами выделения устойчивых словосочетаний, знаком с их преимуществами и недостатками	домашнее задание по выделению устойчивых словосочетаний
Тема 3. Методы разрешения семантической неоднозначности.	лк 4 см 4 ср 16		домашнее задание по применению WordNet
Тема 4. Тематическое моделирование.	лк 4 см 4 ср 10	умеет применять методы тематического моделирования	домашнее задание по тематическому моделированию
Тема 5. Определение семантической близости. Векторные модели.	ср 8	умеет семантически близкие слова	домашнее задание по использованию GenSimдля сравнения разных моделей выделения семантически близких слов
Тема 6. Применение методов семантической обработки к задачам извлечения информации из тек-	26	умеет использовать систему Natashадля написания правил по извлечению именованных сущностей;	домашнее задание по извлечению именованных сущностей

ста.		умеет применять методы автоматической классификации для извлечения тонального лексикона на основе большого корпуса текстов	
Проект.	20	использует методы, обсуждаемые в рамках курса, для задач извлечения информации из текста	проект по извлечению информации из текста
Подготовка к тесту.	12		
Часов по видам учебных занятий:	лк18		
	см 22		
	ср 112		
Итого часов:	152		

Тема 1. Квантитативные характеристики слов и использование их в автоматической обработке.

Автоматический семантический анализ. Методы выделения тематически значимых слов в тексте. Ключевые слова. tf.idf, модификации Векторная модель. Вероятностная модель. (модель, основанная на релевантности, OKAPIBM25). Мера LogLikelihood для выделения лексических единиц, специфичных для коллекции текстов по сравнению с другой коллекцией текстов (выделение терминов предметной области). Мера странности (wierdness). Алгоритм RAKE.

Тема 2. Выделение устойчивых словосочетаний

Понятие устойчивых словосочетаний в лингвистике. Разные подходы. Основанное на частотности определение коллокаций. Параметры задачи: понятие окна, типы выделяемых семантических отношений в зависимости от окна.
Частеречные фильтры. Метод среднего и среднеквадратичного отклонения. T-score. T-scoredля разведения двух близких синонимов. Хи-квадрат. LogLikelihood. Поточечная взаимная информация (PMI). Роль синтаксиса при выделении коллокаций.

Тема 3. Методы разрешения семантической неоднозначности

Методы разрешения семантической неоднозначности, основанные на знаниях. Алгоритм Леска. WordNet. Алгоритмы, основанные на использовании лексикографической базы WordNet. Понятие семантического расстояния в WordNet.

Методы автоматической классификации в задачах разрешения семантической неоднозначности (WSD). Наивный байесовский классификатор.

Обучение без учителя при извлечении значений лексемы из неразмеченного корпуса (wordsenseinduction).

Обучение с частичным применением учителя при разрешении семантической неоднозначности. Алгоритм Яровски

Тема 4. Тематическое моделирование

Латентно-семантический анализ. Метод сингулярного разложения матрицы. Латентное размещение Дирихле.

Тема 5. Определение семантической близости. Векторные модели

Дистрибутивная семантика, векторная модель слова. Эмбеддинги: word2vec, GloVe, AdaGram. Обучение моделей word2vec.

Тема 6. Применение методов семантической обработки к задачам извлечения информации из текста

Обзор задач и систем компьютерной лингвистики
Постановка задач для выполнения курсового проекта: разработка ТЗ для создания системы обработки текста (синтеза текста, машинного перевода и т.п.), использующей модули автоматической обработки текста.

III. ОЦЕНИВАНИЕ

Оценка состоит из:

- 30% - выполнение домашних заданий для самостоятельной работы, чтение литературы,

по каждому из разделов выдаются небольшие домашние задания с использованием языка Python и специальных библиотек для реализации обсуждаемых на занятиях алгоритмов;

- 10% - чтение статей из открытых источников по соответствующим темам и подготовку краткого резюме по статье;
- 35 % - проектная работа,

по каждому из разделов готовится один проект; оценивается разработка ТЗ для создания системы обработки текста; взаимное рецензирование и обсуждение проекта; презентации проектов; окончательная версия проекта; рейтинг при оценке качества (F-меры).

- Итоговый экзамен – 25%

В качестве итогового контроля освоения дисциплины предлагается выполнить тест, включающий теоретические вопросы и практические задания.

Оценка по домашним заданиям вычисляется как среднее по всем домашним задания. Возможна досдача и пересдача домашних заданий в конце модуля. Оценка при этом рассчитывается как 0.7 от полной оценки за выполненное задание.

При пересдаче необходимо устранить недостатки проекта. Предполагается переработка проекта и его защита.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Раздел «Автоматическая обработка естественного языка»
Примерные вопросы/ задания для домашних работ

- Вычислите tf.idf для выбранного Вами текста и выбранного корпуса текстов
- Протестируйте систему выделения ключевых слов. Вычислите точность и полноту
- Разметьте в корпусе текстов глаголы по значениям
- Дано множество контекстов некоторого многозначного слова. Сгруппируйте контексты по семантической близости. Разбейте множество контекстов на группы в соответствии с конкретным значением слова.
- Предложите вариант проекта системы, использующей модули автоматической обработки текста. Обоснуйте актуальность и новизну такой системы

- Выполните анализ аналогов системы
- Составьте предварительное описание проекта
- Проведите рецензирование проекта другой группы

Вопросы для оценки качества освоения дисциплины

Какие методы выделения ключевых слов вы знаете?
Какие методы выделения коллокаций Вы знаете; каковы параметры выделения устойчивых словосочетаний
Какие методы разрешения семантической неоднозначности, основанные на базах знаний, Вы знаете?
Какие статистические методы применяются в задачах разрешения семантической неоднозначности
Назовите основные методы кластеризации. Как эти методы применяются к задача кластеризации текстов
Какие методы классификации применяются в задачах рубрикации текстов
Каковы задачи извлечения именованных сущностей? Какие типы омонимии необходимо разрешать в задачах автоматического извлечения сущностей
В чем заключается задача и извлечения фактов и отношений? Какие два базовых подхода используются в решении данной задачи? Приведите примеры систем.
Перечислите задачи извлечения мнений и анализа тональности.
Назовите основные классы лексем и конструкций, которые необходимо учитывать в автоматическом анализе тональности и извлечении мнений.

V. РЕСУРСЫ

5.1 Основная литература

Manning C.D. Introduction to Information Retrieval. / Christopher D. Manning, Prabhakar-Raghavan and HinrichSchütze. – 2008. – CUP. – URL: http://informationretrieval.org

Автоматическая обработка текстов на естественном языке и компьютерная лингвистика / Е.И. Большакова, Э.С.Клышинский, Д.В. Ландэ, А.А.Носков, О.В. Пескова, Е.В. – Ягунова М.: МИЭМ, 2011 г. –URL:http://window.edu.ru/catalog/pdf2txt/465/78465/59324.

5.2 Дополнительная литература

Manning, Ch. D. Foundations of Statistical Natural Language Processing. / Christopher D. Manning, HinrichSchuetze, and Christopher Manning –MIT Press. – 1999. – URL: https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=3339544.

Jurafsky, D., Martin J. H. Speech and Language Processing, 3 издание.– 2018.– URLhttps://web.stanford.edu/~jurafsky/slp3/

Perkins J. Python Text Processing with NLTK 2.0 Cookbook: Over 80 Practical Recipes for Using Python's NLTK Suite of Libraries to Maximize Your Natural Language Processing Capabilities./ Jacob Perkins ed.– Packt Publishing. – 2010. – URL: https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1126730. – ЭБС ProQuest Ebook Central - Academic Complete.

5.3 Программное обеспечение

№ п/п	Наименование	Условия доступа

1.	Microsoft Windows 10 Microsoft Windows 8.1 Professional RUS	*Из внутренней сети университета (договор)*
2.	MicrosoftOfficeProfessionalPlus 2010	*Из внутренней сети университета (договор)*
3.	Python 3	*Свободный*
4.	Ubuntu 18	*Свободный*

5.4

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

№ п/п	Наименование	Условия доступа
	Профессиональные базы данных, информационно-справочные системы	
1.	ЭБС ProQuest Ebook Central - Academic Complete	URL:https://www.proquest.com/libraries/academic/
2.		
	Интернет-ресурсы (электронные образовательные ресурсы)	
1	Единое окно к образовательным ресурсам [Электронный ресурс]	URL: http://window.edu.ru

5.6 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использованиеи демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

– ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);

– мультимедийный проектор с дистанционным управлением.

Учебные аудитории для лабораторных и самостоятельных занятий по дисциплине оснащены, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.