Программа учебной дисциплины «Автоматическая обработка естественного языка»

Утверждена
Академическим советом ОП
Протокол № 15 от «28» июня 2018 г.

Автор	Толдова С.Ю.
Число кредитов	2
Контактная работа (час.)	30
Самостоятельная работа (час.)	46
Курс	3
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 автоматической обработки текста на разных уровнях лингвистического анализа.

В результате освоения дисциплины студент должен:

знать:

- основные задачи компьютерной лингвистики;
- основные формальные модели, лежащие в основе различных модулей автоматической обработки текста;
- необходимые этапы морфологического анализа и проблемы, возникающие при моделировании каждого из этапов;
- основные алгоритмы, используемые для построения автоматического синтаксического анализа;
- наиболее известные доступные для свободного использования компоненты автоматического анализа, в том числе синтаксические и морфологические парсеры;
- принципы оценки качества таких систем;

уметь:

- создавать модули первичной обработки текста;
- строить формальную модель морфологии для создания системы автоматического морфологического анализа;
- проводить оценку качества систем автоматического морфологического, синтаксического и семантического анализа;
- использовать соответствующие модули в различных приложениях;

владеть:

- разработки программ первичной обработки текста;
- использования систем автоматического морфологического анализа;
- тестирования систем морфологического и синтаксического анализа.

Изучение дисциплины «Автоматическая обработка естественного языка» базируется на следующих дисциплинах:

- курс по теория языка программы подготовки бакалавра
- курс по дискретной математики программы подготовки бакалавра
- начальный курс по программированию программы подготовки бакалавра
- английский язык

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

- владеть базовыми представлениями о грамматических категориях и анализе языковых единиц;
- владеть базовыми знаниями в области теории алгоритмов и основ математики;
- владеть базовыми знаниями в области теории вероятностей и статистики;
- уметь читать научные работы и технические описания на английском языке.

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

- машинный перевод, корпусная лингвистика, онтологии и семантические технологии

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Введение в компьютерную лингвистику

Введение в компьютерную лингвистику. Задачи компьютерной лингвистики. Модель информационного поиска. Новостная агрегация и рубрикация. Извлечение информации из текста. Основные типы ресурсов. Основные формальные модели: конечные автоматы, контекстно-свободные грамматики

Свойства естественного языка, создающие сложности для автоматической обработки: омонимия, отсутствие взаимооднозначного соответствия между формой и смыслом. Цепочка обработки: основные этапы обработки. Основные платформы и пакеты для разработки систем АОТ.

Тема 2. Первичная обработка текста.

Графематический анализ. Сегментация текста. Проблемы токенизации: токены; Стоп-слова; обработка специальных символов; обработка слов с дефисом. Типизация токенов. Оффсеты. Сегментация на предложения. Сегментация текста в библиотеке NLTK.

Тема 3. Модель информационного поиска. Векторизация текста

Модель информационного поиска. Модель мешка слов. Индексация. Матрица терм-документ. N-грамы. tf.idf. Оценка качества. Векторная модель документа, векторная модель слова. Поиск похожих текстов. Косинусная мера близости.

Векторизация текстов в библиотеке scikit-learn

Тема 4. Автоматический морфологический анализ.

Введение в автоматический морфологический анализ. Постановка задачи. Основные типы морфологической обработки.

Явления неконкатенативной морфологии. Конечные автоматы и конченые преобразователи. Примеры построения конченых автоматов для морфологического анализа.

Проблемы морфологической неоднозначности. Методы дизамбигуации. Языковые модели. Скрытые марковские модели. Алгоритм Витерби.

Оценка качества частеречного тегера: практикум.

Тема 5. Автоматический синтаксический анализ

Основные модели автоматического синтаксического анализа: непосредственные составляющие, зависимости. Контекстно-свободные грамматики. Унификационные грамматики.

Синтаксический анализ: основные проблемы автоматического анализа (омонимия, типичные случаи синтаксической омонимии, синтаксические нули).

Контекстно-свободные грамматики. базовые алгоритмы (нисходящий алгоритм, алгоритм спуска, алгоритм Кока-Янгера-Касами)

Зависимостные грамматики. Алгоритмы анализа в терминах зависимостей.

Универсальные зависимости (UD): основные стандарты морфологической и синтаксической разметки в терминах UD. Запуск системы синтаксического анализа в терминах UD (UD-pipe).

III. ОЦЕНИВАНИЕ

Оценки по всем формам текущего контроля выставляются по 10-ти балльной шкале.

Накопленная оценка складывается следующим образом

- Домашние задания - 40%

В течение семестра предлагаются небольшие практические задания по анализу структуры дискурса

- Контрольная работа - 30%

В контрольной проверяется владение базовыми понятиями и методами автоматической обработки текста (вопросы с выбором вариантов ответа или с кратким ответом, практические задания, аналогичные заданиям, разбираемым на семинарах)

- Проектное задание - 30%

Предлагается выполнить проект по запуску одной из систем морфологического или синтаксического анализа и провести ее тестирование, либо разработать систему морфологического анализа.

- Итоговый экзамен – 40% от итоговой оценки

В качестве итогового контроля освоения дисциплины предлагается ответить на два теоретических и один практический вопрос.

Накопленная оценка по 10-ти балльной шкале вычисляется по формуле:

$$O_{\text{накопленная}} = 0{,}4 \cdot O_{\text{текущие тесты}} + 0{,}4 \cdot O_{\text{домашнее задание}} + 0{,}2 \cdot O_{\text{контрольная}}$$

Оценка за итоговый контроль в форме экзамена $O_{\text{экзамен}}$ выставляется по 10-балльной шкале. Результирующая оценка вычисляется по следующей формуле:

$$O_{\text{итоговый}} = 0{,}4 \cdot O_{\text{экзамен}} + 0{,}6 \cdot O_{\text{накопленная}},$$

Способ округления накопленной оценки текущего контроля: арифметический.

Тестирование программы предварительного анализа текста и морфологических анализаторов проходит в формате Форума по оценке систем автоматической обработке текста. Командам выдается тестовый и эталонный корпус. Каждая команда проводит оценку точности и полноты, а также функциональное тестирование соответствующей программы.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Оценочные средства для текущего контроля студента/
Примеры домашних заданий

Создайте список наиболее частотных терминов вашего корпуса.
Разбейте текст на токены и предложения. Составьте список необходимых для сегментации классов символов, типов токенов; составьте список сложных случаев, предложите решение.
Задайте три вопроса к прочитанной статье (главы из учебника, посвященные формализмам, используемым в автоматической обработке текста; статьи, посвященные реальным системам и методам автоматического анализа текстов)
Протестируйте систему сегментации текста
Постройте конечный автомат и конечный преобразователь для описания правил морфонологических чередований и построения словоформ на одном из предложенных языков
Тестирование системы морфологического анализа. Проведите морфологическую дизамбигуацию некоторого текста. Какие типы омонимии вам встретились. Оцените качество предсказаний системы.
Постройте контекстно-свободную грамматику для анализа некоторой синтаксической конструкции
Предложите синтаксическую разметку предложения, основанную на деревьях зависимости
Тестирование системы синтаксического анализа в терминах универсальных зависимостей

Мини-тесты по материалам лекций и прочитанной литературе

Пример вопросов мини-теста:

1. Назовите критерии определения вершин в грамматике зависимостей, приведите примеры. В каких случаях в разметке UD, принципы выделения вершины не соответствуют теоретическим принципам. Какие аргументы.
2. Приведите два разных представления данных для морфологического анализа на примере анализа словоформы "городка"
3. На каком допущении относительно вероятности тега в цепочке тегов базируется метод дизамбигуации, основанный на скрытых марковских моделях.
4. Дана словоформа "данные". Определите, в чем проблема ее лемматизации

Оценочные средства для промежуточной аттестации

1. Какие типы лингвистических данных вам известны?
2. Принципы работы морфологических парсеров
3. Конечные автоматы и конечные преобразователи в морфологическом анализе
4. Методы снятия морфологической неоднозначности
5. Основания оценки качества автоматического морфологического разбора
6. Технология shallow parcing
7. Технология chunking
8. Принципы работы синтаксических парсеров
9. Основания оценки качества автоматического синтаксического разбора

Практические вопросы:
1. С помощью информации из НКРЯ рассчитайте, вероятность какой цепочки тегов выше для Мой три окна: (а) A-Pro V N или (б) V Num N (с учетом лексической вероятности)

2. Приведите глубинное, промежуточное и поверхностное представление для словоформ татарского языка (исходя из принципа двухуровневой морфологии: символу алфавита на одном уровне соответствует только один символ алфавита на другом уровне, грамматический тег – один символ):

bala-lar-ɤbɤz-ga – нашим детям

täräz-lär-ebez-gä – нашим окнам

3. Даны четыре предложения. Постройте для них деревья НС. Извлеките из полученного корпуса грамматику. Переведите ее в нормальную форму Хомского.

Распишите применение алгоритма Кока-Янгера-Касами для разбора предложения

Такие типы стали есть в цехе.

Если Вам не хватает правил построенной Вами грамматики для разбора предложения, допишите необходимые правила.

V. РЕСУРСЫ

5.1 Основная литература

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

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

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.	Microsoft Office Professional Plus 2010	*Из внутренней сети университета (договор)*
3.	Python 3	*Свободный*
4.	Ubuntu 18	*Свободный*

5.	NLTK	*Свободный*

5.4

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

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

5.6 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и 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 семинарских и самостоятельных занятий по дисциплине не требуют специального технического оснащения