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

Утверждена
Академическим советом ООП
Протокол № 18 от «23»_августа____2019__ г.

| | |
|---|---|
| Автор | Толдова С.Ю. |
| Число кредитов | 2 |
| Контактная работа (час.) | 30 |
| Самостоятельная работа (час.) | 84 |
| Курс | 3 курс, ОП бакалавриата «Фундаментальная и компьютерна лингвистика» |
| Формат изучения дисциплины | без использования онлайн курса |

## 1. 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 теории 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;
– владеть базовыми навыками программирования на языке Python.

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

– автоматическая обработка естественного языка: семантика, анализ контента;

а также в исследованиях при написании курсовых работ.

## 2. Содержание учебной дисциплины

| **Тема (раздел дисциплины)** | **Объем в часах** | **Планируемые результаты обучения (ПРО), подлежащие контролю** | **Формы контроля** |
|---|---|---|---|
| Тема 1 **Введение в компьютерную лингвистику**. | лк 2<br>см 4<br>ср 8 | знает основные задачи АОТ | |
| Тема 2. **Первичная обработка текста**. **Модель информационного поиска. Векторизация текста** | лк 4<br>см 4<br>ср 12 | осуществляет первичную обработку текста, разбиение на предложения, распознавание языка | домашнее задание: автоматическое определение языка |
| Тема 3. **Автоматический морфологический анализ**. | лк 4<br>см 4<br>ср 20 | осуществляет морфологическую аннотацию текста, определяет качество морфологического таггера; например, строит конечный автомат для анализа одного из морфонологических явлений в одном из малоресурсных языках | домашнее задание: классификация ошибок таггера |
| Тема 4. **Автоматический синтаксический анализ**. | лк 4<br>см 4<br>ср 20 | запускает синтаксический анализатор; | домашнее задание: синтаксический анализ текста с помощью алгоритма Кока-Янгера-Касами |
| **Тема 5. Проект по АОЕЯ** | ср 24 | | |

| | |
|---|---|
| **Часов по видам учебных занятий:** | лк 14 |
| | см 16 |
| | ср 84 |
| **Итого часов:** | 114 |

**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.

**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 3. 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 тагера: практикум.

**Тема 4. 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).

## 3. Оценивание

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

Оценка складывается следующим образом

- Домашние задания - 40%

В течение семестра предлагаются небольшие практические задания

- 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 – 30%

В качестве итогового контроля освоения дисциплины предлагается тест.

$$О_{итоговая} = 0,4 * О_{ДЗ} + 0,3 * О_{проектное\ задание} + 0,3 * О_{экзамен}$$

Домашнее задание задаётся на неделю или две недели. В случае сдачи работы после дедлайна оценка снижается: а) на балл при опоздании больше, чем на 3 дня, но не меньше, чем не 2 недели и б) на 3 балла при опоздании больше, чем на 2. Пересдача домашних заданий разрешается только при неудовлетворительной оценке и не позднее, чем за неделю до экзамена.

Способ округления накопленной оценки текущего контроля: арифметический.

## 4. 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

Практические вопросы итогового теста:
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-ɣbɤz-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.

## 5. 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<br>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

## 6. Особенности организации обучения для лиц с ограниченными возможностями здоровья и инвалидов

В случае необходимости, обучающимся из числа лиц с ограниченными возможностями здоровья (по заявлению обучающегося) а для инвалидов также в соответствии с индивидуальной программой реабилитации инвалида, могут предлагаться следующие варианты восприятия учебной информации с учетом их индивидуальных психофизических особенностей, в том числе с применением электронного обучения и дистанционных технологий:

*6.1.1. для лиц с нарушениями зрения:* в печатной форме увеличенным шрифтом; в форме электронного документа; в форме аудиофайла (перевод учебных материалов в аудиоформат); в печатной форме на языке Брайля; индивидуальные консультации с привлечением тифлосурдопереводчика; индивидуальные задания и консультации.

*6.1.2. для лиц с нарушениями слуха:* в печатной форме; в форме электронного документа; видеоматериалы с субтитрами; индивидуальные консультации с привлечением сурдопереводчика; индивидуальные задания и консультации.

*6.1.3. для лиц с нарушениями опорно-двигательного аппарата:* в печатной форме; в форме электронного документа; в форме аудиофайла; индивидуальные задания и консультации.