

**Правительство Российской Федерации**

**Федеральное государственное автономное образовательное учреждение  
высшего профессионального образования  
«Национальный исследовательский университет  
«Высшая школа экономики»**

Факультет Компьютерных наук  
Департамент больших данных и информационного поиска  
Базовая кафедра Яндекс

УТВЕРЖДАЮ  
Академический руководитель  
образовательной программы  
«Науки о данных»  
по направлению 01.04.02  
«Прикладная математика и информатика»  
С.О. Кузнецов

\_\_\_\_\_ 2015 г.

**Программа дисциплины «Автоматическая обработка текстов»**  
для направления 01.04.02 "Прикладная математика и информатика" подготовки магистра  
для магистерской программы "Науки о данных"

**Автор программы:**  
Зобнин А.И. ( alzobnin@yandex-team.ru )

Одобрена на заседании базовой кафедры Яндекс  
2015 г.

«\_\_» \_\_\_\_\_

Заведующий кафедрой

\_\_\_\_\_ М.А. Бабенко

Рекомендована Академическим советом образовательной программы  
«Науки о данных»

«\_\_» \_\_\_\_\_ 2015 г.

Менеджер базовой кафедры Яндекс

\_\_\_\_\_ И.И. Алескерова

Москва, 2015

*Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения подразделения разработчика программы.*



## Пояснительная записка

### Автор программы

К. ф.-м. н., доцент базовой кафедры Яндекса Зобнин Алексей Игоревич.

### Аннотация

Дисциплина «Автоматическая обработка текстов» предназначена для подготовки магистров 01.04.02 – Прикладная математика и информатика.

Курс посвящен прикладным задачам, связанным с обработкой текста на естественном языке. Студенты познакомятся с задачами на разных уровнях: от работы с отдельными символами (кодировки, определение языка) они перейдут к уровню слов (морфология, снятие морфологической неоднозначности, выделение именованных сущностей), синтаксису и семантике. Отдельно будут рассмотрены модели дистрибутивной семантики (представления слов в виде векторов в подходящем пространстве). На семинарах планируется знакомство с инструментами для обработки текста - специальными модулями для Python (NLTK, gensim и т. д.), трансдюсерами (OpenFST), морфологическими и синтаксическими парсерами (mystem, tomita), моделью word2vec и т. д.

Программа курса предусматривает лекции (30 часов) и практические занятия (30 часов).

### Учебные задачи курса

Целью данного курса является ознакомление слушателей с методами обработки текста на естественном языке. Предполагается знакомство с методами извлечения отношений, анализа тональности, аннотирования и кластеризации текстов, а также с существующими программными реализациями этих методов. Все темы курса снабжены практическими заданиями, призванными дать возможность сопоставить теорию с практикой.

### Компетенции, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

- знать основные уровни анализа текста — символичный, морфологический, синтаксический, семантический;
- владеть основными математическими моделями анализа текста: языковыми моделями на основе n-грамм, скрытыми марковскими моделями, марковскими моделями максимальной энтропии, синтаксическими моделями, моделями векторного пространства и т. д.
- уметь создавать программы с помощью языка Python и специализированных пакетов, решающие задачи анализа текста с помощью применения языковых моделей.

В результате освоения дисциплины студент осваивает следующие компетенции:

Компетенция	Код по ФГОС/ НИУ	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
Способность строить и решать математические модели в соответствии с направлением подготовки и специализацией	ИК-М7.2пми	Знание основных математических моделей языка	Лекции, практические занятия, домашние задания
Способность применять в исследовательской и прикладной деятельности современные языки программирования и языки манипулирования данными, операционные системы, пакеты программ и т.д.	ИК-М7.5пми	Владение языком программирования Python, способность написать реализацию языковой модели для решения задач компьютерной лингвистики	Лекции, практические занятия, домашние задания
Способность публично представлять результаты профессиональной деятельности (в том числе с использованием информационных технологий)	ИК-М2.5пми	Способность переработать и изложить в виде доклада материалы научной статьи по компьютерной лингвистике	Лекции, практические занятия, домашние задания

### Место дисциплины в структуре образовательной программы

Для освоения дисциплины предполагаются базовые знания по таким разделам математики и информатики, как «Теория вероятностей и математическая статистика», «Информатика и программирование», «Алгоритмы и структуры данных» – соответствующие дисциплины входят в программу обучения бакалавра по направлению 01.03.02 «Прикладная математика и информатика».

### Тематический план дисциплины «Автоматическая обработка текстов»

№	Название темы	Всего часов по дисциплине	Аудиторные часы		Самостоятельная работа
			Лекции	Сем. и практика	
1	Введение в автоматическую обработку текстов	14	2	2	10
2	Стандарт Unicode	14	2	2	10
3	Языковые модели	12	2	2	8
4	Регулярные языки	12	2	2	8

5	Морфология	12	2	2	8
6	Скрытые марковские модели	12	2	2	8
7	Марковские модели максимальной энтропии	12	2	2	8
8	Контекстно-свободные грамматики	12	2	2	8
9	Синтаксические деревья зависимостей	12	2	2	8
10	Извлечение коллокаций	14	2	2	10
11	Дистрибутивная семантика	12	2	2	8
12	Формальная семантика	12	2	2	8
13	Извлечение коллокаций	14	2	2	10
14	Анализ тональности текста	12	2	2	8
15	Распознавание и синтез речи	14	2	2	10
	Итого	190	30	30	130

### Формы контроля и структура итоговой оценки

Текущий контроль — две (теоретические) домашние работы в первом и втором модулях соответственно, две (практические) контрольные работы в первом и втором модулях соответственно (по 80 мин.).

Итоговый контроль – устный экзамен (120 мин.), включающий вопросы и задачи по пройденным темам.

Итоговая оценка вычисляется следующим образом:

$$0,1 * \partial з1 + 0,1 * \partial з2 + 0,2 * кр1 + 0,2 * кр2 + 0,4 * экзамен.$$

### Таблица соответствия оценок по десятибалльной и системе зачет/незачет

Оценка по 10-балльной шкале	Оценка по 5-балльной шкале
1	Незачет
2	
3	
4	Зачет
5	
6	
7	
8	
9	
10	

**Таблица соответствия оценок по десятибалльной и пятибалльной системе**

<b>По десятибалльной шкале</b>	<b>По пятибалльной системе</b>
1 – неудовлетворительно 2 – очень плохо 3 – плохо	неудовлетворительно – 2
4 – удовлетворительно 5 – весьма удовлетворительно	удовлетворительно – 3
6 – хорошо 7 – очень хорошо	хорошо – 4
8 – почти отлично 9 – отлично 10 – блестяще	отлично – 5

## **Программа дисциплины «Автоматическая обработка текстов»**

### **1. Введение в автоматическую обработку текстов.**

Задачи, возникающие при обработке текстов на естественном языке. Обзор применяемых методов.

### **2. Стандарт Unicode.**

Однобайтные национальные кодировки. Принципы Unicode. Кодировки UTF-8, UTF-16 и UTF-32. Токенизация и нормализация.

### **3. Языковые модели.**

Символьные и словные языковые модели на основе n-грамм. Сглаживания языковых моделей. Задача распознавания языка. Прюнинг и сжатие языковых моделей.

### **4. Регулярные языки.**

Регулярные языки, регулярные выражения и конечные автоматы. Теорема Клини. Лемма о накачке.

### **5. Морфология.**

Компьютерная морфология. Грамматический словарь А. А. Зализняка. Лемма и парадигма. Лемматизация и стемминг. Конечные трансдюсеры как модель для морфологических преобразований. Программа Mystem.

### **6. Скрытые марковские модели.**

Определение скрытых марковских моделей. Прямые и обратные вероятности. Алгоритм Витерби. Обучение скрытых марковских моделей с помощью алгоритма Баума-Велша. Задача снятия неоднозначностей при определении частей речи.

## **7. Марковские модели максимальной энтропии.**

Генеративные и дискриминативные модели. Логистическая регрессия. Принцип максимальной энтропии. Модификация алгоритма Витерби для МЕММ. Задача распознавания именованных сущностей.

## **8. Контекстно-свободные грамматики.**

Иерархия формальных грамматик Хомского. Нисходящие и восходящие синтаксические парсеры. Алгоритмы Earley и Cocke–Younger–Kasami. Томита-парсер. Вероятностные контекстно-свободные грамматики.

## **9. Синтаксические деревья зависимостей.**

Деревья зависимостей и деревья составляющих. Data-driven-подходы к разметке синтаксиса. Технология MaltParser.

## **10. Извлечение коллокаций.**

Биграммы. Распределение расстояний между словами. Взаимная информация. Отношения правдоподобий. Алгоритмы извлечения ключевых слов.

## **11. Дистрибутивная семантика.**

Модель векторных пространств. Латентно-семантический анализ. Тематическое моделирование. Глубокое обучение. Модель word2vec и ее применения. Языковые модели на основе рекуррентных нейронных сетей.

## **12. Формальная семантика.**

Логика первого порядка. Лямбда-исчисление. Предикаты в вершинах дерева синтаксического разбора.

## **13. Извлечение отношений.**

Отношения в тексте между именованными сущностями. Поиски совпадений по образцу. Подходы, основанные на машинном обучении. Bootstrapping.

## **14. Анализ тональности текста.**

Подходящие свойства текста. Наивный байесовский подход. Алгоритм Turney. Составление лексикона тональностей.

## **15. Распознавание и синтез речи.**

Модель зашумленного канала. Акустическая модель. Декодирование. Технология Яндекс SpeechKit.

## **Список литературы**

### **Основная литература**

1. D. Jurafsky, J. Martin. Speech and Language Processing. 2nd edition. Prentice-Hall. 2008.
2. C. Manning, P. Raghavan, H. Schütze. Introduction to Information Retrieval. Cambridge University Press. 2008. Перевод: Введение в информационный поиск. Вильямс, 2011.
3. C. Manning, H. Schütze. Foundations of Statistical Natural Language Processing. MIT Press. Cambridge, MA: May 1999.

### **Образовательные технологии**

1. Лекции в виде презентаций выкладываются на вики-страницу курса для самостоятельной работы студентов;
2. Материалы семинаров и разбор задач семинаров оформляются в формате IPython notebook и также выкладываются на вики-странице.
3. Практические (лабораторные) работы студентов предусматривают самостоятельную работу по решению конкретных задач компьютерной лингвистики с помощью изученных моделей.
4. Теоретическое домашнее задание закрепляет знания студентов, полученные на лекции.

### **Оценочные средства для текущего и итогового контроля**

#### **Примеры теоретических заданий:**

1. Составить регулярное выражение, удовлетворяющее заданным требованиям.
2. Доказать, что заданный формальный язык является контекстно-свободным, но не является регулярным.
3. Построить наиболее вероятную цепочку тегов (скрытых состояний) в заданной скрытой марковской модели по указанному предложению.
4. Восстановить цепочку действий алгоритма MaltParser для получения заданного синтаксического дерева зависимостей.
5. Вывести формулу для коэффициентов заданного алгоритма сглаживания n-граммной языковой модели.

#### **Примеры практических заданий:**

1. Построить символьную триграммную языковую модель по википедии и с ее помощью построить распознаватель языка документа.
2. Вычислить перплексию n-граммной языковой модели с заданным сглаживанием.

3. На основе заданной обучающей выборки построить марковскую модель максимальной энтропии для выделения заданных именованных сущностей (имен собственных, географических названий и т. д.) из текста.

4. Построить морфологический трансдюсер для заданного языка по данному лексикону и указанному набору грамматических правил.

5. Применить модель word2vec для автоматического поиска морфологически близких слов.

**Примеры вопросов на экзамене:**

1. В чем особенности кодировок UTF-8, UTF-16 и UTF-32?

2. Что такое конечный трансдюсер? Всегда ли недетерминированный трансдюсер эквивалентен детерминированному?

3. Какие методы сглаживания языковых моделей Вы знаете?

4. В чем отличие скрытых марковских моделей от марковских моделей максимальной энтропии?

5. Какие существуют методы векторного представления слов?

Автор программы:

/ <Зобнин А. И.> /