**Федеральное государственное учреждение**
**«Федеральный исследовательский центр «Информатика и управление»**
**Российской академии наук»**
**(ФИЦ ИУ РАН)**

**Утверждена**
Ученым советом ФИЦ ИУ РАН,
протокол № 1 от «27» ноября 2015 г.
Председатель Ученого совета,
директор ФИЦ ИУ РАН
_____И.А. Соколов
«30» ноября 2015 г.

# РАБОЧАЯ ПРОГРАММА

УЧЕБНОЙ ДИСЦИПЛИНЫ
«Основы обработки текстовой информации»

**Направление подготовки**
09.06.01 Информатика и вычислительная техника

**Профиль (направленность программы)**
05.13.11 Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей;
05.13.18 Математическое моделирование, численные методы и комплексы программ

**Квалификация выпускника**
Исследователь. Преподаватель-исследователь

**Форма обучения**
очная

Москва, 2015

Направление подготовки: 09.06.01 Информатика и вычислительная техника

Профиль (направленность программы): 05.13.01 Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей;
05.13.18 Математическое моделирование, численные методы и комплексы программ

Дисциплина: «Основы обработки текстовой информации»

Форма обучения: очная

Рабочая программа составлена с учетом ФГОС ВО по направлению подготовки 09.06.01 Информатика и вычислительная техника, утвержденного приказом Министерства образования и науки Российской Федерации от 30 июля 2014 года № 875, зарегистрировано в Минюсте Российской Федерации 20 августа 2014 года № 33685.

РАБОЧАЯ ПРОГРАММА РЕКОМЕНДОВАНА

отделом Систем математического обеспечения ФИЦ ИУ РАН

Руководитель отдела _____ / Серебряков В.А. /

«___»_____ 2015г.

ИСПОЛНИТЕЛИ (разработчики программы):

Серебряков В.А., ФИЦ ИУ РАН, зав. отделом Систем математического обеспечения ФИЦ ИУ РАН, д.ф.-м.н., профессор.

Рабочая программа зарегистрирована в аспирантуре под учетным номером _____ на правах учебно-методического издания.

Начальник отдела докторантуры и аспирантуры_____ / Клименко С..И. /

_____

Оглавление

## АННОТАЦИЯ

Дисциплина «Основы обработки текстовой информации» реализуется в рамках Блока 1 дисциплина по выбору Основной профессиональной образовательной программы высшего образования - программам подготовки научно-педагогических кадров в аспирантуре Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН) по направлению подготовки 09.06.01 Информатика и вычислительная техника, профиль (направленность программы) 05.13.11 «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» и 05.13.18 «Математическое моделирование, численные методы и комплексы программ» аспирантам очной формы обучения.

Рабочая программа разработана с учетом требований ФГОС ВО по направлению подготовки 09.06.01 Информатика и вычислительная техника, утвержденного приказом Министерства образования и науки Российской Федерации от 30 июля 2014 года № 875, зарегистрировано в Минюсте Российской Федерации 20 августа 2014 года № 33685.

Основным источником материалов для формирования содержания программы являются: материалы конференций, симпозиумов, семинаров, Интернет-ресурсы, научные издания и монографические исследования и публикации.

Общая трудоемкость дисциплины по учебному плану составляет - 2 зач.ед. (72 часов), из них лекций - 36 час., семинарских занятий – 0 час., практических занятий – 0 час. и часов самостоятельной работы – 36 час. Дисциплина реализуется на 3 курсе, 5 семестре, продолжительность обучения – 1 семестр.

Текущая аттестация проводится не менее 2 раз в соответствии с заданиями и формами контроля, предусмотренные настоящей программой.

Промежуточная оценка знания осуществляется в период зачетно-экзаменационной сессии в форме: зачета.

## 1. ЦЕЛИ И ЗАДАЧИ

Цель курса - освоение аспирантами фундаментальных знаний в области обработки и анализа текстовой информации, а также изучение основных проблем компьютерной обработки текстов и современных подходов к их решению.

Задачами данного курса являются:

- формирование базовых знаний в области компьютерной обработки текстовой информации как дисциплины, обеспечивающей технологические основы современных инновационных сфер деятельности;
- обучение аспирантов принципам решения задач обработки естественного языка на основе методов машинного обучения;
- формирование подходов к выполнению аспирантами исследований в области обработки естественного языка.

## 2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Процесс изучения дисциплины «Основы обработки текстовой информации» направлен на формирование компетенций или отдельных их элементов в соответствии с ФГОС ВО по направлению подготовки 09.06.01 Информатика и вычислительная техника, профиль (направленность программы) 05.13.11 «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» и 05.13.18 «Математическое

моделирование, численные методы и комплексы программ» аспирантам очной формы обучения:

а) универсальных (УК)

- способность к критическому анализу и оценке современных научных достижений, генерированию новых идей при решении исследовательских и практических задач, в том числе в междисциплинарных областях (УК-1);
- готовность участвовать в работе российских и международных исследовательских коллективов по решению научных и научно-образовательных задач (УК-3);
-способность следовать этическим нормам в профессиональной деятельности (УК-5);
- способность планировать и решать задачи собственного профессионального и личностного развития (УК-6).

б) общепрофессиональных (ОПК):

-владение методологией теоретических и экспериментальных исследований в области профессиональной деятельности (ОПК-1);
-владение культурой научного исследования, в том числе с использованием современных информационно-коммуникационных технологий (ОПК-2);
-способность к разработке новых методов исследования и их применению в самостоятельной научно-исследовательской деятельности в области профессиональной деятельности (ОПК-3);
-готовность организовать работу исследовательского коллектива в области профессиональной деятельности (ОПК-4);
-способность представлять полученные результаты научно-исследовательской деятельности на высоком уровне и с учетом соблюдения авторских прав (ОПК-6);
-владение методами проведения патентных исследований, лицензирования и защиты авторских прав при создании инновационных продуктов в области профессиональной деятельности (ОПК-7);
-готовность к преподавательской деятельности по основным образовательным программам высшего образования (ОПК-8).

в) профессиональных (ПК):

- готовность использовать знание основных методов системного программирования в последующей профессиональной деятельности в качестве научных сотрудников, преподавателей вузов, инженеров, технологов (ПК-1);
- готовность выявить естественнонаучную сущность проблем, возникающих в ходе профессиональной деятельности в области моделирования и анализа сложных естественных и искусственных систем (ПК-3);
-способность к созданию математических и информационных моделей исследуемых процессов, явлений и объектов, относящихся к профессиональной сфере (ПК-4);
-способность применять на практике умения и навыки в организации исследовательских работ и проводить научные исследования, готовность к участию в инновационной деятельности (ПК-5).

В результате освоения дисциплины «Основы обработки текстовой информации» обучающийся должен:

**Знать:**

- место и роль общих вопросов науки в научных исследованиях;
- современные проблемы математики, физики и экономики;
- теоретические модели рассуждений, поведения, обучения в когнитивных науках;
- новейшие открытия в области когнитивных наук;
- постановку проблем математического и информационного моделирования сложных систем;
- взаимосвязь и фундаментальное единство естественных наук.

**Уметь:**

– эффективно использовать на практике теоретические компоненты науки: понятия, суждения, умозаключения, законы;

– представить панораму универсальных методов и законов современного естествознания;

– работать на современной электронно-вычислительной технике;

– абстрагироваться от несущественных факторов при моделировании реальных природных и общественных явлений;

– планировать процесс моделирования и вычислительного эксперимента.

**Владеть:**

– научной картиной мира;

– методами постановки задач и обработки результатов компьютерного моделирования;

– навыками самостоятельной работы в лаборатории на современной вычислительной технике;

– методами математического моделирования поведения, рассуждений и обучения.

## 3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

### 3.1. Структура дисциплины

**Распределение трудоемкости дисциплины по видам учебных работ**

| Вид учебной работы | Трудоемкость | | | | | |
|---|---|---|---|---|---|---|
| | общая | | Из них | | | |
| | Зач. Ед. | Час. | Лекц. | Прак. | Сем. | Сам.р. |
| **ОБЩАЯ ТРУДОЕМКОСТЬ** по Учебному плану | 2 | 72 | 36 | | | 36 |
| ***Аудиторные занятия*** | | | | | | |
| Лекции (Л) | 1 | 36 | 36 | | | |
| Практические занятия (ПЗ) | | | | | | |
| Семинары (С) | | | | | | |
| ***Самостоятельная работа (СР) без учёта промежуточного контроля:*** | | | | | | |
| Самоподготовка (проработка и повторение лекционного материала и материала учебников и учебных пособий, подготовка к семинарским и практическим занятиям) и самостоятельное изучение тем дисциплины | 1 | 36 | | | | 36 |
| ***Вид контроля:*** зачет (является составной частью кандидатского экзамена) | | | | | | |

### 3.2. Содержание разделов дисциплины

**Общее содержание дисциплины**

| | Разделы и темы | Содержание | Объем (зачетные единицы - часы) | | Общее количество часов |
|---|---|---|---|---|---|
| | | | Аудиторная работа | Самостоятельная работа | |

| | | | | | |
|---|---|---|---|---|---|
| 1 | Задачи обработки текста. | Многозначность при обработке текста. Проблема понимания | 2 | 2 | 4 |
| 2 | Регулярные выражения и конечные автоматы (КА) | Регулярные выражения, КА, распознавание языка с помощью КА. Регулярные языки и конечные автоматы. Построение КА для регулярных выражений | 2 | 2 | 4 |
| 3 | Методы поиска словосочетаний | Проверка статистических гипотез для поиска словосочетаний, t-тест, критерий хи-квадрат, отношение правдоподобия,информационнотеоретический подход к поиску словосочетаний | 4 | 4 | 8 |
| 4 | Методы поиска словосочетаний | Проверка статистических гипотез для поиска словосочетаний, t-тест, критерий хи-квадрат, отношение правдоподобия,информационнотеоретический подход к поиску словосочетаний | 4 | 4 | 8 |
| 5 | Методы обучения с учителем и задачи обработки текстов | Использование скрытой марковской модели для определения частей речи. Скрытые марковские модели, Вероятность последовательности, Прямой алгоритм. Алгоритм Витерби, Наивный байесовский классификатор, Логистическая регрессия. Модель максимальной энтропии, Марковская модель максимальной энтропии. | 4 | 4 | 8 |
| 6 | Контексто-свободные граммитики и синтаксический анализ | Типы грамматик. Грамматика составляющих. Грамматика зависимостей. Категориальная грамматика. Контекстно- свободные грамматики. КС грамматики и регулярные языки. Банк деревьев. Синтаксический разбор. Разбор сверху вниз и снизу вверх. Алгоритм Кока-Янгера-Касами. Эквивалентность КС грамматик. Фрагментирование | 4 | 4 | 8 |

| | | | | | |
|---|---|---|---|---|---|
| 7 | Статистические методы синтаксического анализа | Стохастические КС грамматики. Разрешение синтаксической многозначности. Моделирование языка. Обучение стохастических КС грамматик. Вероятностная версия алгоритма Кока-Янгера- Касами. Оценка качества. Проблемы стохастический КС грамматик. Алгоритм Коллинза. Оценка качества | 4 | 4 | 8 |
| 8 | Лексическая семантика. | Лексическая семантика. WordNet. Значения слов. Разрешение лексической многозначности. Алгоритмы классификации. Самонастройка. Методы оценки качества. Методы основанные на словарях и тезаурусах. Варианты алгоритма Леска. Методы оценки качества. Семантическая близость слов. Подходы на основе тезаурусов. Подходы на основе статистик | 4 | 4 | 8 |
| 9 | Вопросно-ответные системы и автоматическое реферирование | Вопросно-ответные системы. Общая архитектура. Обработка запроса. Извлечение фрагментов текста. Автоматическое реферирование. Общая архитектура | 4 | 4 | 8 |
| 10 | Машинный перевод | Машинный перевод. Классические подходы. Статистический машинный перевод. Модель зашумленного канала. Модель перевода на основе фраз. Выравнивание фраз (если слова выровнены). Декодирование. Выравнивание слов.Модель IBM Model 1. Тренировка моделей выравнивания. Методы оценки качества.BLUE. | 4 | 4 | 8 |
| ВСЕГО | | | 36 | 36 | 72 |

**Перечень разделов дисциплины и распределение времени по темам**

| № темы и название | Количество часов |
|---|---|
| 1. Задачи обработки текстов | 6 |

| | |
|---|---|
| 2. Регулярные выражения и конечные автоматы | 6 |
| 3. Методы поиска словосочетаний | 6 |
| 4. Языковые модели и задача определения частей речи | 6 |
| 5. Методы обучения с учителем и задачи обработки текстов | 8 |
| 6. Контекстно-свободные грамматики и синтаксический анализ | 8 |
| 7. Статистические методы синтаксического анализа | 8 |
| 8. Лексическая семантика | 8 |
| 9. Вопросно-ответные системы и автоматическое реферирование | 8 |
| 10. Машинный перевод | 8 |
| ВСЕГО( зач. ед.(часов)) | 72 часов |

**Лекции**

| № темы и название | Количество часов |
|---|---|
| 1. Задачи обработки текстов | 2 |
| 2. Регулярные выражения и конечные автоматы | 2 |
| 3. Методы поиска словосочетаний | 2 |
| 4. Языковые модели и задача определения частей речи | 2 |
| 5. Методы обучения с учителем и задачи обработки текстов | 4 |
| 6. Контекстно-свободные грамматики и синтаксический анализ | 4 |
| 7. Статистические методы синтаксического анализа | 4 |
| 8. Лексическая семантика | 4 |
| 9. Вопросно-ответные системы и автоматическое реферирование | 4 |
| 10. Машинный перевод | 4 |
| ВСЕГО( зач. ед.(часов)) | 36 часа |

### 3.3. Семинарские занятия

Не предусмотрены

### 3.4. Практические занятия

Не предусмотрены

### 3.5. Самостоятельная работа аспирантов

Внеаудиторная самостоятельная работа аспирантов включает следующие виды деятельности:

- конспектирование и реферирование первоисточников и другой научной и учебной литературы;
- проработку учебного материала (по конспектам, учебной и научной литературе);
- изучение учебного материала, перенесенного с аудиторных занятий на самостоятельную проработку;
- написание рефератов;
- выполнение переводов научных текстов с иностранных языков;
- индивидуальные домашние задания расчетного, исследовательского и т.п. характера

**Содержание и объем самостоятельной работы аспирантов**

| | Темы | Трудоёмкость в зач. ед.(количество часов) |
|---|---|---|
| 1 | Проработка и повторение лекционного материала и материала рекомендованной литературы – выполняется самостоятельно каждым аспирантом по итогам каждой из лекций, результаты | 10 |

| | | |
|---|---|---|
| | контролируются преподавателем на лекционных занятиях, используются конспект лекций, учебники, рекомендуемые данной программой | |
| 2 | Самостоятельное изучение отдельных подразделовпрограммы – выполняется каждым аспирантом по заданию преподавателя, результаты контролируются преподавателем на лекционных занятиях, используются материалы, рекомендуемые данной программой | 26 |
| | ВСЕГО (часов) | 36 |

## 4. ТЕКУЩАЯ И ПРОМЕЖУТОЧНАЯ АТТЕСТАЦИЯ. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

**Текущая аттестация аспирантов**. Текущая аттестация аспирантов проводится в соответствии с локальным актом ФИЦ ИУ РАН - Положением о текущей, промежуточной и итоговой аттестации аспирантов ФИЦ ИУ РАН по программам высшего образования – программам подготовки научно-педагогических кадров в аспирантуре и является обязательной.

Текущая аттестация по дисциплине проводится в форме опроса, а также оценки вопроса-ответа в рамках участия обучающихся в дискуссиях и различных контрольных мероприятиях по оцениванию фактических результатов обучения, осуществляемых преподавателем, ведущим дисциплину.

Объектами оценивания выступают:

- учебная дисциплина – активность на занятиях, своевременность выполнения различных видов заданий, посещаемость занятий;
- степень усвоения теоретических знаний и уровень овладения практическими умениями и навыками по всем видам учебной работы, проводимых в рамках семинаров, практических занятий и самостоятельной работы.

Оценивание обучающегося на занятиях осуществляется с использованием нормативных оценок по 4-х бальной системе (5-отлично, 4-хорошо, 3-удовлетворительно, 2-не удовлетворительно).

**Промежуточная аттестация аспирантов.** Промежуточная аттестация аспирантов по дисциплине проводится в соответствии с локальным актом ФИЦ ИУ РАН - Положением о текущей, промежуточной и итоговой аттестации аспирантов ФИЦ ИУ РАН по программам высшего образования – программам подготовки научно-педагогических кадров в аспирантуре и является обязательной.

Промежуточная аттестация по дисциплине осуществляется в форме зачета в период зачетно-экзаменационной сессии в соответствии с Графиком учебного процесса по приказу (распоряжению заместителя директора по научной работе). Аспирант допускается к зачету в случае выполнения аспирантом всех учебных заданий и мероприятий, предусмотренных настоящей программой. В случае наличия учебной задолженности (пропущенных занятий и (или) невыполненных заданий) аспирант отрабатывает пропущенные занятия и выполняет задания.

**Оценивание аспиранта на промежуточной аттестации в форме зачета.**

| **Оценка зачета** (нормативная) | **Требования к знаниям и критерии выставления оценок** |
|---|---|
| *Зачтено* | Аспирант при ответе демонстрирует содержание тем учебной |

| | |
|---|---|
| | дисциплины, владеет основными понятиями, имеет представление об особенностях теории вычислительных систем, обладает навыком по концептуальному проектированию вычислительных систем, изучил основные методы проектирования программных комплексов. Информирован и способен делать анализ проблем и намечать пути их решения |
| *Не зачтено* | Аспирант при ответе демонстрирует плохое знание значительной части основного материала в области теории вычислительных систем. Не информирован или слабо разбирается в проблемах, и или не в состоянии наметить пути их решения. |

**Перечень контрольных вопросов:**

1. Задачи обработки текста. Многозначность при обработке текста. Проблема понимания
2. Регулярные выражения
3. Конечные автоматы, распознавание языка с помощью КА
4. Регулярные языки и конечные автоматы. Построение КА для регулярных выражений
5. Проверка статистических гипотез для поиска словосочетаний. Проверка по критерию Стьюдента.
6. Проверка статистических гипотез для поиска словосочетаний. Критерий согласия Пирсона
7. Проверка статистических гипотез для поиска словосочетаний. Отношение правдоподобия
8. Проверка статистических гипотез для поиска словосочетаний. Информационно-теоретический подход к поиску словосочетаний
9. Модель N-грамм. Оценка вероятности высказывания
10. Модель N-грамм. Сглаживание (Лапласа и Откат)
11. Модель N-грамм. Оценка качества. Тренировочный и проверочный корпуса
12. Задача определения частей речи. Существующие подходы
13. Использование скрытой марковской модели для определения частей речи
14. Скрытые марковские модели. Вероятность последовательности. Прямой алгоритм
15. Скрытые марковские модели. Наиболее правдоподобное объяснение. Алгоритм Витерби
16. Модели классификации. Наивный байесовский классификатор
17. Модели классификации. Логистическая регрессия
18. Модели классификации. Модель максимальной энтропии
19. Модели классификации. Марковская модель максимальной энтропии
20. Типы грамматик. Грамматика составляющих. Грамматика зависимостей. Категориальная грамматика
21. Контекстно-свободные грамматики. КС грамматики и регулярные языки. Банк деревьев.
22. Синтаксический разбор. Разбор сверху вниз и снизу вверх
23. Синтаксический разбор. Алгоритм Кока-Янгера-Касами (CKY parsing). Эквивалентность КС грамматик
24. Фрагментирование
25. Стохастические контекстно-свободные грамматики. Разрешение синтаксической многозначности
26. Моделирование языка. Обучение стохастических КС грамматик
27. Вероятностная версия алгоритма Кока-Янгера-Касами. Оценка качества
28. Проблемы стохастический КС грамматик. Алгоритм Коллинза. Оценка качества
29. Лексическая семантика. WordNet. Значения слов
30. Разрешение лексической многозначности. Алгоритмы классификации. Самонастройка. Методы оценки качества

31. Разрешение лексической многозначности. Методы основанные на словарях и тезаурусах. Варианты алгоритма Леска. Методы оценки качества
32. Семантическая близость слов. Подходы на основе тезаурусов. Методы оценки качества
33. Семантическая близость слов. Подходы на основе статистик. Методы оценки качест ва
34. Вопросно-ответные системы. Общая архитектура. Обработка запроса
35. Вопросно-ответные системы. Общая архитектура. Извлечение фрагментов текста
36. Вопросно-ответные системы. Общая архитектура. Обработка ответа.
37. Автоматическое реферирование. Общая архитектура
38. Машинный перевод. Классические подходы
39. Статистический машинный перевод. Модель зашумленного канала. Модель перевода на основе фраз. Выравнивание фраз (если слова выровнены). Декодирование
40. Статистический машинный перевод. Выравнивание слов. Модель IBM Model 1
41. Статистический машинный перевод. Выравнивание слов. Тренировка моделей выравнивания
42. Статистический машинный перевод. Методы оценки качества. BLUE.

## 5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

**Основная литература**

1. Daniel Jurafsky and James H. Martin. 2008. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Second Edition. Prentice Hall.

2. Christopher D. Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press.

3. Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. O'Reilly Media, 2009 (http://www.nltk.org/book)

**Электронные ресурсы, включая доступ к базам данных и . т.д.**

**Информационные ресурсы:** Журналы по обработке текстовой информации (ComputationalLinguistics, ACL Journal), труды конференций (ACL, EACL, COLING, EMNLP, Диалог), доступные через Internet, электронные конспекты лекций, разработанные для данного курса.

## 6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

**Необходимое оборудование для лекций и практических занятий:** Компьютер и мультимедийное оборудование (проектор, звуковая система)

**Необходимое программное обеспечение:** ОС Microsoft Windows, Linux, MS Office, включая MS PowerPoint, любой браузер для доступа в Интернет

**Обеспечение самостоятельной работы** - базы данных по журналам Computational Linguistics, ACL Journal

Программу составил д.ф.-м.н. Серебряков В.А.