РОССИЙСКАЯ АКАДЕМИЯ НАУК

Федеральное государственное бюджетное учреждение науки
Институт системного программирования
Российской академии наук

«УТВЕРЖДАЮ»

Директор ИСП РАН
академик РАН,
д.ф.-м.н., профессор
_____ В.П.Иванников
«___»_____ 2012 г.

РАБОЧАЯ ПРОГРАММА

УЧЕБНОЙ ДИСЦИПЛИНЫ
«Основы обработки текстовой информации»

для подготовки аспирантов по специальности
05.13.11 - Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Москва 2012

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. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП ПОСЛЕВУЗОВСКОГО ПРОФЕССИОАЛЬНОГО ОБРАЗОВАНИЯ (АСПИРАНТУРА)

Дисциплина «Основы обработки текстовой информации» относится к дисциплинам по выбору учебного плана подготовки аспирантов по научной специальности 05.13.11 «Математическое обеспечение вычислительных машин, комплексов и компьютерных сетей».

Изучение данной дисциплины базируется на следующих дисциплинах подготовки бакалавров или специалистов:

- «Линейная алгебра»,
- «Теория вероятности и математическая статистика»;
- «Программирование и основы алгоритмизации»;
- «Базы данных»,
- «Искусственный интеллект»,
- «Методы оптимизации»,

а также на дисциплинах подготовки магистра:

- «Современные проблемы информатики и вычислительной техники»;
- «История и методология информатики и вычислительной техники»;
- «Компьютерные технологии в науке и образовании».

Для успешного изучения курса аспиранту необходимо знать общесистемное программное и техническое обеспечения автоматизированных систем, а также уметь работать с персональной ЭВМ.

Основные положения дисциплины будут использованы при подготовке к кандидатскому экзамену по научной специальности 05.13.11 «Математическое обеспечение вычислительных машин, комплексов и компьютерных сетей», в научно-исследовательской работе и при выполнении диссертации на соискание ученой степени кандидата физико-математических наук.

3. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ СОДЕРЖАНИЯ ДИСЦИПЛИНЫ

В результате изучения дисциплины «Основы обработки текстовой информации» аспирант должен:

– **иметь представление**: о месте и роли дисциплины «Основы обработки текстовой информации» в своей будущей научной и практической деятельности, о взаимосвязи дисциплины с другими дисциплинами, наукой и техникой; о современных автоматизированных системах, используемых для обработки текстов;

– **знать**: модели и алгоритмы, применяемые для обработки текстовой информации; современные проблемы обработки текстовой информации; подходы к экспериментальному исследованию качества решения задач обработки текстовой информации;

– **уметь**: решать задачи из области обработки текстов; проводить самостоятельные научные исследования по теме дисциплины; применять изученные модели и алгоритмы для решения поставленных задач.

4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Структура преподавания дисциплины

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 час.

ВИД ЗАНЯТИЙ
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))	32 часа

ВИДЫ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

№ п.п.	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 контролируются преподавателем на лекционных занятиях, используются конспект лекций, учебники, рекомендуемые данной программой	32
2.	Самостоятельное изучение отдельных подразделов программы – выполняется каждым аспирантом по заданию преподавателя, результаты контролируются преподавателем на лекционных занятиях, используются материалы, рекомендуемые данной программой	26
ВСЕГО (зач. ед.(часов))		58 часов

Содержание дисциплины

Развёрнутые темы и вопросы по разделам

№ п/п	Название модулей	Разделы и темы лекционных занятий	Содержание	Объем	
				Аудиторная работа (зачетные единицы/часы)	Самостоятельная работа (зачетные единицы/часы)
1		Задачи обработки текстов	Задачи обработки текста. Многозначность при обработке текста. Проблема понимания	2	1
2		Регулярные выражения и конечные автоматы	Регулярные выражения, Конечные автоматы, распознавание языка с помощью КА. Регулярные языки и конечные автоматы. Построение КА для регулярных выражений	2	5
3		Методы поиска словосочетаний	Проверка статистических гипотез для поиска словосочетаний, t-тест, критерий хи-квадрат, отношение правдоподобия, информационно-теоретический подход к поиску словосочетаний	2	5
4		Языковые модели и задача опреде-	Модель N-грамм. Оценка вероятности высказывания, методы	2	5

		ления частей речи	сглаживания Лапласа и Отката, Оценка качества, Тренировочный и проверочный корпуса, Задача определения частей речи, существующие подходы		
5		Методы обучения с учителем и задачи обработки текстов	Использование скрытой марковской модели для определения частей речи, Скрытые марковские модели, Вероятность последовательности, Прямой алгоритм, Наиболее правдоподобное объяснение, Алгоритм Витерби, Наивный байесовский классификатор, Логистическая регрессия, Модель максимальной энтропии, Марковская модель максимальной энтропии	4	7
6		Контекстно-свободные грамматики 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	7
7		Статистические методы синтаксического анализа	Стохастические контекстно-свободные 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	7
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 осно-	4	7

			ве статистик.		
9		Вопросно-ответные системы и автоматическое реферирование	Вопросно-ответные системы. Общая архитектура. Обработка запроса. Извлечение фрагментов текста. Автоматическое реферирование. Общая архитектура	4	7
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	7

5. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

В учебном процессе используются следующие образовательные технологии:

№ п/п	Вид занятия	Форма проведения занятий	Цель
1	Лекция	Изложение теоретического материала	Получение теоретических знаний по дисциплине
2	Лекция	Изложение теоретического материала с помощью презентаций	Повышение степени понимания материала
3	Лекция	Разбор конкретных примеров применения современных технологий обработки текстов	Осознание связей между теорией и практикой, а также взаимозависимостей разных дисциплин
4	Самостоятельная работа аспиранта	Самостоятельное изучение отдельных подразделов программы. Самоподготовка (проработка и повторение лекционного материала и материала рекомендованной литературы)	Повышение степени понимания материала

6. ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ, ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ИТОГАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ И УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ АСПИРАНТОВ

Форма контроля знаний:
- кандидатский экзамен по специальности.

Контрольно-измерительные материалы

На кандидатском экзамене аспирант должен продемонстрировать знания в объеме основной программы кандидатского экзамена по специальности 05.13.11 «Математическое обеспечение вычислительных машин, комплексов и компьютерных сетей», а также дополнительной программы, в которую, в зависимости от выбранной аспирантом специализации, могут входить вопросы, рассматриваемые в данном курсе.

Перечень контрольных вопросов для дополнительной программы:

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 качества
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

7. 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

8. 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 (Computational Linguistics, ACL Journal), труды конференций (ACL, EACL, COLING, EMNLP, Диалог), доступные через Internet, электронные конспекты лекций, разработанные для данного курса.

Программу составил к.ф.–м.н. Турдаков Д.Ю.

Программа принята на заседании Ученого Совета ИСП РАН
протокол № 2012-5 от 23 мая 2012 г.