Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего профессионального образования "Национальный исследовательский университет "Высшая школа экономики"

Факультет гуманитарных наук
Школа лингвистики

Программа дисциплины «Информационный поиск и извлечение данных»

для образовательной программы «Фундаментальная и компьютерная лингвистика»
направления 45.03.03 «Фундаментальная и прикладная лингвистика»
подготовки бакалавра

Авторы программы:
А.Б. Кутузов, доцент, akutuzov72@gmail.com
А.А. Бонч-Осмоловская, кандидат филологических наук, abonch@gmail.com

Одобрена на заседании школы лингвистики «24» апреля 2015 г.
Руководитель школы Е.В. Рахилина _____

Рекомендована Академическим советом образовательной программы
«28» апреля 2015 г., Протокол № 3

Утверждена «21» мая 2015 г.
Академический руководитель образовательной программы
Ю.А. Ландер _____

Москва, 2015

1. Область применения и нормативные ссылки

Настоящая программа учебной дисциплины устанавливает минимальные требования к знаниям и умениям студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и бакалавров направления подготовки 45.03.03 «Фундаментальная и прикладная лингвистика» Школы лингвистики Факультета гуманитарных наук.

Программа разработана в соответствии с:

- Образовательным стандартом федерального государственного автономного образовательного учреждения высшего профессионального образования национального исследовательского университета «Высшая школа экономики», в отношении которого установлена категория «национальный исследовательский университет»
- Учебным планом университета по направлению подготовки 45.03.03 «Фундаментальная и прикладная лингвистика» для подготовки бакалавра, утвержденным в 2015 г.

.

2. Цели освоения дисциплины

Целями освоения дисциплины являются:

1. ознакомление с современным состоянием исследований в области информационного поиска и извлечения данных;
2. формирование представлений о способах построения программных средств для информационного поиска;
3. углубление знаний о методах автоматической обработки текста.

3. Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

- Знать основные способы разработки систем информационного поиска и их использования;
- Уметь перечислить и охарактеризовать наиболее известные системы информационного поиска;
- Владеть научными методами, необходимыми для разработки поисковых систем;
- Уметь выявлять основные типы информации, необходимые в конкретном случае для разработки системы информационного поиска.

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
-способен рефлексировать (оценивать и перерабатывать) освоенные научные методы и способы деятельности; - способен к самостоятельному освоению но-	СК-1 СК-3	Способен к обобщению, анализу, восприятию информации	Формы обучения: -лекции, -семинарские занятия, -самостоятельная работа, -реферирование научной литературы.

Компетенция	Код по ФГОС/ НИУ	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
вых методов исследования; - способен совершенствовать и развивать свой интеллектуальный и культурный уровень, строить траекторию профессионального развития и карьеры	СК-4	Умеет самостоятельно получать информацию о новых методах исследования и оценивать сферу и возможности ее применения Имеет представление о смежных научных направлениях (антропологии, типологии, когнитивных науках и др.)	
Способен самостоятельно разработать методический инструментарий для осуществления исследовательской и проектной деятельности в области фундаментальной и прикладной лингвистики	ПК-6	Владеет методическим инструментарием для описания грамматической системы (подсистем) языка, выявления семантических эффектов	Методы обучения: - анализ кейсов - упражнения для самостоятельной работы - работа с литературой
Способен проводить анализ качества языковых данных, корпусов, систем, использующихся для автоматической обработки естественного языка	ПК-9	Использует данные корпусов, информационных систем, данные экспериментальной лингвистики, психолингвистики и нейролингвистики для эмпирически обоснованного объяснения фактов языка; способен оценивать валидность данных	Методы обучения: - анализ кейсов, - построение объяснительной модели.

4. Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к блоку обязательных теоретических дисциплин программы подготовки бакалавров.

Изучение данной дисциплины базируется на следующих дисциплинах:

- «Теория языка» и другие дисциплины теоретического блока программы подготовки бакалавра «Фундаментальная и прикладная лингвистика» или
- «Компьютерные инструменты лингвистического исследования»

5. Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторные часы			Самостоятельная работа
			Лекции	Семинары	Практические занятия	
1	Введение в информационный поиск	2	4	2		
2	Поиск в Интернете	30	2			

3	Слова в матрицах: взвешивание терминов и модель векторного пространства	40		2		10
4	Введение в fact extraction	4		2		
5	Создание шаблонов для выделения высказываний в неструктурированных текстах	4	2			
6	Предобработка текста для задач fact extraction	8	2	2		8
7	Пишем собственный анализатор текстов, выделяющий высказывания	20		2		20
8	Оцениваем качество анализаторов (точность, полнота, F-мера)			2		8
9	Презентация анализаторов			2		10
10	Системы извлечения информации: принципы работы		2			
11	Система Томита: модули и компоненты		2			10
12	Система Томита: правила грамматики и вывода информации		2			12
	Итого	108	16	14		78

6. Формы контроля знаний студентов

Тип контроля	Форма контроля	1 год				Параметры **
		1	2	3	4	
	Домашнее задание	1				Включает написание собственного анализатора текстов, выделяющего в них высказывания.
Итоговый	Зачет		1			Устный зачет, 120 мин

6.1. Критерии оценки знаний, навыков

Оценка 10-8 баллов выставляется при качественной подготовке домашних заданий, активном участии в работе на семинарах, высоком качестве защищаемой модели анализатора, аналитическом подходе к изучаемым темам, наборам данных, проектам, отсутствии ошибок в ответах на зачёте (допускаются незначительные ошибки, которые студент сам исправляет в беседе с преподавателем).

Оценка 6-7 баллов выставляется при наличии отдельных неточностей при выполнении текущих заданий и на зачете, а также недочётах в работе анализатора. Допускается частичная неполнота ответа и незначительное количество ошибок в теоретических вопросах и практическом анализе данных при выполнении проектной деятельности и на зачете.

Оценка в 4-5 баллов выставляется, если не наблюдалось активности в аудиторных занятиях и при самостоятельной подготовке, к работе анализатора есть серьёзные претензии, обнаружены значительные пробелы в области теоретических знаний и принципиальные ошибки в проектной деятельности и анализе данных на зачете.

Оценка в 3 балла выставляется при наличии лишь отдельных положительных моментов в ответе на теоретические вопросы, в проектной деятельности и анализе данных на зачете.

Оценка в 2 балла выставляется при полном отсутствии знаний.

Оценка 1;0 – неправильные ответы сопровождаются демонстративными проявлениями безграмотности или неэтичного отношения к теме и предмету в целом.

6.2. Порядок формирования оценок по дисциплине

Итоговая накопленная оценка состоит из следующих составляющих:

- посещаемость и активное участие в лекциях и семинарах (Qакт) - 15%
- выполнение домашних заданий для самостоятельной работы, чтение литературы (Qсам) - 30%
- доклады на семинарах и коллоквиуме (Qвыст) - 30%
- проектная работа (Qпроект) - 25%

Каждый из параметров оценивается от 0 до 10 баллов. Таким образом:

Qнакопл = 0.15*Qакт + 0.30*Qсам + 0.30*Qвыст + 0.25*Qпроект

Итоговая оценка Qитог вычисляется по формуле:

- итоговая накопленная оценка (Qнакопл) - 60%
- зачет (Qзачет) - 40%

Таким образом:

Qитог = 0.60*Qнакопл + 0.40*Qзачет

Округление оценок при выставлении в ведомость производится в пользу студента. Для более точного определения балла итоговая оценка по дисциплине вычисляется, исходя из неокругленной накопленной оценки (например, 7,8 или 8,25). Текущие оценки и итоговую накопленную оценку студенты могут узнать из ведомости оценок.

7. Содержание дисциплины

1. **Введение в информационный поиск.** Постановка проблемы. История вопроса. Области применения. Необходимая терминология. Индексы: прямой и обратный. Булев поиск. Ранжированный поиск. Мастерская по строительству обратного индекса

Литература

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. 2008.
Introduction to Information Retrieval. CUP. http://informationretrieval.org

2. **Поиск в Интернете.** Поисковые машины Рунета. Почему искать в Интернете сложно? Экономическая модель поисковика. Пользователи и их запросы. Как расправиться с дубликатами. Обкачка: пауки в паутине. Веб как граф. Алгоритмы ссылочного ранжирования.

Литература

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. 2008.
Introduction to Information Retrieval. CUP. http://informationretrieval.org

3. Слова в матрицах: взвешивание терминов и модель векторного пространства. Ранжированная выдача. Взвешивание по частоте термов. Обратная документная частота. Формула tf-idf. Модель векторного пространства. Запрос как вектор. Сравнение векторов.

Литература

Апресян Ю.Д. Лексическая семантика. М., 1996. С. 95-113 («Требования к толкованиям и толкуемым выражениям»), 119-133 («Семантические валентности слова»). http://bookre.org/reader?file=680272

Апресян Ю.Д. (отв. ред.). Новый объяснительный словарь синонимов. М., 2003. Введение, глава «Лингвистическая терминология словаря» (с. 22-52 (XXII-LII)). http://www.lrc-lib.ru/ruslang/noss/intro.pdf

Апресян Ю.Д. Трехуровневая теория управления: лексикографический аспект // Апресян Ю.Д. и др. Теоретические проблемы русского синтаксиса. Взаимодействие грамматики и словаря. М., 2010. (см. Ридинг в LMS)

4. Введение в fact extraction.

Ч. Филлмор: падежная грамматика, фреймовая теория. Аргументная структура, модель управления. Роли и валентности. Агенс – Пациенс.

Литература

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. 2008. Introduction to Information Retrieval. CUP. http://informationretrieval.org

5. Создание шаблонов для выделения высказываний в неструктурированных текстах.

6. Предобработка текста для задач fact extraction. Типы обработки исходного текста: предобработка (преобразование форматов, кодировок и др.; графематический анализ, разбиение текста на составные части; морфологический анализ; лемматизация). Токенизация. Прагматическая обработка.

Литература

Manning, C. D. (1999). Foundations of statistical natural language processing. H. Schütze (Ed.). MIT press.

Martin, J. H., & Jurafsky, D. (2000). Speech and language processing. International Edition.

7. Пишем собственный анализатор текстов, выделяющий высказывания.

8. Оцениваем качество анализаторов (точность, полнота, F-мера). Понятие точности. Понятие полноты. Понятие F-меры. Согласованность критериев и конфликты между ними.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. 2008. Introduction to Information Retrieval. CUP. http://informationretrieval.org

Manning, C. D. (1999). Foundations of statistical natural language processing. H. Schütze (Ed.). MIT press.

9. Презентация анализаторов. Студенты рассказывают о составленных ими анализаторах и демонстрируют их работу..

10. Системы извлечения информации: принципы работы. Основные современные системы извлечения информации из текста: Apachi, RCO Fact Extractor, Томита-парсер и др. Параметры описания систем: язык текста, типы извлекаемых сущностей, набор модулей анализа.

Литература

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. 2008. Introduction to Information Retrieval. CUP. http://informationretrieval.org

Manning, C. D. (1999). Foundations of statistical natural language processing. H. Schütze (Ed.). MIT press.

Martin, J. H., & Jurafsky, D. (2000). Speech and language processing. International Edition.

11. Система Томита: модули и компоненты. GLR-парсинг и алгоритмы системы Томита. Способы использования парсера Томита. Основные компоненты Томита: словарь, грамматика.

Литература

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. 2008. Introduction to Information Retrieval. CUP. http://informationretrieval.org

Martin, J. H., & Jurafsky, D. (2000). Speech and language processing. International Edition.

12. Система Томита: правила грамматики и вывода информации. Устройство грамматических правил системы Tomita. Общий вид правила. Форма записи слов. Знаки пунктуации. Устройство синтаксических правил, типы составляющих. Запись пути к исходному тексту. Преобразование цепочек в факты.

8. Образовательные технологии

Образовательные технологии включают лекции, семинары, встречи с ведущими специалистами-экспертами, проектную деятельность, самостоятельную работу студентов, тесты для проверки усвоения материала. Инновационные образовательные технологии включают работу с современными электронными лингвистическими ресурсами (словари, корпуса, интернет-порталы) и инструментарием для проведения экспериментов. Материалы по курсу размещены в системе LMS, в ней же студенты могут получить и отправить задания, узнать текущие оценки и обсудить учебные вопросы в форуме.

При проведении занятий используются интерактивные формы (не менее 60% от аудиторной нагрузки, включая коллоквиум, обсуждение проектов, обсуждение презентаций студентов, обсуждение моделирования экспериментов и описания данных, брифы по вопросам студентов и т.д.). Кроме того, обучение предполагает большую долю самостоятельной внеаудиторной работы, направляемой и контролируемой преподавателем. Студенты имеют возможность получать консультации во внеаудиторное время через систему LMS, а также по электронной почте.

9. Оценочные средства для текущего контроля и аттестации студента

a. Тематика заданий текущего контроля

Примерные задания для самостоятельной внеаудиторной работы:

1. Построение собственного анализатора для выделения высказываний в текстах.
2. Оценка работы поисковых систем.
3. Оценка сложности того или иного типа текстов для информационного поиска.
4. Выработка критериев оценки работы анализаторов.
5. Проверка критериев на материале текстов разных функциональных стилей.

b. Вопросы для оценки качества освоения дисциплины (зачёт)

Примерные вопросы для зачета:

1. Назовите основные русскоязычные системы поиска. В чём состоят их достоинства и недостатки?

2. Каковы основные сферы применения технологий информационного поиска?

3. Какие сложности возникают при поиске в интернете?

4. Какое место в описании языка должно занимать изменение языковой системы в течение жизни у носителя языка? Варьирование системы по диалектам; идиолектам; состояние языка у носителей и их потомков, проживающих за рубежом?

5. В чём состоят основные особенности системы Tomita? Какие модули и компоненты в ней выделяются?

6. Какие системы извлечения информации вы знаете?

7. Назовите основные типы информации, извлекаемой из текста. В каких случаях нужен каждый из них?

10. Учебно-методическое и информационное обеспечение дисциплины

Основная литература:

Alani, H., Kim, S., Millard, D. E., Weal, M. J., Lewis, P. H., Hall, W., & Shadbolt, N. R. (2003). Automatic extraction of knowledge from web documents.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. 2008. Introduction to Information Retrieval. CUP. http://informationretrieval.org

Manning, C. D. (1999). Foundations of statistical natural language processing. H. Schütze (Ed.). MIT press.

Martin, J. H., & Jurafsky, D. (2000). Speech and language processing. International Edition.

Pasca, M., Lin, D., Bigham, J., Lifchits, A., & Jain, A. (2006, July). Organizing and searching the world wide web of facts-step one: the one-million fact extraction challenge. In AAAI (Vol. 6, pp. 1400-1405).

Zobel, Justin, and Alistair Moffat. 2006. Inverted files for text search engines. ACM Computing Surveys 38(2)

Дополнительная литература

11. Материально-техническое обеспечение дисциплины

Для лекций и семинаров используется компьютер/ноутбук; проектор; экран.