**Правительство Российской Федерации**

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

Факультет филологии

**Программа дисциплины** Мастер-класс по компьютерной лингвистике

для направления 035800.68 «Фундаментальная и прикладная лингвистика» для магистерской программы «Компьютерная лингвистика»

Авторы программы:

Бонч-Осмоловская А.А., к.ф.н, abonch@gmail.com

Одобрена на заседании профессорско-преподавательского состава направления подготовки «Фундаментальная и прикладная лингвистика» «04» апреля 2013 г.

Руководитель направления: Рахилина Е.В.

Рекомендована секцией УМС по лингвистике, филологии и журналистике «18» мая 2013 г.

Председатель: Поливанов К.М.

Утверждена УС факультета филологии «14» февраля 2014 г.

Ученый секретарь Никишина Е.А. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ [подпись]

Москва, 2012

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

# 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 студентов направления подготовки/ специальности 035800.68 «Фундаментальная и прикладная лингвистика» , обучающихся по магистерской программе «Компьютерная лингвистика» изучающих дисциплину «Мастер-класс по компьютерной лингвистике».

Программа разработана в соответствии с:

* Стандартом НИУ ВШЭ по направлению подготовки 035800.68 «Фундаментальная и прикладная лингвистика» подготовки магистра для магистерской программы «Компьютерная лингвистика»;
* Рабочим учебным планом университета по направлению подготовки 035800.68 «Фундаментальная и прикладная лингвистика» подготовки магистра для магистерской программы «Компьютерная лингвистика» утвержденным в 2012г.

# Цели освоения дисциплины

Целями освоения дисциплины Мастер класс по компьютерной лингвистике являются знакомство с основными проблемами, методами и подходами в области компьютерной лингвистики, основными инструментами и технологиями, ключевыми задачами и диапазоном имеющихся решений.

# Компетенции обучающегося, формируемые в результате освоения дисциплины

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 к самостоятельному освоению новых методовисследования, изменению научного и научно-производственногопрофиля своей деятельности | СК-М 3 | понимает основные направления современной компьютерной лингвистики, знает возможности применения автоматического анализа языка в смежных дисциплинах и для неязыковых задач, знает основные ресурсы, аггрегирующие информацию о современном состоянии компьютерной лингвистики  | - чтение специальной литературы- дискуссия с приглашенными докладчиками- обзор ключевых интернет ресурсов |
| Способен проводить практические занятия по дисциплинам профильной направленности в образовательных учреждениях общего и среднего профессионального образования, используя современные интерактивные методики обучения и информационно-коммуникационные технологии | ПК-12 | Умеет организовать рассказ о своей деятельности в виде доступного мастер-класса. Умеет использовать для иллюстрации интерактивные открытые ресурсы по компьютерной лингвистике | - подготовка научно-популярной статьи по одной из тем современной компьютерной лингвистики, подбор актуальный открытых интерактивных ресурсов |
| Способен задавать, транслировать правовые и этические нормы в профессиональной и социальной деятельности | ПК-23 |  | - чтение специальной литературы- дискуссия с приглашенными докладчиками |
| Способен определять, транслировать общие цели в профессиональной и социальной деятельности  | ПК-17 | знает основные лингвистические ресурсы, владеет цепочкой лингвистической обработкой текста, умеет использовать среды разработки лингвистических систем | - чтение специальной литературы- дискуссия с приглашенными докладчиками |

# Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к циклу профессиональных дисциплин, обязательных для изучения.

Изучение данной дисциплины базируется на следующих дисциплинах:

* Введение в лингвистику (адаптационный курс) или курс по теория языка программы подготовки бакалавра

Для освоения учебной дисциплины, студенты должны владеть следующими знаниями и компетенциями:

* Владеть базовыми представлениями о грамматических категориях и анализе языковых единиц

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

* Машинный перевод, корпусная лингвистика, онтологии и семантические технологии

# Тематический план учебной дисциплины

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Название раздела | Всего часов  | Аудиторные часы | Самостоя­тельная работа |
| Лекции | Семинары | Практические занятия |
| 1 | Лингвистические задачи поисковых систем | 54 |  | 18 |  | 36 |
| 2 | Лингвистические проблемы в машинном переводе | 24 |  | 12 |  | 12 |
| 3 | Современные лингвистические технологии | 74 |  | 18 |  | 56 |
|  | ИТОГО | 152 |  | 48 |  | 104 |

# Формы контроля знаний студентов

|  |  |  |  |
| --- | --- | --- | --- |
| Тип контроля | Форма контроля | 1 год | Параметры \*\* |
| 1 | 2 | 3 | 4 |
| Текущий(неделя) | Домашняя работа |  |  |  | 1 | Подготовка статьи в открытый энциклопедический ресурс nlpub.ru (20 часов).  |
| Итоговый | Зачет |  |  |  | 1 | Письменный зачет |

## Критерии оценки знаний, навыков

Активность на занятиях, самостоятельная подготовка к темам занятиям, чтение специальной литературы, знакомство с основными направлениями компьютерной лингвистики, обсуждавшимися на занятиях, подходами и методами, выполнение домашнего задания.

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

## Порядок формирования оценок по дисциплине

*Отекущий* = *Одз*;

Результирующая оценка за итоговый контроль в форме зачета выставляется по следующей формуле, где *Озачет* – оценка за работу непосредственно на зачете:

 *Оитоговый = k1·Озачет + k2·О дз*

При этом удельный вес форм контроля распределяется следующим образом:

*k1 = 0,5*

*k2 = 0,5*

Способ округления накопленной оценки итогового контроля в форме зачета: в пользу студента.

# Содержание дисциплины

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Название | часы аудиторной работы | самостоятельная работа (часы)  | самостоятельная работа (вид работы) |
| Раздел 1 | Лингвистические технологии поисковых систем | 18 | 36 | чтение литературы |
| Тема 1 | Лингвистические технологии, используемые в разработке поисковика Яндекс: алгоритмы индексации документов. Нормализация лингвистических данных | 6 | 12 | чтение литературы |
| Тема 2 | Лингвистические технологии, используемые в разработке НКРЯ: алгоритмы создания и индексации корпуса, виды корпусных разметок | 6 | 12 | чтение литературы |
| Тема 3 | Лингвистические технологии, используемые в разработке поисковика mail.ru: алгоритмы машинного обучения для оценки качества поисковика. | 6 | 12 | чтение литературы |
| Раздел 2 | Лингвистические технологии систем машинного перевода | 12 | 12 | чтение литературы |
| Тема 4 | Лингвистические технологии, используемые в разработках ABBYYCompreno: формализмы синтаксического анализа, формализмы семантического анализа, извлечение фактов, настройка факт-экстарктора | 6 | 12 | чтение литературы |
| Раздел 3 | Ключевые технологии лингвистики и веба | 18 | 56 | чтение литературы |
| Тема 5 | Тенденции развития интернет среды: современные технологии и автоматический анализ языка, социальные сети, краудсорсинг | 6 | 12 | чтение литературы |
| Тема 6 | Методы кластеризации текста: основные алгоритмы – k-means, Байесовские алгоритмы, метод опорных векторов, метод LDA | 6 | 12 | чтение литературы |
| Тема 7 | Ведение лингвистических проектов: рабочее место лингвиста, примеры, функционал, возможности | 3 | 12 | чтение литературы |
| Тема 8 | Создание лингвистических онтологий и тезаурусов: Semantic web, языки owl rdf, LOD, лексические онтологии Раскина | 3 | 20 | подготовка домашнего задания |

Литература по разделу:

Раздел 1 . Лингвистические технологии поисковых систем

Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск //М.: Вильямс. – 2011. Главы 4-9,

Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009 ( глава по выбору студента)

Раздел 2. Лингвистические технологии машинного перевода

Базовый учебник: Daniel Jurafsky and James H. Martin (2008): Speech and Language Processing (2nd edition) ch. 23

Раздел 3.

Базовый учебник: Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск //М.: Вильямс. – 2011. Главы 11-19

Heath T., Bizer C. Linked data: Evolving the web into a global data space //Synthesis lectures on the semantic web: theory and technology. – 2011. – Т. 1. – №. 1. – С. 1-136., доступна по ссылке http://linkeddatabook.com/editions/1.0/,

Все разделы:

Материалы конференции «Диалог» <http://www.dialog-21.ru/>

Материалы конференции РОМИП <http://romip.ru/>

Материалы журнала Computational Linguistics <http://www.mitpressjournals.org/loi/coli>

Материалы ассоциации компьютерной лингвистики <http://www.aclweb.org/>

# Образовательные технологии

В рамках учебного курса могут быть предусмотрены встречи с российскими или зарубежными практикующими специалистами, занятыми в области разработки лингвистических систем, часть занятий проводится на территории компаний, специализирующихся в области компьютерной лингвистики. Часть занятий может проводиться с помощью технологий дистанционных семинаров в формате skype-conference.

# Оценочные средства для текущего контроля и аттестации студента

## Примеры заданий итогового контроля

Примерные вопросы итогового письменного зачета

1. Опишите основные модули поисковой системы и их функционал.
2. Опишите основные модули систем машинного перевода и их функционал
3. Опишите основные принципы индексации разметки НКРЯ и его подкорпусов
4. Опишите лингвистические задачи, где применяются методы кластеризации текста?
5. Опишите возможности применения лингвистического анализа в платформах социальных сетей?
6. Опишите алгоритмы и модели улучшения качества лингвистического анализа с помощью социальных сетей и краудсорсинга.
7. Опишите принципы сбора открытых связанных данных.
8. Опишите использование лингвистического анализа в архитектуре семантического веба.
9. Опишите основной функционал рабочего места лингвиста для составления тезаурусов и словарей
10. Опишите основной функционал рабочего места лингвиста для снятия морфологической и синтаксической омонимии
11. Опишите основной функционал рабочего места лингвиста для семантического анализа

## Вопросы для оценки качества освоения дисциплины

1. Типы поисковых систем и их архитектура
2. Основные алгоритмы машинного перевода.
3. Принципы создания и индексации лингвистических корпусов.
4. Методы кластеризации и классификации текстов
5. Функционал рабочего места лингвиста для решения задач разных языковых уровней
6. Семантический веб и лексические онтологии
7. Связанные открытые данные и способы их извлечения

# Учебно-методическое и информационное обеспечение дисциплины

## Базовый учебник

Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск //М.: Вильямс. – 2011.

Daniel Jurafsky and James H. Martin (2008): Speech and Language Processing (2nd edition)

## Основная литература

Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009

Heath T., Bizer C. Linked data: Evolving the web into a global data space //Synthesis lectures on the semantic web: theory and technology. – 2011. – Т. 1. – №. 1. – С. 1-136., доступна по ссылке http://linkeddatabook.com/editions/1.0/,

Материалы конференции «Диалог» <http://www.dialog-21.ru/>

Материалы конференции РОМИП <http://romip.ru/>

Материалы журнала Computational Linguistics <http://www.mitpressjournals.org/loi/coli>

Материалы ассоциации компьютерной лингвистики <http://www.aclweb.org/>

## Дистанционная поддержка дисциплины

Дистанционная поддержка курса не предусмотрена

# Материально-техническое обеспечение дисциплины

Для проведения занятий необходим проектор и компьютерные классы