Нижегородский филиал
Федерального государственного автономного образовательного учреждения высшего профессионального образования
"Национальный исследовательский университет
"Высшая школа экономики"

Департамент прикладной лингвистики и иностранных языков

Рабочая программа дисциплины

«Автоматическая обработка естественного языка»
(Natural Language Processing)

для образовательной программы «Фундаментальная и прикладная лингвистика»
направления подготовки 45.03.03 «Фундаментальная и прикладная лингвистика»
уровня бакалавр

Разработчик программы:
Малафеев А.Ю., к.филол.н., aumalafeev@hse.ru

Одобрена на заседании департамент прикладной лингвистики и иностранных языков
«___»_____ 2016 г.
Зав. департаментом В.М. Бухаров_____

Утверждена «___»_____ 2016 г.
Академический руководитель образовательной программы
А.Ю.Малафеев _____

Нижний Новгород
2016

The course covers the basics of natural language processing (NLP) used in many real-world applications such as language modeling, text classification, sentiment analysis, summarization and machine translation. The students will not only use some of the existing NLP libraries and software packages, but also learn about the principles behind their design, and about the mathematical models underlying modern computational linguistics. The course also involves completing practical programming assignments in Python and conducting experiments on texts written in English and Russian.

Syllabus:

1. Introduction to natural language processing

Structural features of texts in natural language; ambiguity on all levels of language; the main challenges of natural language processing; basic approaches to problem solving: manually written rules and machine learning.

2. Basic text processing and edit distance

Preprocessing: tokenization and segmentation; normalization of words: stemming, lemmatization, morphological analyzers; regular expressions; edit distance.

3. Language models

N-grams; perplexity; methods of smoothing; the use of language models: input prediction, error correction, speech recognition, text generation.

4. Tagging problems and hidden Markov models

POS tagging; named entity recognition as a tagging problem; hidden Markov models, their advantages and disadvantages; the Viterbi algorithm.

5. Text classification and sentiment analysis

Classification problems; naive Bayes classifier; text classification; sentiment analysis.

6. Evaluation

Performance measures: accuracy, precision, recall, F-measure; state-of-the-art.

7. Parsing

Constituency and dependency trees; context-free grammar; probabilistic approach to parsing; lexicalized PCFGs; CKY algorithm.

8. Machine translation

Classical approaches: direct, transfer-based, interlingual; statistical machine translation; IBM model; alignment; parameter estimation in IBM models; phrase-based translation models.

9. Computational semantics

Word senses and meanings; WordNet; semantic similarity measures: thesaurus-based and distributional methods.

10. Text summarization

Extractive and abstractive summarization; multiple-document summarization; query-based summarization; supervised and unsupervised learning; evaluation of summarization systems; ROUGE.

Prerequisites: knowledge of general linguistics and corpus linguistics; programming skills (Python).
Author: Alexey Malafeev, Foreign Languages Department, Associate Professor.
Exam type: oral.

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 студентов направления 45.03.03 «Фундаментальная и прикладная лингвистика», изучающих дисциплину «Автоматическая обработка естественного языка».

Программа разработана в соответствии с:

- образовательным стандартом НИУ ВШЭ для направления 45.03.03 «Фундаментальная и прикладная лингвистика»;
- образовательной программой направления 45.03.03 «Фундаментальная и прикладная лингвистика»;
- учебным планом университета по направлению подготовки направления 45.03.03 «Фундаментальная и прикладная лингвистика», утвержденным в 2016 г.

2 Цели освоения дисциплины

Дисциплина нацелена на овладение студентами основами автоматической обработки текстов, написанных на естественном языке. Это предполагает не только умение использовать готовые приложения для лингвистического анализа, но и понимание принципов их работы, а также знакомство с базовыми математическими моделями, лежащими в основе современной компьютерной лингвистики. Дисциплина также включает выполнение практических заданий с помощью языка программирования Python. Дисциплина преподается на английском языке, поэтому дополнительная цель ее изучения – развитие навыков профессионального общения на английском языке.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

- Знать структурные особенности текстов на естественном языке и принципы их обработки с помощью компьютера с целью получения лингвистической (морфологической, синтаксической, семантической) информации; иметь представление о методах, применяемых для решения сложных прикладных задач автоматической обработки текста, в частности, информационного поиска, автоматического реферирования, анализа тональности, машинного перевода; понимать ограничения существующих компьютерных моделей автоматической обработки текстов.
- Уметь применять существующие системы автоматической обработки, определять преимущества и недостатки этих систем, оценивать и сравнивать результаты их работы.
- Иметь навыки (приобрести опыт) решения конкретных задач автоматической обработки текста, в том числе при помощи языка программирования Python, а также проведения экспериментов на текстовом материале.

В результате освоения дисциплины студент осваивает следующие компетенции:

Компетенция	Код по ОС НИУ ВШЭ	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
Профессиональные компетенции:			

Компетенция	Код по ОС НИУ ВШЭ	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
Знает основы математических дисциплин, которые используются при формализации лингвистических знаний и процедур анализа и синтеза лингвистических структур: теории множеств, теории вероятностей и математической статистики, дискретной математики, математической логики, теории автоматов и формальных грамматик.	ПК-2	Использует математические методы решения задач компьютерной лингвистики.	Семинары, самостоятельная работа, групповая работа, проектная деятельность
Способен свободно вести профессиональное письменное и устное общение на первом иностранном языке.	ПК-6	Формулирует и отвечает на вопросы по тематике дисциплины на английском языке.	Лекции, семинары, групповые дискуссии
Умеет использовать лингвистические технологии для проектирования систем анализа и синтеза естественного языка, анализа и синтеза мультимодальных языковых систем, в том числе лингвистических компонентов интеллектуальных и информационных электронных систем.	ПК-18	Выполняя задания на программирования, реализует системы анализа и синтеза естественного языка.	Семинары, самостоятельная работа, групповая работа, проектная деятельность
		Успешно собирает и систематизирует данные, полученные в ходе групповых и индивидуальных мини-исследований.	Семинары, самостоятельная работа, групповая работа, проектная деятельность
Умеет проводить квалифицированное тестирование эффективности лингвистически ориентированного программного продукта.	ПК-19	Оценивает и тестирует системы автоматической обработки текстов.	Семинары, самостоятельная работа
Общекультурные компетенции:			
Стремится к саморазвитию, повышению своей квалификации и мастерства.	ОК-6	Обосновывает для себя необходимость изучения и практики в области корпусной лингвистики для профессионального и личностного развития.	Лекции, семинары, групповые дискуссии
Способен работать с ин-	ОК-13	Получает информацию, необходи-	Проекты, самостоятель-

Компетенция	Код по ОС НИУ ВШЭ	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
формацией в глобальных компьютерных сетях.		мую для выполнения заданий и проектов, из сети Интернет.	ная работа

4 Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к профессиональному циклу для направления 45.03.03 «Фундаментальная и прикладная лингвистика». Дисциплина изучается на третьем курсе в 1, 2 и 3 модулях.

Изучение данной дисциплины базируется на следующих дисциплинах: «Введение в лингвистику», «Теория языка», «Компьютерные инструменты лингвистического исследования». При практической работе на семинарах студенты используют навыки, полученные в рамках освоения дисциплины «Программирование».

Основные положения дисциплины должны быть использованы в дальнейшем при изучении дисциплин «Машинное обучение», «Анализ и синтез звучащей речи», полученные практические навыки – при освоении дисциплины «Программирование для лингвистов», а также при подготовке выпускной квалификационной работы.

5 Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторные часы			Самостоятельная работа
			Лекции	Семинары	Практические занятия	
1	Введение в автоматическую обработку естественного языка	19	3		4	12
2	Базовая обработка текста и дистанция редактирования	24	2		6	16
3	Языковые модели	20	2		4	14
4	Задачи разметки текста и скрытые марковские модели	25	3		6	16
5	Классификация текстов и анализ тональности	20	2		4	14
6	Информационный поиск	24	2		6	16
7	Парсинг	25	3		6	16
8	Машинный перевод	21	3		4	14
9	Компьютерная семантика	25	3		6	16
10	Автоматическое реферирование	25	3		6	16
	Всего	228	26		52	150

Количество зачетных единиц на дисциплину – 6.

6 Формы контроля знаний студентов

Тип	Форма	Параметры

контроля	контроля	2	3	
Текущий	Контрольная работа	+		Тесты, 20 вопросов, 40 минут на выполнение
Итоговый	Экзамен		+	Устный, 2 вопроса на билет, 30 минут на подготовку, 10 минут на ответ. Дополнительные вопросы в случае спорной оценки (без времени на подготовку)

6.1 Критерии оценки знаний, навыков

Студент должен продемонстрировать знание основных понятий и актуальных проблем корпусной лингвистики в объеме, достаточном для осуществления практической деятельности в области корпусных технологий и исследований. Студент должен обладать навыками эффективного использования компьютерных инструментов, изучаемых в рамках дисциплины.

Оценки по всем формам текущего контроля выставляются по 10-ти балльной шкале. Оценка за контрольную работу выводится по формуле: количество правильных ответов (из 20) * 0,5. При оценивании выполненного домашнего задания учитывается правильность применённого алгоритма, его быстродействие, а также удобочитаемость кода.

Несколько студентов освобождаются от экзамена "автоматом" при условии 80% посещения лекций и отличной работы на семинарах. "Автомат" подразумевает оценку не ниже 8 баллов.

6.2 Порядок формирования оценок по дисциплине

Преподаватель оценивает работу студентов на практических занятиях: результаты проверочных работ, активность в дискуссиях, ответы на вопросы преподавателя, выполнение мини-проектов и командных заданий. Оценки за работу на практических занятиях преподаватель выставляет в рабочую ведомость. Накопленная оценка по 10-ти балльной шкале (среднее арифметическое) за работу на практических занятиях определяется перед итоговым контролем.

Результирующая оценка за дисциплину рассчитывается следующим образом:

$$О_{результ} = 0,5 * О_{накопл} + 0,5 * О_{экз},$$

$$где\ О_{накопл.} = 0,7 * О_{текущий} + 0,3 * О_{аудит}$$

$$О_{текущий} = 0,5 * О_{к/р} + 0,5 * О_{д/з}$$

Способ округления – арифметический.

На пересдаче студенту не предоставляется возможность получить дополнительный балл для компенсации оценки за текущий контроль.

7 Содержание дисциплины

В содержание дисциплины «Автоматическая обработка естественного языка» входит изучение следующих вопросов: структурные особенности текстов на естественном языке; неоднозначность на всех уровнях языка; основные задачи автоматического анализа текстов: классификация текстов, анализ тональности, информационный поиск, извлечение знаний, автоматическое рефери-

рование, машинный перевод; state-of-the-art (новейшие достижения) автоматической обработки текстов; основные подходы к решению задач: правила, написанные вручную и машинное обучение; показатели качества: точность, полнота, F-мера; регулярные выражения; языковые модели; предобработка текста: токенизация и сегментация; нормализация слов: стеммеры, лемматизаторы, морфологические анализаторы; морфологические тэггеры; синтаксические парсеры; извлечение именованных сущностей; векторные модели текстов; мера tf.idf; лексические базы данных (на примере WordNet); семантическая близость слов, текстов; библиотека NLTK для языка программирования Python.

1. Введение в автоматическую обработку естественного языка
   Структурные особенности текстов на естественном языке; неоднозначность на всех уровнях языка; основные задачи автоматического анализа текстов; основные подходы к решению задач: правила, написанные вручную и машинное обучение; показатели качества: точность, полнота, F-мера; state-of-the-art.

2. Базовая обработка текста и дистанция редактирования
   Предобработка текста: токенизация и сегментация; нормализация слов: стеммеры, лемматизаторы, морфологические анализаторы; регулярные выражения; дистанция редактирования.

3. Языковые модели
   N-граммы; перплексия; методы сглаживания; линейная интерполяция; применение языковых моделей: предсказание ввода, исправление ошибок правописания, распознавание речи, порождение текста.

4. Задачи разметки текста и скрытые марковские модели
   Разметка по частям речи; извлечение именованных сущностей как задача разметки; скрытые марковские модели, их достоинства и недостатки; модификации скрытых марковских моделей.

5. Классификация текстов и анализ тональности
   Задачи классификации; наивный байесовский классификатор; проблемы классификации текстов; анализ тональности; извлечение аспектов; соревнования по классификации текстов и анализу тональности.

6. Информационный поиск
   Векторные модели текстов; матричное представление; обратный индекс; фразовые запросы; ранжированный информационный поиск; коэффициент Жаккара; tf-idf; методы оценки поисковых машин.

7. Парсинг
   Синтаксис составляющих и синтаксис зависимостей; контекстонезависимые грамматики; вероятностный подход к парсингу; лексикализованные вероятностные грамматики; алгоритм CKY; применение парсинга в различных задачах.

8. Машинный перевод
   Классические подходы: пословный, трансферный, интерлингвальный; статистический машинный перевод; модели IBM; выравнивание текстов; оценка параметров в моделях IBM; фразовые модели; извлечение фразовых лексиконов; алгоритм декодирования.

9. Компьютерная семантика
   Значение и смысл; WordNet и аналогичные лексические базы данных; измерение семантической близости; тезаурусные методы; дистрибуционные (корпусные) методы.

10. Автоматическое реферирование
    Экстрактивное и абстрактивное реферирование; реферирование нескольких документов; реферирование, основанное на запросе; обучение с учителем и без учителя в контексте автоматического реферирования; оценка систем реферирования; ROUGE.

8 Образовательные технологии

Проектная деятельность, практическая работа с компьютерными инструментами, компьютерные симуляции, мастер-классы экспертов.

8.1 Методические указания студентам

Самостоятельная работа студентов осуществляется в соответствии с «Методическими рекомендациями по организации самостоятельной работы студентов НИУ ВШЭ – Нижний Новгород», утвержденными УМС от 30.04.2014, протокол № 4.

9 Оценочные средства для текущего контроля и аттестации студента

9.1 Тематика заданий текущего контроля

1) Структурные особенности текстов на естественном языке;
2) неоднозначность на всех уровнях языка;
3) основные задачи автоматического анализа текстов;
4) основные подходы к решению задач: правила, написанные вручную и машинное обучение;
5) точность;
6) полнота;
7) F-мера;
8) state-of-the-art;
9) предобработка текста;
10) токенизация;
11) сегментация;
12) нормализация слов;
13) стеммеры;
14) лемматизаторы;
15) морфологические анализаторы;
16) регулярные выражения;
17) дистанция редактирования;
18) N-граммы;
19) перплексия;
20) методы сглаживания;
21) линейная интерполяция;
22) применение языковых моделей;
23) разметка по частям речи;

24) извлечение именованных сущностей как задача разметки;
25) скрытые марковские модели, их достоинства и недостатки;
26) модификации скрытых марковских моделей;
27) задача классификации текстов;
28) наивный байесовский классификатор;
29) проблемы классификации текстов;
30) анализ тональности;
31) извлечение аспектов;
32) соревнования по классификации текстов и анализу тональности;
33) векторные модели текстов;
34) матричное представление (термины-документы);
35) обратный индекс;
36) фразовые запросы;
37) ранжированный информационный поиск;
38) коэффициент Жаккара;
39) tf-idf;
40) методы оценки поисковых машин;
41) синтаксис составляющих и синтаксис зависимостей;
42) контекстонезависимые грамматики;
43) вероятностный подход к парсингу;
44) лексикализованные вероятностные грамматики;
45) алгоритм CKY;
46) применение парсинга;
47) пословный МП;
48) трансферный МП;
49) интерлингвальный МП;
50) статистический МП;
51) модели IBM для МП;
52) выравнивание текстов;
53) оценка параметров в моделях IBM;
54) фразовые модели МП;
55) извлечение фразовых лексиконов для МП;
56) алгоритм декодирования в МП;
57) значение и смысл;
58) WordNet и аналогичные лексические базы данных;
59) измерение семантической близости;
60) тезаурусные методы;
61) дистрибуционные (корпусные) методы;
62) экстрактивное и абстрактивное реферирование;
63) реферирование нескольких документов;
64) реферирование, основанное на запросе;
65) обучение с учителем и без учителя в контексте автоматического реферирования;
66) оценка систем реферирования;
67) ROUGE.

9.2 Пример домашнего задания.

Напишите простую реализацию языковых моделей:

```
def tokenize(text):
      """Возвращает список токенов; игнорирует знаки препинания"""
```

```
	def get_ngram_counts(words, n):
		"""Пройдя по words (списку токенов без знаков препинания), функция возвращает словарь вида {ngram: count} для всех n-граммов, а также n-граммов низшего порядка.
		Например, get_ngram_counts(words, 3) возвращает словарь частотности всех триграммов, биграммов и юниграммов, считая их по списку words."""

	def get_prob(text, ngrams, n, corp_size):
		"""Возвращает вероятность строки text, основываясь на словаре ngrams и оценивая параметры модели с помощью максимального правдоподобия. n - длина n-граммов в модели (напр. 1, 2 или 3), corp_size - количество токенов в корпусе (для оценки параметров юниграмм-модели)."""
```

Пример работы (на основе одного из слайдов лекции):

```
	corpus = 'I saw a cat and a dog. The cat was sleeping, and the dog was awake. I woke up the cat.'
	words = tokenize(corpus)
	corp_size = len(words)
	ngrams = get_ngram_counts(words, 2)
	tests = ('a cat', 'the cat', 'the dog', 'the woke', 'the cat was awake')
	for t in tests:
		print(t, get_prob(t, ngrams, 2, corp_size))
```

Ожидаемый вывод:

```
	a cat 0.5
	the cat 0.6666666666666666
	the dog 0.3333333333333333
	the woke 0.0
	the cat was awake 0.1111111111111111
```

9.3 Вопросы для оценки качества освоения дисциплины

1. Автоматическая обработка естественного языка в кругу смежных дисциплин.
2. Особенности естественного языка и возможности его автоматической обработки.
3. Основные задачи автоматического анализа текстов и подходы к их решению.
4. Оценка систем автоматической обработки текстов.
5. Предобработка текста. Регулярные выражения.
6. Стеммеры, лемматизаторы, морфологические анализаторы.
7. Дистанция редактирования и ее применение в задачах АОТ.
8. Языковые модели: мотивация и применение. N-граммы.
9. Проблемы языковых моделей и способы их решения. Методы оценки языковых моделей.
10. Задачи разметки текста, применение разметки.
11. Скрытые марковские модели, их достоинства и недостатки. Модификации ММ.
12. Классификация текстов: формулировка задачи и методы решения.
13. Наивный байесовский классификатор. Проблемы классификации текстов.
14. Анализ тональности, извлечение аспектов. Соревнования по классификации текстов и анализу тональности.
15. Информационный поиск и векторные модели текстов.
16. Фразовые запросы в информационном поиске; ранжированный информационный поиск.
17. Коэффициент Жаккара; метрика tf-idf; методы оценки поисковых машин.
18. Задача парсинга, его применение. Синтаксис составляющих и синтаксис зависимостей. Контекстонезависимые грамматики.

19. Вероятностный подход к парсингу. Лексикализованные вероятностные грамматики.
20. Машинный перевод. Классические подходы к МП: пословный, трансферный, интерлингвальный.
21. Статистический машинный перевод; модели IBM.
22. Фразовые модели МП. Извлечение фразовых лексиконов. Алгоритм декодирования.
23. Компьютерная семантика. Значение и смысл слов. WordNet и аналогичные лексические базы данных.
24. Измерение семантической близости: тезаурусные и дистрибуционные (корпусные) методы.
25. Автоматическое реферирование. Виды автоматического реферирования.
26. Обучение с учителем и без учителя в контексте автоматического реферирования; оценка систем реферирования; ROUGE.

10 Учебно-методическое и информационное обеспечение дисциплины

10.1 Основная литература

1. Большакова, Е.И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика [Электронный ресурс]: учеб. пособие / Большакова, Е.И., Клышинский, Э.С., Ландэ, Д.В., Носков, А.А., Пескова, О.В., Ягунова, Е.В. - М.: МИЭМ, 2011. - 272 с - ISBN 978–5–94506–294–8. – Режим доступа: http://www.hse.ru/data/2012/04/05/1251263483/пособие%20школа%20по%20компьютерной%20лингвистике%20-%20копия.pdf. – Загл. с экрана.

10.2 Дополнительная литература

1. ACL Anthology: A Digital Archive of Research Papers in Computational Linguistics [Электронный ресурс]. – Режим доступа: http://www.aclweb.org/anthology/. Загл. с экрана.
2. Bird, S., Natural language processing with Python [Электронный ресурс] / Bird, S., Klein, E., Loper, E. – O'Reilly Media, Inc., 2009. Режим доступа: http://www.nltk.org/book/. Загл. с экрана
3. Computational Linguistics [Электронный ресурс]. – MIT Press. Режим доступа: http://www.mitpressjournals.org/loi/coli. Загл. с экрана

10.3 Литература для самостоятельного изучения студентами

1. Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”.
2. Jurafsky D., Martin J. Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing. – Prentice Hall, 2008.
3. Manning C. D., Schütze H. Foundations of statistical natural language processing. – MIT press, 1999.
4. NLPub — каталог лингвистических ресурсов для обработки русского языка. https://nlpub.ru
5. Автоматическая Обработка Текста. http://www.aot.ru/
6. Болховитянов А.В., Гусев А.В., Чеповский А.М. Морфологические модели компьютерной лингвистики: учеб. пособие – М. МГУП, 2010.
7. Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.
8. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во Московского университета, 2011.

9. Марчук Ю. Н. Компьютерная лингвистика: учебное пособие. – М.: АСТ. – 2007.
10. Мельчук И.А. Язык: от смысла к тексту – М.: Языки славянской культуры, 2012.
11. Пиотровский Р.Г. , Бектаев К.Б., Пиотровская А.А. Математическая лингвистика. – М.: Высшая школа, 1977.
12. Чатуев М.Б., Чеповский А.М. Частотные методы в компьютерной лингвистике: учеб. пособие – М. МГУП, 2011.

11 Материально-техническое обеспечение дисциплины

Для лекций и практических занятий используется компьютер/ноутбук, проектор, экран. Для практических занятий студентам необходимо иметь цифровые устройства (ноутбуки, планшеты) с доступом в Интернет. Возможно использование стационарных компьютеров, подключенных к Интернету, в компьютерном классе.

Разработчик Малафеев А.Ю.