Санкт-Петербургский филиал федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет "Высшая школа экономики"»

Факультет Санкт-Петербургская школа экономики и менеджмента
Национального исследовательского университета
«Высшая школа экономики»

Департамент прикладной математики и бизнес-информатики

Рабочая программа дисциплины
«Информационный поиск и обработка текстов на естественном языке»

для образовательной программы «Анализ больших данных в бизнесе, экономике и обществе»
направления подготовки 01.04.02«Прикладная математика и информатика»
уровень магистратура

Разработчики программы
Сироткин А.В., avsirotkin@hse.ru
Суворова А.В., asuvorova@hse.ru

Утверждена Академическим руководителем образовательной программы
«31» августа 2017г.

А.В. Сироткин _____

Санкт-Петербург, 2017

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

1 Область применения и нормативные ссылки

Настоящая рабочая программа дисциплины устанавливает минимальные требования к образовательным результатам, а также определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих дисциплину “Информационный поиск и обработка текстов на естественном языке”, учебных ассистентов и студентов направления подготовки 01.04.02 «Прикладная математика и информатика», обучающихся по образовательной программе «Анализ больших данных в бизнесе, экономике и обществе». Рабочая программа дисциплины разработана в соответствии с:

- Образовательным стандартом НИУ ВШЭ, утвержденным ученым советом Национального исследовательского университета «Высшая школа экономики», протокол от протокол от 06.12.2013 г. № 50. https://www.hse.ru/data/2016/11/02/1111123560/01.04.02%20%D0%9F%D1%80%D0%B8%D0%BA%D0%BB%D0%B0%D0%B4%D0%BD%D0%B0%D1%8F%20%D0%BC%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0%20%D0%B8%20%D0%B8%D0%BD%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0.pdf
- Образовательной программой «Анализ больших данных в бизнесе, экономике и обществе», направление подготовки 01.04.02 «Прикладная математика и информатика»;
- Объединенным учебным планом университета по образовательной программе «Анализ больших данных в бизнесе, экономике и обществе».

2 Цели освоения дисциплины

Целью освоения дисциплины «Информационный поиск и обработка текстов на естественном языке» является ознакомление слушателей с методами обработки текста на естественном языке, а также методами обработки слабоструктурированных данных и извлечения информации. Предполагается знакомство с методами извлечения отношений, анализа тональности, аннотирования и кластеризации текстов, а также с существующими программными реализациями этих методов.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент осваивает следующие компетенции:

Компетенция	Код по ОС ВШЭ	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
Способен принимать управленческие решения и готов нести за них ответственность	СК-5	Способен выбирать направление проекта с учетом имеющихся ресурсов, предлагать путь достижения цели и следовать ему.	Групповые выступления на семинарах, подготовка командного проекта	Домашнее задание

		Знает основные системы хранения данных и способен выбирать наилучшие для проекта.		
Способен анализировать, верифицировать, оценивать полноту информации в ходе профессиональной деятельности, при необходимости восполнять и синтезировать недостающую информацию.	СК-6	Способен отбирать источники информации в соответствии с планируемой задачей, оценивать их полноту и необходимость привлечения других источников	Лекции, самостоятельная работа	Домашнее задание, экзамен
Способен коммуницировать со специалистами в области математических моделей и информационных технологий, а также с экспертами из прикладных областей с использованием различных формальных языков и нотаций.	СК-8	Способен сформулировать задачу обработки текста в терминах как предметной области, так и области формальных объектов	Лекции, практические занятия, самостоятельная работа	Контрольная работа, экзамен
Способен создавать междисциплинарные тексты с использованием языка и аппарата прикладной математики и информатики	ПК-11	Умеет писать документацию и готовить отчеты	Практические занятия, самостоятельная работа	Домашнее задание, экзамен
Способен осуществлять целенаправленный многокритериальный поиск информации о новейших научных и технологических достижениях в сети Интернет и других источниках.	ПК-13	Знает основные существующие пакеты и способен находить новые. Умеет производить отбор программных библиотек, предлагающих решения отдельных частей задачи, с целью наилучшего решения с точки зрения, как	Практические занятия	Домашняя работа, контрольная работа, экзамен

		эффективности алгоритмов, так и с точки зрения времени создания программы.		

4 Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к вариативной части цикла дисциплин магистерской программы.

Изучение данной дисциплины базируется на следующих дисциплинах:

- Современные методы анализа данных
- Современные методы принятия решений

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

- Системный анализ и разработка сложных информационных систем;
- при выполнении проектов, подготовке ВКР

5 Тематический план учебной дисциплины

Курс рассчитан на 48 часов аудиторной нагрузки, из них 16 часов лекций,16 часов семинаров и 16 часов практических занятий, общим объемом 4 зачетные единицы (152 часа).

№	Название раздела	Всего часов	Аудиторные часы			Самостоятельная работа
			Лекции	Семинары	Практические занятия	
1	Введение в обработку естественного языка	44	4	4	4	20
2	Классификация и кластеризация текстов	32	2	2	2	14
3	Информационный поиск	32	2	2	2	14
4	Введение в машинный перевод	32	2	2	2	14
5	Введение в извлечение информации	32	2	2	2	14
6	Методы машинного обучения в задаче извлечения информации	28	2	2	2	14
7	Извлечение мнений	28	2	2	2	14
	Итого:	228	16	16	16	104

6 Содержание дисциплины

1. Введение в обработку естественного языка

Этапы анализа текста. Обзор основных приложений автоматического анализа текста (АОТ) (машинный перевод, информационный поиск, и т.д.). Слова, фразы, предложения, корпусы. Языковые модели. Автоматический морфологический анализ и синтез. Виды морфологического анализа: стемминг, лемматизация, полный морфоанализ. Принципы морфоанализа на базе словаря основ или словаря словоформ. Морфологические процессоры для русского языка

2. Классификация и кластеризация текстов.

Классификация текстов как типичная задача обработки текстов в области TextMining. Обзор методов машинной классификации. Выбор признаков и метрик. Особенности кластеризации текстов. Рубрицирование текстовых документов. Обзор задач АОТ, решаемых на основе классификации текстов. Модели и методы автоматической классификации и кластеризации текстовой информации. Иерархические и вероятностные подходы. Интеллектуальный анализ данных

3. Информационный поиск.

Индексирование текстов для информационного поиска. Векторная модель документа. Булевский поиск, ранжированный поиск. Оценка релевантности документа. Поиск в сети Интернет, принципы работы поисковых машин. Автоматическое реферирование и аннотирование документов как смежные задачи информационного поиска. Основные стратегии сжатия текста. Типы аннотаций. Обзорное реферирование. Оценка качества аннотаций

4. Введение в машинный перевод

Стратегии машинного перевода, основанного на лингвистических правилах. Статистический машинный перевод: особенности и виды. Принципы создания статистического переводчика.

5. Введение в извлечение информации

Основные способы представления смысла текста и модели представления знаний в искусственном интеллекте: семантические сети, язык предикатов.Семантический анализ текста на основе семантико-синтаксических моделей управления. Разметка частей речи. Выделение именованных сущностей. Извлечение информации и отношений из текста. Извлечение информации и знаний из текстов: особенности задачи и типы извлекаемых объектов. Понятие лингвистического шаблона для извлечения информации. Инструментальные программные средства для построения систем извлечения информации из текстов. Извлечение знаний под управлением онтологий в системах класса OntosMiner.

6. Методы машинного обучения в задаче извлечения информации.

Формальные методы определения автора текста. Лингвостатистические параметры. Статистические методы атрибуции. Авторский инвариант и лингвистические спектры. Применение методов кластеризации и классификации для установления авторства текстов. Методы обнаружения спама: вероятностные и статистические, байесовский классификатор

7. Извлечение мнений.

Автоматический анализ тональности текстов и извлечение мнений из текстов: особенности и подходы к решению. Анализ тональности как задача классификации

7 Оценочные средства

7.1 Формы контроля знаний студентов

Тип контроля	Форма контроля	Модуль	Параметры
		2	
Текущий	Домашнее задание	*	Подготовка технического задания для проекта
	Контрольная работа	*	Письменная работа, 80 минут.
Итоговый	Экзамен	*	Экзамен в форме защиты проекта

7.2 Критерии и шкалы оценки знаний, примеры заданий

7.2.1 Оценочные средства для оценки качества освоения дисциплины в ходе текущего контроля

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

Домашнее задание

Задание представляет собой подготовку технического задания на выполнение проекта, связанного с обработкой текста. При оценке работы учитывается корректность постановки

проблемы, подробность и детальность описания проекта и требования к разрабатываемой системе.

Критерии оценивания домашнего задания:

Оценка	Критерии выставления оценки
«Отлично» (8-10)	Задание выполнено в полном объеме. Описание проекта, требований к разрабатываемой системе, сроков выполнения подробно и детально. Выбор методов и использованные оценки обоснованы в полном объеме. Предоставлено техническое задание в письменной форме
«Хорошо» (6-7)	Задание выполнено в полном объеме. Описание проекта, требований к разрабатываемой системе, сроков выполнения подробно и детально. Выбор методов и использованные оценки частично обоснованы. Предоставлено техническое задание в письменной форме.. Имеются замечания / неточности.
«Удовлетворительно» (4-5)	Задание выполнено частично. Описание проекта, требований к разрабатываемой системе, сроков выполнения требует пересмотра. Выбор методов и использованные оценки обоснованы частично. Предоставлено техническое задание в письменной форме
«Неудовлетворительно» (0-3)	Задание выполнено частично. Описание проекта, требований к разрабатываемой системе, сроков выполнения требует значительного пересмотра либо отсутствует часть разделов (не указаны сроки / требования). Нет обоснования выбора методов или оценок. Или не представлено техническое задание в письменной форме

Примеры домашних заданий

- подготовить техническое задание на программу, собирающую отзывы и упоминания заданного объекта/компании из социальных медиа и сайтов отзывов, проводящую анализ тональностей и выделяющих положительные и отрицательные отзывы;
- подготовить техническое задание на программу, собирающую описания вакансий в определенной области с сайтов размещения вакансий, проводящую анализ требуемых навыков в этой области

Контрольная работа

Контрольная работа представляет собой тест с закрытыми и открытыми вопросами (теоретическими и практическими). Тест может содержать от 15 до 25 заданий, покрывающих рассмотренные на занятиях темы. Оценка определяется подсчетом выполненных заданий. Способ округления арифметический.

Критерии оценивания контрольной работы

Оценка		Критерии
«Отлично»	10	Процент выполненных заданий 95-100%
	9	Процент выполненных заданий 85-94%

	8	Процент выполненных заданий 75-84%
«Хорошо»	7	Процент выполненных заданий 65-74%
	6	Процент выполненных заданий 55-64%
«Удовлетворительно»	5	Процент выполненных заданий 45-54%
	4	Процент выполненных заданий 35-44%
«Неудовлетворительно»	3	Процент выполненных заданий 25-34%
	2	Процент выполненных заданий 15-24%
	1	Процент выполненных заданий 5-14%
	0	Процент выполненных заданий 0-4%

Примеры заданий контрольной работы

1. Составить регулярное выражение, удовлетворяющее заданным требованиям.
2. Построить наиболее вероятную цепочку тегов (скрытых состояний) в заданной скрытой марковской модели по указанному предложению.
3. Вывести формулу для коэффициентов заданного алгоритма сглаживания n-граммной языковой модели.
4. Построить символьную триграммную языковую модель по заданному корпусу и с ее помощью построить распознаватель языка документа.
5. Вычислить перплексию n-граммной языковой модели с заданным сглаживанием.
6. На основе заданной обучающей выборки построить марковскую модель максимальной энтропии для выделения заданных именованных сущностей (имен собственных, географических названий и т. д.) из текста.

7.2.2 Итоговый контроль по дисциплине

Итоговый экзамен представляет собой защиту творческого проекта. Творческий проект является реализацией того проекта, для которого в рамках домашнего задания составлялось техническое задание

Критерии оценивания экзамена

Оценка	Критерии выставления оценки
«Отлично» (8-10)	Проект выполнен в полном объеме. Идеи проекта оригинальны и проработаны по всем блокам тем, которые включены в тематический план курса. Выбор методов и инструментов обоснован в полном объеме. Презентация выполнена и представлена на итоговом занятии. Студент презентовал проект и ответил на все дополнительные вопросы. Предоставлен письменный отчет по

	проекту
«Хорошо» (6-7)	Проект выполнен в полном объеме. Идеи проекта оригинальны и частично проработаны по всем блокам тем, которые включены в тематический план курса (допускается проработка проекта на 80% от требуемого объема). Выбор методов и инструментов обоснован в полном объеме. Имеются замечания / неточности. Презентация выполнена и представлена на итоговом занятии. Студент презентовал проект и ответил на все дополнительные вопросы. Предоставлен письменный отчет по проекту
«Удовлетворительно» (4-5)	Проект выполнен частично. Идеи проекта оригинальны и частично проработаны по блокам тем, которые включены в тематический план курса (допускается проработка проекта на 60% от требуемого объема). Есть замечания по обоснованию применения конкретных методов. Презентация выполнена и представлена на итоговом занятии. Студент презентовал проект и ответил на все дополнительные вопросы. Предоставлен письменный отчет по проекту
«Неудовлетворительно» (0-3)	Проект выполнен частично. Идеи частично проработаны по блокам тем, которые включены в тематический план курса (менее 60% от требуемого объема). Нет обоснования выбора методов. Или презентация проекта не сделана и не представлена на итоговом занятии, или не предоставлен письменный отчет по проекту

Примеры заданий итогового проекта

- реализовать программу, собирающую отзывы и упоминания заданного объекта/компании из социальных медиа и сайтов отзывов, проводящую анализ тональностей и выделяющих положительные и отрицательные отзывы;
- реализовать программу, собирающую описания вакансий в определенной области с сайтов размещения вакансий, проводящую анализ требуемых навыков в этой области

7.3 Порядок формирования оценок по дисциплине

Накопленная оценка по дисциплине «Информационный поиск и обработка текстов на естественном языке» рассчитывается следующим образом:

$$\mathbf{О_{накопл.}} = 0{,}5\ Окр + 0{,}5\ Одом.задание$$

где

Окр – оценка за контрольную работу,

Одом.задание - оценка за домашнее задание.

Результирующая оценка по дисциплине «Информационный поиск и обработка текстов на естественном языке» рассчитывается следующим образом:

$$\mathbf{О_{результ.}} = 0{,}4 * О_{накопл.} + 0{,}6 * О_{экз.}$$

При накопленной оценке более 8 баллов по желанию студента она может быть засчитана в качестве результирующей. Способ округления экзаменационной и результирующей оценок: арифметический.

8 Образовательные технологии

Основными образовательными технологиями являются: работа в группах на семинарских занятиях, проектный метод.

9 Учебно-методическое и информационное обеспечение дисциплины

9.1 Основная литература

1. Aggarwal C. C., Zhai C. X. (ed.). Mining text data. – Springer Science & Business Media, 2012. Режим доступа: https://proxylibrary.hse.ru:2258/toc.aspx?bookid=54151
2. Ceri S. et al. Web Information Retrieval. Springer, 2013. 287 p. Режим доступа: https://proxylibrary.hse.ru:2258/toc.aspx?bookid=77020

9.2 Дополнительная литература

1. Munzert S., Rubba C., Meißner P., Nyhuis D. Automated data collection with R: A practical guide to web scraping and text mining. – John Wiley & Sons, 2014. Режим доступа: https://proxylibrary.hse.ru:2258/toc.aspx?bookid=72676
2. Goker A., Davies J. (ed.). Information retrieval: searching in the 21st century. – John Wiley & Sons, 2009. Режим доступа: https://proxylibrary.hse.ru:2258/toc.aspx?bookid=33746

9.3 Ресурсы информационно-телекоммуникационной сети «Интернет»

Для извлечения информации используются следующие сайты:
wikipedia.org – онлайн энциклопедия
twitter.com – сервис блогов
vk.com – социальная сеть с богатым API для доступа к информации
www.tripadvisor.ru – сайт отзывов

10 Рекомендации для самостоятельной работы студентов

Самостоятельная работа может рассматриваться как организационная форма обучения – система педагогических условий, обеспечивающих управление учебной деятельностью по освоению знаний и умений в области учебной деятельности без посторонней помощи. Студенту нужно четко понимать, что самостоятельная работа – не просто обязательное, а необходимое условие для получения знаний по дисциплине и развитию компетенций, необходимых в будущей профессиональной деятельности.
Самостоятельная работа проводится с целью:

- систематизации и закрепления полученных на лекциях теоретических знаний;
- углубления и расширения теоретических знаний;
- формирования умений использовать нормативную, правовую, справочную документацию и специальную литературу;
- развития познавательных способностей и активности студентов: творческой инициативы, самостоятельности, ответственности и организованности;
- формирования самостоятельности мышления, способностей к саморазвитию, самосовершенствованию и самореализации;
- формирования практических (общеучебных и профессиональных) умений и навыков;
- развития исследовательских умений;
- получения навыков эффективной самостоятельной профессиональной (практической и научно-теоретической) деятельности.

11 Материально-техническое обеспечение дисциплины и информационные технологии, используемые при осуществлении образовательного процесса по дисциплине, включая перечень программного обеспечения информационных справочных систем (при необходимости).

Практические занятия проводятся в компьютерных классах. На лекциях и практических занятиях используется проектор. Для успешного освоения дисциплины, студент использует следующие программные средства: Python (пакеты scipy и numpy, сборка Anaconda, Pandas, Scikit-learn и др.), инструменты R и RStudio.

12 Особенности организации обучения для лиц с ограниченными возможностями здоровья

В случае необходимости, обучающимся из числа лиц с ограниченными возможностями здоровья (по заявлению обучающегося) могут предлагаться следующих варианты восприятия учебной информации с учетом их индивидуальных психофизических особенностей, в том числе с применением электронного обучения и дистанционных технологий:

1) для лиц с нарушениями зрения: в печатной форме увеличенным шрифтом; в форме электронного документа; в форме аудиофайла (перевод учебных материалов в аудиоформат); индивидуальные консультации с привлечением тифлосурдопереводчика; индивидуальные задания и консультации.

2) для лиц с нарушениями слуха: в печатной форме; в форме электронного документа; видеоматериалы с субтитрами; индивидуальные консультации с привлечением сурдопереводчика; индивидуальные задания и консультации.

3) для лиц с нарушениями опорно-двигательного аппарата: в печатной форме; в форме электронного документа; в форме аудиофайла; индивидуальные задания и консультации.