**Майнор «Интеллектуальный анализ данных»**

**Программа учебной дисциплины «Прикладные задачи анализа данных»**

Утверждена

Проректором Рощиным С.Ю.

от «10» февраля 2016 г.

|  |  |
| --- | --- |
| Автор  | Игнатов Д.И., к.т.н., доцент, dignatov@hse.ruЧерняк Е.Л., старший преподаватель, echernyak@hse.ruСоколов Е.А., старший преподаватель, esokolov@hse.ru |
| Число кредитов  | 5 |
| Контактная работа (час.)  | 76 |
| Самостоятельная работа (час.)  | 114 |
| Курс  | 3 |
| Формат изучения дисциплины | Без использования онлайн курса |

1. **ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ**

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, выбравших майнор «Интеллектуальный анализ данных ».

Программа разработана в соответствии с образовательным стандартом Федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики» по направлению подготовки 01.03.02 Прикладная математика и информатика для квалификации «бакалавр».

**Целями** освоения дисциплины «Прикладные задачи анализа данных» являются овладение студентами моделями и методами интеллектуального анализа данных и машинного обучения в задачах поиска информации (Information Retrieval), обработки и анализа данных, в частности, обнаружения знаний и поиска закономерностей в данных (knowledge Discovery in Databases and Data Mining), а также приобретение навыков исследователя данных (data scientist) и разработчика математических моделей, методов и алгоритмов анализа данных.

В результате освоения дисциплины студент должен:

● **Знать** основные модели и методы машинного обучения и разработки

данных

● **Уметь** адекватно применять указанные модели и методы, а также

программные средства, в которых они реализованы

● **Владеть** навыками (приобрести опыт) анализа реальных данных с помощью

изученных методов

В результате освоения дисциплины студент осваивает следующие компетенции:

Для специализаций подготовки бакалавров настоящая дисциплина является неотъемлемой дисциплиной в рамках майнора «Интеллектуальный анализ данных», который предоставляет дополнительную специализацию.

Изучение данной дисциплины базируется на следующих дисциплинах:

● Введение в программирование

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

● Необходимо владеть основами программирования на языке Python (изучается в

рамках первой дисциплины курса «Введение в программирование») и знаниями

математики в объеме программы средней школы.

● полученными во второй дисциплины курса «Введение в машинное обучение»

● полученными в третьей дисциплины курса «Интеллектуальные системы»

Основные положения дисциплины могут быть использованы в дальнейшем при изучении следующих дисциплин: введение в машинное обучение и майнинг данных, анализ больших данных, глубинное обучение, статистический анализ данных.

1. **СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ**

**Тема 1. Введение в автоматическую обработку текстов. Морфологический анализ.**

Методы выделения ключевых слов и словосочетаний. Синтаксический анализ. Краткий экскурс в историю автоматической обработки текстов (АОТ). Основные исследовательские и прикладные задачи. 7-уровневая модель языка. Основные этапы обработки текстов: графематический, морфологический и синтаксический анализ. Токенизация текста по правилам. Задачи морфологического анализа: лемматизация и определение частеречных характеристик. Задача выделения ключевых слов и словосочетаний. Меры связности. Меры контрастности. Морфологические шаблоны. Синтаксический анализ. Грамматика составляющих. Универсальные зависимости.

***Литература для освоения темы***

● Jurafsky Daniel, Martin James H. Speech and language processing an introduction to natural language processing, computational linguistics, and speech, 2000, Главы 10, 12, 14

**Тема 2. Векторная модель документа. Дистрибутивная семантика. Методы снижения**

**размерности.**

Векторная модель документа и ее использование в задачах информационного поиска. Модель “мешок слов”. Вычисление косинусной близости между векторами. TF-IDF преобразование. Геометрическая интерпретация косинусной меры сходства. Простейший поиск по запросу. Векторная модель семантики. Дистрибутивная семантика и ее приложения. Понятие “контекста”. Снижение размерности. Сингулярное разложение. Латентно-семантический анализ (ЛСА). Использование ЛСА в задаче поиска по запросу. Использование ЛСА для визуализации дистрибутивной модели.

***Литература для освоения темы***

● Jurafsky Daniel, Martin James H. Speech and language processing an introduction to natural language processing, computational linguistics, and speech, 2000, Главы 15, 16

**Тема 3. Задачи классификации в области автоматической обработки текстов.** Классификация документов. Классификация последовательностей. Задачи классификации, которые возникают в области АОТ: классификация документов и классификация последовательностей. Частеречный разбор как задача классификации последовательностей. Общая формулировка задачи извлечения именованных сущностей. IOB-разметка. Общие подходы к задаче классификации документов: напоминание. Метод наивного Байеса. Логистическая регрессия. Генеративные и дискриминативные методы классификации. Общие подходы к задаче классификации последовательностей. Скрытые цепи Маркова. Условные случайные поля.

***Литература для освоения темы***

● Jurafsky Daniel, Martin James H. Speech and language processing an introduction to natural language processing, computational linguistics, and speech, 2000, Главы 6, 7, 9

***Тема 4. Глубинное обучение и автоматическая обработка текстов. Сверточные и***

***рекуррентные нейронные сети.***

Почему так сложно формулировать лингвистические признаки. Почему глубинное обучение помогает находить скрытые закономерности данных. Обзор задач автоматической обработки текстов, решаемых с помощью глубинного обучения. Современные показатели качества моделей глубинного обучения. Модель однослойного перцептрона Маккалока-Питтса. Модель многоуровневого перцептрона. Алгоритм обратного распространения ошибки и градиентный спуск. Классификация документов и проблема нефиксированной длины входа. Сверточная нейронная сеть. Фильтр. Макспулинг. Вероятностные языковые модели. Рекуррентная нейронная сеть. Проблема исчезающих градиентов. Модели с длинной долгосрочной памятью. Фреймворки глубинного обучения: Theano, Tensorflow, Keras.

***Литература для освоения темы***

● Ian Goodfellow and Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press, 2016, Часть 2, главы 6-12

***Тема 5. Анализ частых множеств признаков и ассоциативных правил (повторение и дополнение)***

Задача анализа потребительской корзины. Поддержка и достоверность. Частые, замкнутые и максимальные частые множества. Алгоритмы Априори и FP-growth. Импликации и ассоциативные правила как объектно-признаковые зависимости. Меры “интересности правил”. Прикладные задачи: анализ посещаемости сайтов и рекомендация контекстной рекламы. Программные средства: Orange, SPMF, ConExp.

***Литература для освоения темы***

● Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014 (<http://www.dataminingbook.info/>), главы 8, 9, 12.

● Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman: Mining of Massive Datasets, 2nd

Ed. Cambridge University Press 2014 (http://www.mmds.org), глава 6.

**Тема 5. Анализ последовательностей (sequence mining).**

Задача поиска частых последовательностей как развитие идеи поиска частых множеств. Форматы данных и алгоритмы (SPADE, PrefixSpan). Реализация алгоритмов в SPMF. Понятие о “контрастных” закономерностях (emerging patterns) в задачах классификации. Прикладные задачи: анализ демографических последовательностей, конкурс Recommender Systems Challenge 2015.

***Литература для освоения темы***

● Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014 (<http://www.dataminingbook.info/>), глава 10.

**Тема 6. Анализ формальных понятий (АФП) и его приложения**

Базовые понятия АФП: контекст, операторы Галуа, формальное понятие, решетка понятий, импликации. Связь импликаций и ассоциативных правил. Связь частых замкнутых множеств и понятий. Диаграмма решетки понятия как средство визуализации и исследования данных. Исследование признаков (Attribute Exploration). Задача построения таксономий предметных областей. Приложения АФП в информационном поиске и анализе текстов. Поиск документов-дубликатов.

***Литература для освоения темы***

● Dmitry I. Ignatov: Introduction to Formal Concept Analysis and Its Applications in Information Retrieval and Related Fields. RuSSIR 2014: 42-141 (https://arxiv.org/pdf/1703.02819.pdf или https://www.researchgate.net/publication/287483929\_Introduction\_to\_Formal\_Concept\_Analysis\_and\_Its\_Applications\_in\_Information\_Retrieval\_and\_Related\_Fields)

**Тема 7. Рекомендательные системы и алгоритмы**

Классификация рекомендательных систем. Контентные рекомендательные системы. Коллаборативная фильтрация: сходство по пользователям (user-based) и сходство по признакам (item-based). Оценка качества рекомендательных систем (Precision, Recall, F1-мера, HitRate, MAE, RMSE, NDCG). Бимодальная кроссвалидация. Прикладная задача: рекомендация фильмов на примере данных проекта MovieLens. Альтернативные подходы оценки качества в бизнес-задачах.

***Литература для освоения темы***

● Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman: Mining of Massive Datasets, 2nd Ed. Cambridge University Press 2014 (http://www.mmds.org), глава 9.

**Тема 8. Бикластеризация. Мультимодальная кластеризация**

Методы бикластеризация. Объектно-признаковая бикластеризация. Кластеризация по подпространству. Системы совместного пользования ресурсами и фолксономии (Folksonomy). Трикластеризация. Прикладные задачи: анализ данных экспрессии генов, социальных сетей и систем совместного пользования ресурсами. Понятие о мультимодальной кластеризации.

**Тема 9. Понижение размерности (SVD, BMF, NMF) и поиск скрытых факторов в рекомендательных системах.**

Разложение матриц на основе сингулярных чисел (Singular Value Decomposition, SVD). Неотрицательная матричная факторизация (Non-negative Matrix Factorisation). Булева матричная факторизация (Boolean Matrix Factorisation). SVD-подобные модели (Funk SVD, Koren's SVD, Matrix Factorization Machines). Связь задачи рекомендаций, регрессии и классификации. Прикладная задача: рекомендация фильмов на основе матричной факторизации. Гибридные рекомендательные системы на примере рекомендации радиостанций.

***Литература для освоения темы***

● Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman: Mining of Massive Datasets, 2nd Ed. Cambridge University Press 2014 (http://www.mmds.org), глава 9, 10.

● Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014 (http://www.dataminingbook.info/),

глава 7.

**Тема 10. Спектральная кластеризация**

Кластеризация графов. Понятие о минимальном разрезе графа (min-cut). Преобразование Лапласа. Коэффициент Релея, вектор и число Фидлера. Иерархическая спектральная кластеризация. Прикладные задачи: поиск сообществ в соцсетях, сегментация рынка контекстной рекламы. Связь спектральной кластеризации с SVD.

*Литература для освоения темы*

● Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman: Mining of Massive Datasets, 2nd Ed. Cambridge University Press 2014 (http://www.mmds.org), глава 10.

● Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014 (<http://www.dataminingbook.info/>), глава 16.

**Тема 11. Анализ связей (алгоритм Page Rank)**

Задача ссылочного ранжирования. Алгоритмы HITS и PageRank.

***Литература для освоения темы***

● Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman: Mining of Massive Datasets, 2nd Ed. Cambridge University Press 2014 (http://www.mmds.org), глава 5.

1. **ОЦЕНИВАНИЕ**

**Критерии оценки знаний, навыков**

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

**Порядок формирования оценок по дисциплине**

ИО =0,5 ДЗ + 0.1 ТЗ + 0.2 ПЗ + 0.2 ПП, где

ДЗ – домашние задания (по усмотрению семинариста оценка по домашнему заданию может быть скорректирована за счет учета работы на семинарах и практических занятиях),

ТЗ – техническое задание к проекту (индивидуальный или групповой проект, допускающий не более 3 участников),

ПЗ – пояснительная записка (письменный отчет по проекту),

ПП – презентация проекта (защита проекта).

Cпособ округления итоговой оценки по дисциплине: арифметический. В случае особых обстоятельств студент может получить возможность пересдать низкие результаты за проектную работу.

На пересдаче студенту не предоставляется возможность получить дополнительный балл для компенсации оценки за домашние задания.

На защите проекта студент может получить дополнительные вопросы (дополнительную практическую задачу, решить к пересдаче домашнее задание), ответ на каждый из который оценивается в 1 балл.

В диплом выставляется результирующая оценка по курсу Майнор “Интеллектуальный анализ данных”, определяемая программой первого года курса.

Дисциплина "Прикладные задачи анализа данных" не дает вклад более 0,25 в результирующую оценку по курсу Майнор “Интеллектуальный анализ данных”.

1. **ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ**

Преподавателю рекомендуется использовать демонстрацию работы изучаемых методов анализа данных с помощью предустановленных программных продуктов (PyData) и других открытых средств во время практических и лекционных занятий. В рамках семинаров возможно решение задач, а домашние задания предлагается составлять практико-ориентированными.

Для проектов по анализу данных рекомендуется предлагать доступные источники данных, например, UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/ или https://www.openml.org). Для проектов по анализу текстов доступны следующие источники: коллекции текстов на русском языке – материалы конкурсов Диалог (http://www.dialog-21.ru/evaluation/), Открытый корпус (opencorpora.org), материалы конкурсов BSNLP (<http://bsnlp-2017.cs.helsinki.fi/shared_task.html>).

# ***Оценочные средства для текущего контроля студента***

Оценочные средства для текущего контроля и аттестации студента могут включать учет работы в рамках семинарских/практических занятий.

Тематика заданий текущего контроля

Домашнее задание 1. Пакет NLTK для первичной обработки текстов.

Домашнее задание 2. Пакет Gensim для снижения размерности в векторной модели.

Домашнее задание 3. Пакет Keras для обучения нейронных сетей.

Домашнее задание 4. Поиск частых множеств и ассоциативных правил.

***Оценочные средства для промежуточной аттестации***

Вопросы для оценки качества освоения дисциплины. Совпадает с примерным перечнем вопросов к экзамену (защите проекта):

1. Назовите две основные задачи морфологического анализа
2. По данному предложению постройте дерево составляющих
3. По данному предложению постройте дерево зависимостей
4. Что оценивает мера связности биграмм? Какие меры связности вы знаете?
5. Зачем нужно tf-idf преобразование?
6. Почему возникает необходимость снижения размерности в векторной модели?
7. Что такое word embedding?
8. Какие виды word embedding’ов вы знаете?
9. Приведите пример задачи классификации документов
10. Приведите пример задачи классификации последовательностей
11. Что общего между методом наивного Байеса и скрытыми Марковскими цепями?
12. Что общего между логистической регрессией и условными случайными полями?
13. Что такое персептрон?
14. В чем заключается метод обратного распространения ошибки?
15. Как использовать сверточные нейронные сети для классификации документов?
16. Как использовать рекуррентные нейронные сети для генерации текстов?
17. **РЕСУРСЫ**
	1. **Основная литература**

Базовые учебники

● Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014 (http://www.dataminingbook.info/pmwiki.php/Main/BookDownload)

● Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman: Mining of Massive Datasets, 2nd Ed. Cambridge University Press 2014 (http://www.mmds.org)

● Jurafsky Daniel, Martin James H. Speech and language processing an introduction to natural language processing, computational linguistics, and speech, 2000. (https://web.stanford.edu/~jurafsky/slp3/)

Основная литература

1. Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014 (http://www.dataminingbook.info/pmwiki.php/Main/BookDownload)

2. Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman: Mining of Massive Datasets, 2nd Ed. Cambridge University Press 2014 (http://www.mmds.org)

3. Manning, Christopher D., and Hinrich Schütze. Foundations of statistical natural language processing. Cambridge: MIT press, 1999 (http://nlp.stanford.edu/fsnlp/)

4. Ian Goodfellow and Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press, 2016 (http://www.deeplearningbook.org)

5. Francesco Ricci, Lior Rokach, Bracha Shapira: Recommender Systems Handbook. Springer 2015 (http://link.springer.com/book/10.1007%2F978-1-4899-7637-6)

* 1. **Дополнительная литература**
1. Òscar Celma: Music Recommendation and Discovery - The Long Tail, Long Fail, and Long Play in the Digital Music Space. Springer 2010 (http://link.springer.com/book/10.1007%2F978-3-642-13287-2)
2. Boyd, Vandenberghe. Convex Optimization (http://stanford.edu/~boyd/cvxbook/)
3. Dekking, F.M., Kraaikamp, C., Lopuhaä, H.P., Meester, L.E., A Modern Introduction to Probability and Statistics (http://www.ewi.tudelft.nl/index.php?id=50508 и http://www.springer.com/gp/book/9781852338961)
4. Boris Mirkin. Core Concepts in Data Analysis: Summarization, Correlation, Visualization. 2010 (<http://www.hse.ru/data/2010/10/14/1223126254/Mirkin_All.pdf>)
	1. **Программное обеспечение**

|  |  |  |
| --- | --- | --- |
| **№ п/п** | **Наименование** | **Условия доступа** |
|   1. |  Microsoft Windows 7 Professional RUSMicrosoft Windows 10Microsoft Windows 8.1 Professional RUS | *Из внутренней сети университета (договор)* |
| 2. | Microsoft Office Professional Plus 2010 | *Из внутренней сети университета (договор)* |
| 3. | Язык программирования Python и его библиотеки NumPy, SciPy, Pandas, Scikit-Learn. | *Свободно-распространяемое (при наличии подключения к сети Интернет)* |
| 4. | Свободно-распространяемые библиотеки анализа данных и машинного обучения, такие как Orange, Weka, SPMF, Concept Explorer и др. | *Свободно-распространяемое (при наличии подключения к сети Интернет)* |

* 1. **Профессиональные базы данных, информационные справочные системы,
	интернет-ресурсы (электронные образовательные ресурсы)**

|  |  |  |
| --- | --- | --- |
| **№ п/п** | **Наименование** | **Условия доступа** |
|  | ***Профессиональные базы данных, информационно-справочные системы*** |
| 1. | Elesevier, Springer, ACM. | *Из внутренней сети университета (договор)* |
|  | ***Интернет-ресурсы (электронные образовательные ресурсы)*** |
| 1. | Портал http://www.machinelearning.ru | *Свободный (при наличии подключения к сети Интернет)* |

* 1. **Материально-техническое обеспечение дисциплины**

Используется проектор (для лекций или семинаров), слайды мультимедийных презентаций и компьютеры с предустановленным программным обеспечением и доступом в Интернет.