Программа учебной дисциплины «Анализ неструктурированной информации»

Утверждена
Академическим советом ООП
Протокол №__ от «__» _____ 2019 г.

Разработчик	Бекларян Армен Левонович
Число кредитов	5
Контактная работа (час.)	64
Самостоятельная работа (час.)	126
Курс	1 курс магистратуры, Образовательная программа «Бизнес-информатика»
Формат изучения дисциплины	без использования онлайн курса

РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целью освоения учебной дисциплины является формирование у студентов комплекса теоретических знаний, методологических основ и практических навыков в области анализа неструктурированной информации. В том числе технологий извлечения знаний из текстовых данных (Text Mining) и технологии, лежащей на пересечении извлечения знаний из баз данных, эффективного поиска информации, искусственного интеллекта, машинного обучения и обработки естественных языков (Web Mining).

Задачи курса:

– Формирование теоретических и методологических основ в области анализа неструктурированной информации, а также практических навыков, использования алгоритмов интеллектуального анализа данных, реализованных в специализированных программных продуктах.

– Формирование теоретических основ и навыков использования парадигмы распределенных вычислений MapReduce и концепции баз данных NoSQL.

– Формирование навыков проведения сравнительного анализа основных моделей, включая методы индукции правил, сети Кохонена и ассоциативные правила.

В результате освоения дисциплины студент должен:

знать:

– характеристики рынка систем анализа неструктурированной информации и перспективы развития сегмента информационно-технологической отрасли «Большие данные» (Big Data), основные методы анализа, применяемые в «Больших данных», а также основные классы и принципы построения информационных систем, применяемых для практической реализации этих методов.

уметь:

– применять для анализа неструктурированной информации эвристические алгоритмы поиска, эволюционное вычисление, генетические алгоритмы, алгоритмы ненаправляемого обучения (Unsupervised Learning).

владеть:

– навыками использования систем анализа неструктурированной информации для решения задач сквозного поиска по источникам, выявления закономерностей на основании анализа текстовых данных, извлечения ключевых факторов из неструктурированных текстов.

Изучение дисциплины «Анализ неструктурированной информации» базируется на следующих дисциплинах:

– Проектирование информационных систем;
– Системный анализ и проектирование;
– Управление данными;
– Теория вероятностей и математическая статистика.

Для освоения учебной дисциплины студенты должны знать концептуальные основы архитектуры предприятия, основные классы информационных систем управления бизнесом, лучшие практики и современные стандарты в сфере информационных технологий.

Также студенты должны владеть методами проектирования информационных систем, уметь систематизировать и обобщать информацию, разрабатывать конкретные предложения по результатам исследований, готовить справочно-аналитические материалы для принятия управленческих решений в сфере информационных технологий.

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

– Системы интеллектуального анализа данных;
– Системы бизнес интеллекта;
– Системы поддержки принятия решений.

I. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Название раздела	Аудиторные часы		Самостоятельная работа	Планируемые результаты обучения (ПРО), подлежащие контролю	Формы контроля
	Лекции	Практические занятия			
Тема 1. Концепция «Больших Данных»	4	2	16	Понимание основных терминов	Практическая работа в компьютерном классе по применению in-memory analytics
Тема 2. Неструктурированная информация	10	8	20	Самостоятельное построение топологии рекуррентной нейронной сети	Практическая работа в компьютерном классе по решению задачи анализа текста с помощью нейронной сети

Тема 3. Аппаратное и программное обеспечение «Больших Данных»	6	8	22	Самостоятельное разворачивание NoSQL хранилища	Практическая работа в компьютерном классе по решению задачи подсчета количества слов на базе Apache Hadoop. Реферат по анализу существующих решений класса Big Data в рамках заданной предметной области
Тема 4. Масштабирование и многоуровневое хранение «Больших Данных»	4	6	18	Самостоятельное разворачивание облачной топологии для решения задачи анализа комментариев в реальном времени	Практическая работа в компьютерном классе по созданию сервиса анализа комментариев на базе IBM Watson Cloud
Тема 5. Практическое применение «Больших Данных»	4	12	50	Самостоятельная постановка задач машинного обучения на базе основных эвристических алгоритмов	Практическая работа в компьютерном классе по решению разнообразных задач машинного обучения. Домашнее задание по комплексному анализу на базе заданного датасета
Итого часов	**28**	**36**	**126**		

Тема 1. Концепция «Больших Данных»

Что такое «Большие данные», и что они нам сулят. Разница между бизнес-аналитикой и «Большими данными». Устаревание информации. Рост объемов данных на фоне вытеснения аналоговых средств хранения. Корректная интерпретация информационных потоков. Обработка информационных потоков. Предпосылки применения контент-анализа в различных исследованиях.

Необходимость в аналитической работе с большими данными. Явная (выраженная) и скрытая (структурная) информация. Количественная и качественная стратегия анализа текстов. Возможности и ограничения каждого из подходов. Процедура контент-анализа. Определение круга проблем для контент-анализа. Начальный этап исследования: формулирование целей и задач исследования, выбор эмпирического материала, выдвижение рабочих гипотез. Операциональный этап исследования: определение категорий и подкатегорий, выбор единиц анализа, установление правил кодирования. Этап счета. Этап интерпретации результатов. Презентация результатов. Типичные ошибки при проведении контент-анализа.

Технические признаки, характеризующие «Большие данные». Принцип V3 – Volume (объём данных), Variety (разнообразие данных) и Velocity (скорость генерации и работы с дан-

ными). Интеграция, миграция и построение хранилищ данных. Высокопроизводительные вычисления (High Performance Computing, HPC) при выполнении аналитических исследований. Grid computing (распределенные вычисления на нескольких серверах), in-database analytics (частичный перевод нагрузки при аналитических вычислениях в СУБД, а также регламентное применение готовых аналитических моделей к новым данным полностью на стороне СУБД) и in-memory analytics (применение аналитики прямо в оперативной памяти сервера СУБД).

Тема 2. Неструктурированная информация

Эвристические алгоритмы поиска, эволюционное вычисление, этапы генетического алгоритма: задание целевой функции (приспособленности) для особей популяции, создание начальной популяции, размножение (скрещивание), мутирование, вычисление значения целевой функции для всех особей, формирование нового поколения (селекция).

Задача кластеризации, методы кластеризации, иерархическая кластеризация, алгоритм k-средних, зонтичная кластеризация, методы ненаправляемого обучения (Unsupervised Learning). Постановка задачи классификации, подходы и применения, построение и обучение классификатора, оценка качества классификации, рубрикации тренировочных данных (Training Data Set), методы управляемого (направляемого) обучения (Supervised Learning).

Методы распознавания образов, дискриминантный анализ, нелинейная оптимизация, этапы формирования нейронных сетей: сбор данных для обучения, подготовка и нормализация данных, выбор топологии сети, экспериментальный подбор характеристик сети, экспериментальный подбор параметров обучения, собственно обучение, проверка адекватности обучения, корректировка параметров, окончательное обучение, вербализация сети с целью дальнейшего использования.

Совместное использование компьютерных технологий и лингвистики для создания алгоритмов, позволяющих анализировать естественные (человеческие) языки. Применение методов обработки естественных языков и других аналитических методов для выявления и извлечения из анализируемого текста субъективной информации, характеризующей настроения, мнения, отношение людей к проблеме. Рассмотрение следующих основных задач: синтез речи, распознавание речи, анализ текста, синтез текста, машинный перевод, вопросно-ответные системы, информационный поиск, извлечение информации, анализ тональности текста, анализ высказываний, упрощение текста.

Тема 3. Аппаратное и программное обеспечение «Больших Данных»

Вычисления некоторых наборов распределенных задач с использованием большого количества компьютеров, образующих кластер. Шаги Map и Reduce. Предварительная обработка входных данных и свёртка данных. Концепция параллелизма. Шаблоны доступа к данным, хеш-таблица, деревья, таксономия NoSQL, колоночные СУБД, bigtable.

Разработка и выполнение распределённых программ, расширение вычислительных мощностей посредством добавления в кластер дополнительных узлов, технология Hadoop, распределённая файловая система HDFS (Hadoop Distributed File System), интеграция с NoSQL и MapReduce.

Тема 4. Масштабирование и многоуровневое хранение «Больших Данных»

Модели развёртывания: частное облако, публичное облако, гибридное облако, общественное облако. Модели обслуживания: программное обеспечение, платформа, инфраструкту-

ра. Экономические аспекты центров обработки данных. Безопасность при хранении и пересылке данных. Проблема «последней мили».

Обработка Fast Data, подтверждение и корректировка априорных знаний и гипотез, синхронизация скорости работы с ростом объема данных. Получение знаний посредством Big Analytics, преобразования зафиксированной в данных информации в новое знание, принцип «обучения с учителем». Высший уровень работы с данными Deep Insight, обучение без учителя (unsupervised learning), использование современных методов аналитики, а также различные способы визуализации, обнаружение знаний и закономерностей, априорно неизвестных.

Тема 5. Практическое применение «Больших Данных»

Практическое применение решений IBM Cognos Analytics и ресурсов платформы IBM Bluemix. Понятие шаблона, создание правил и категорий. Персональная база данных, фразовый поиск, нечеткий поиск. Возможности уточнения результатов запросов с учетом структуры текста. Анализ совместной встречаемости (collocate analysis) и коэффициент связи категорий (Z-score).

Практическое применение решений векторизации текста. Контент-анализ массовой корреспонденции и социологических опросов. Прямые пропорциональные закономерности, аддитивные закономерности, мультипликативные закономерности.

III. ОЦЕНИВАНИЕ

Формами текущего контроля являются реферат и домашнее задание. Каждая из форм текущего контроля оценивается по 10-балльной шкале. Общая оценка за текущий контроль (по 10-балльной шкале) рассчитывается по формуле:

$$O_{текущий} = 0{,}4 \cdot O_{реф} + 0{,}6 \cdot O_{дз},$$

где $O_{реф}$ – оценка за реферат;

$O_{дз}$ – оценка за домашнее задание.

При определении накопленной оценки (по 10-балльной шкале) самостоятельная внеаудиторная работа не оцениваются. Поэтому накопленная оценка формируется из оценки за текущий контроль и оценки за работу на аудиторных занятиях, и рассчитывается по формуле:

$$O_{накопленная} = 0{,}7 \cdot O_{текущий} + 0{,}3 \cdot O_{ауд} + 0{,}0 \cdot O_{сам.работа},$$

где $O_{текущий}$ – оценка за текущий контроль;

$O_{ауд}$ – оценка за аудиторную работу;

$O_{сам.работа}$ – оценка за самостоятельную работу.

Оценка за аудиторную работу выставляется на основе пропорции посещаемости студента к общему числу проведенных занятий, исходя из максимума в 10 баллов.

Результирующая оценка (выставляется в диплом) формируется на основе итоговой оценки за экзамен (по 10-балльной шкале) и накопленной оценки. Результирующая оценка рассчитывается по формуле:

$$O_{результ} = 0{,}3 \cdot O_{экз} + 0{,}7 \cdot O_{накопленная},$$

где $O_{экз}$ – оценка за итоговый контроль (экзамен);

$О_{накопленная}$ – накопленная оценка.

$$О_{экз} = 0{,}5 \cdot О_{экз1} + 0{,}5 \cdot О_{экз2},$$

где $О_{экз1}$ – оценка за тестовую часть экзамена;

$О_{экз2}$ – оценка за письменную часть экзамена.

Оценка за тестовую часть экзамена представляет собой сумму баллов за каждый вопрос, при этом вопросы дают вклад либо 1 балл, в случае полностью верного данного ответа, либо 0 баллов, в противном случае. Для вопросов с множественным выбором правильным считается тот ответ, в рамках которого выбраны все правильные варианты ответа и только они.

Оценивание письменной части экзамена основано на глубине и корректности предложенных шагов анализа.

При формировании результирующей оценки на основе весовых коэффициентов применяется арифметическое округление до целого числа. В случае точного равенства дробной части пяти десятым округление применяется в большую сторону.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

4.1 Содержание заданий контроля

Выполнение домашнего задания предусматривает построение моделей анализа неструктурированной информации, выявление регулярных выражений, построение аналитических срезов и фильтров, выделение корреляций между срезами, отображение взаимосвязей и визуализацию итогов анализа в системах IBM Cognos Analytics или Microsoft Power BI.

Реферат формируется на основе лекционных материалов, отечественных и зарубежных (англоязычных) публикаций по вопросам теории и практики систем анализа неструктурированной информации.

Экзаменационная работа состоит из двух частей: тест и письменная часть. Тест представляет из себя 10 вопросов закрытого типа, письменная часть – описание шагов анализа в рамках заданной предметной области и проблематики.

4.2 Пример тематик реферата

1. Распознавание лица с помощью нейронных сетей.
2. Влияние GDPR на сбор и обработку больших данных.
3. Разработка лекарственных препаратов с помощью искусственного интеллекта.
4. Влияние технологии смарт-контрактов на межорганизационные процессы.

4.3 Пример вопросов тестовой части экзамена

1. В постулаты контент-анализа НЕ входит утверждение, что:

А. Исследовать можно только то, что зафиксировано. То, что не зафиксировано, не существует.

Б. Статистический метод исследования имеет два варианта: интуитивный и формализованный.

В. Исследовать что-либо можно только двумя способами: аналитическим и статистическим.

Г. Формализованный метод имеет только линейное частотное распределение.

2. Большая часть данных генерируется при взаимодействии:

А. Машин между собой.
Б. Окружающего мира и машины.
В. Человека и машины.
Г. Правильный ответ не известен.

4.4 Методические указания студентам по подготовке домашнего задания

В рамках домашнего задания студентам необходимо подготовить аналитическую отчетность по выбранной предметной области с использованием BI инструментов и систем статистического анализа.

1. Источник данных
Источник данных студент выбирает самостоятельно также, как и предметную область. Среди обязательных требований наличие не менее 10000 записей, а также геопривязок данных любого уровня грануляции.

2. Модификация данных и первичный анализ
Модификация исходных данных, а также предварительный анализ в форме отчета верхнего уровня формируется либо в системе Microsoft Power BI, либо IBM Cognos Analytics на выбор студента.

3. Внедрение вычислимых полей
Вычислимые поля могут представлять из себя как промежуточные вычисления в рамках статистического и глубинного анализа (п.4), так и введенные KPI или иные меры на исходных данных.

4. Проведение статистического и глубинного анализа
Статистический и глубинный анализ проводятся с целью выявления скрытых зависимостей или для подтверждения гипотез, сформированных по итогам первичного анализа (п.2), в частности, допустимо проведение регрессионного анализа. Можно использовать следующие системы анализа данных: R, Python, SPSS Statistics, EViews или Stata.

5. Создание итогового видео ролика
Видео ролик создается средствами Power Map и представляет из себя вынесенные на карту статистические данные, а также итоги проведенного анализа.

Отчетные материалы:
1. Исходные данные (csv, xml,json,xlsx и др.)
2. Проект Power BI или Cognos Analytics
3. Скрипты статистического и глубинного анализа
4. Проект Power Map (xlsx)
5. Видео ролик
6. Сопроводительная документация

Итоговые материалы необходимо выложить в облачном сервисе и отправить ссылку преподавателю по электронной почте.

Оценивание:
Выполнение пп.1 и 2 дает возможность получения оценки 4 и является обязательным минимум при выполнении задания.
Выполнение п.3 добавляет максимум 1 балл.
Выполнение п.4 добавляет максимум 3 балла.
Выполнение п.5 добавляет максимум 2 балла.

4.5 Вопросы для оценки качества освоения дисциплины

Вопросы к Теме 1. Концепция «Больших Данных»

1. В чем принципиальное отличие концепции Big Data от традиционного подхода BI?
2. Понятие явной (выраженной) и скрытой (структурной) информации.
3. Определение контент-анализа.
4. Каковы основные понятия контент-анализа?
5. Какие существуют виды контент-анализа?
6. Какие существуют этапы контент-анализа?
7. Каковы основные признаки, характеризующие «Большие данные»?

Вопросы к Теме 2. Неструктурированная информация

1. Сущность и задачи кластеризации.
2. Основные понятия, принципы и предпосылки генетических алгоритмов.
3. Достоинства и недостатки генетических алгоритмов.
4. Классификация нейронных сетей и принципы построения.
5. Искусственная нейронная сеть прямого прохода.
6. Использование генетических алгоритмов для обучения искусственных нейронных сетей
7. Кластеризация как инструмент предварительной обработки данных для искусственной нейронной сети
8. Какова цель синтаксического анализа?
9. Общая схема алгоритма синтаксического анализа «сверху-вниз» и «снизу-вверх».

Вопросы к Теме 3. Аппаратное и программное обеспечение «Больших Данных»

1. Схема работы фаз map(f, c) и reduce(f, c).
2. Преимущества, ограничения и недостатки парадигмы MapReduce.
3. Какие бывают модели данных и запросов в NoSQL?
4. Какие бывают системы хранения данных в NoSQL?
5. Основные принципы работы фреймворка Hadoop.
6. Репликация данных в распределенной файловой системе HDFS.

Вопросы к Теме 4. Масштабирование и многоуровневое хранение «Больших Данных»

1. Модели развертывания облачных хранилищ.
2. Модели обслуживания облачных хранилищ.
3. Постановка и описание проблемы «последней мили».
4. Безопасность, производительность и надежность при работе с облачными данными.
5. Экономическая составляющая облачных подходов.
6. Способы машинного обучения.
7. Основные фазы обработки «больших данных».

Вопросы к Теме 5. Практическое применение «Больших Данных»

1. Чем отличаются текстовая и персональная базы данных?
2. Метод анализа комбинации слов (collocate analysis).
3. Понятие «сила связи».
4. Статистическая мера совместной встречаемости слов и категорий (Z-score).
5. Реализация закономерностей в системе IBM Cognos Analytics.

V. РЕСУРСЫ

5.1 Основная литература

1. Бессмертный, И. А. Системы искусственного интеллекта: учебное пособие для академического бакалавриата / И. А. Бессмертный. – 2-е изд., испр. и доп. – Москва: Издательство Юрайт, 2019. — 157 с. Режим доступа: https://biblio-online.ru/bcode/423120 (дата обращения: 20.06.2019).
2. Маккинли У. Python и анализ данных / У. Маккинли; Пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2015. – 799 с.
3. Миркин, Б. Г. Введение в анализ данных: учебник и практикум / Б. Г. Миркин. – Москва: Издательство Юрайт, 2019. – 174 с. Режим доступа: https://biblio-online.ru/bcode/432851 (дата обращения: 20.06.2019).
4. Анализ данных: учебник для академического бакалавриата / В. С. Мхитарян [и др.]; под редакцией В. С. Мхитаряна. – Москва: Издательство Юрайт, 2019. – 490 с. Режим доступа: https://biblio-online.ru/bcode/432178 (дата обращения: 20.06.2019).
5. Ратникова Т. А. Анализ панельных данных и данных о длительности состояний: учеб. пособие / Т. А. Ратникова, К. К. Фурманов. – М.: Изд. дом Высшей школы экономики, 2014.
6. Федоров, Д. Ю. Программирование на языке высокого уровня python: учебное пособие для прикладного бакалавриата / Д. Ю. Федоров. – 2-е изд., перераб. и доп. – Москва: Издательство Юрайт, 2019. – 161 с. Режим доступа: https://biblio-online.ru/bcode/437489 (дата обращения: 20.06.2019).
7. Черткова, Е. А. Статистика. Автоматизация обработки информации: учебное пособие для вузов / Е. А. Черткова; под общей редакцией Е. А. Чертковой. – 2-е изд., испр. и доп. – Москва: Издательство Юрайт, 2019. – 195 с. Режим доступа: https://biblio-online.ru/bcode/437242 (дата обращения: 20.06.2019).

5.2 Дополнительная литература

1. Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. Введение в статистическое обучение с примерами на языке R. ДМК Пресс, 2016.
2. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. Погружение в мир нейронных сетей. – Питер, 2018. – с. 480.
3. Рашка С. Python и машинное обучение. М.: ДМК Пресс 2017. 418 с.
4. Ричарт В., Коэльо П.Л. Построение систем машинного обучения на языке Python. ДМК Пресс, 2015.
5. Силен Д., Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных. Питер, 2017.
6. Ульман Дж., Раджараман А., Лесковец Ю. Анализ больших наборов данных. ДМК Пресс, 2016.
7. Флах П. Машинное обучение. ДМК Пресс, 2015.
8. Cambria E., Das D., Bandyopadhyay S., Feraco A. A Practical Guide to Sentiment Analysis. – Springer, Cham, 2017.
9. Pozzi F.A., Fersini E., Messina E., Liu B. Sentiment Analysis in Social Networks. – Morgan Kaufmann, 2017. – 284 p.
10. Rafaels R.J. Cloud Computing: From Beginning to End. – CreateSpace Independent Publishing Platform, 2015. – 152 p.
11. Zhang L., Liu B. Sentiment Analysis and Opinion Mining. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA, 2017.

5.3 Справочники, словари, энциклопедии

1. Тезаурус социологии. Книга 2. Методология и методы социологических исследований. Тематический словарь-справочник. Под редакцией: Тощенко Ж.Т. – М.: Юнити-Дана, 2013. – с. 416.
2. Big Data: The Next Frontier for Innovation, Competition, and Productivity. – McKinsey Global Institute, May 2011.
3. Big Data: What It Is and Why You Should Care. White Paper. – IDC, 2011.

5.4 Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Microsoft Windows 7 Professional или более новая версия	Из внутренней сети университета (договор)
2.	Microsoft Office Professional Plus 2013 или более новая версия	Из внутренней сети университета (договор)
3.	Microsoft SQL Server 2014 Enterprise Edition или более новая версия	Из внутренней сети университета (договор)
4.	Microsoft Power BI	Свободно распространяемое ПО
5.	JDK 8	Свободно распространяемое ПО
6.	Notepad++	Свободно распространяемое ПО
7.	R 3.1.2 или более новая версия	Из внутренней сети университета (договор)
8.	RStudio	Из внутренней сети университета (договор)
9.	Anaconda 3 x64	Из внутренней сети университета (договор)
10.	Faronics Insight	Из внутренней сети университета (договор)

5.5 Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
	Профессиональные базы данных, информационно-справочные системы	
1.	Электронно-библиотечная система Юрайт	URL: https://biblio-online.ru/
	Интернет-ресурсы (электронные образовательные ресурсы)	
1.	Веб-сервис для хостинга IT-проектов и их совместной разработки	URL: https://github.com

5.6 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

– ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);

– мультимедийный проектор с дистанционным управлением.

Учебные аудитории для лабораторных и самостоятельных занятий по дисциплине оснащены ПЭВМ, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ, а также с установленным требуемым программным обеспечением, в количестве одна единица на каждого слушателя дисциплины.