УТВЕРЖДАЮ

Ректор ФГБОУ ВО «Башкирский государственный университет»

_____ Н.Д. Морозкин

«__5__» ____12____ 20__20__ г.

Паспорт Образовательной программы

«Интеллектуальный анализ текста на основе машинного обучения»

Версия программы	**2**
Дата Версии	**05.10.2020**

1. Сведения о Провайдере

1.1	Провайдер	ФГБОУ ВО "Башкирский государственный университет"
1.2	Логотип образовательной организации	Загружен на платформу
1.3	Провайдер ИНН	0274011237
1.4	Ответственный за программу ФИО	Лакман Ирина Александровна
1.5	Ответственный должность	Заведующий лабораторией исследования социально-экономических проблем регионов
1.6	Ответственный Телефон	+7-927-9655655
1.7	Ответственный E-mail	Lackmania@mail.ru

2. Основные Данные

№	Название	Описание
2.1	Название программы	Интеллектуальный анализ текста на основе машинного обучения
2.2	Ссылка на страницу программы	http://sdo.bashedu.ru/mod/page/view.php?id=93378
2.3	Формат обучения	Онлайн
4	Подтверждение от ОО наличия возможности реализации образовательной программы с применением	Подтверждено

	электронного обучения и (или) дистанционных образовательных технологий с возможностью передачи данных в форме элементов цифрового следа	
2.4	Уровень сложности	Базовый
2.5	Количество академических часов	72
	Практикоориентированный характер образовательной программы: не менее 50 % трудоёмкости учебной деятельности отведено практическим занятиям и (или) выполнению практических заданий в режиме самостоятельной работы	53 ак.часа или 74% трудоёмкости учебной деятельности отведено практическим занятиям и (или) выполнению практических заданий в режиме самостоятельной работы
2.6	Стоимость обучения одного обучающегося по образовательной программе, а также предоставление ссылок на 3 (три) аналогичные образовательные программы иных организаций, осуществляющих обучение, для оценки объективности стоимости или обоснование уникальности представленной образовательной программы в случае отсутствия аналогичных образовательных программ на рынке образовательных услуг	**26 000 рублей.** Проведённый обзор показал, что в настоящее время на рынке образовательных услуг не представлено курса повышения квалификации по интеллектуальному анализу текста (ИАТ) в объёме на 72 часа. Существующие предложения в формате переподготовки (длительностью 1.5 года), в которых есть разделы по ИАТ. Таким образом, отталкиваясь от сложившейся стоимости всей программы и учитывая объем охватываемого материала, стоимость курса оценивается в 26000 руб.
2.7	Минимальное количество человек на курсе	**15 человек**
2.8	Максимальное количество человек на курсе	**105 человек**
2.9	Данные о количестве слушателей, ранее успешно прошедших обучение по образовательной программе	38
2.10	Формы аттестации	**Зачёт по итоговому комплексному проекту**

	Указание на область реализации компетенций цифровой экономики, к которой в большей степени относится образовательная программа, в соответствии с Перечнем областей	Искусственный интеллект

3. Аннотация программы

Цель курса дать систематизированное представление о современных подходах к интеллектуальному анализу текста средствами машинного обучения, познакомить с основными принципами лингвостатистики, научить навыкам применения технологий интеллектуального анализа текста, относящегося к неструктурированной информации, в современных информационных средах (RStudio), в конечном итоге сформировать на базовом уровне компетенцию компетенции цифровой экономики: Управление информацией и данными. В рамках изучения курса у слушателей будет сформирована компетенция на базовом уровне:

Способность управлять неструктурированной информацией и данными:

В результате изучения дисциплины слушатель должен:

знать: основные метрики лингвостатистики; основные законы лингвостатистики; (Хипса, Ципфа); основные принципы разметки текста; способы векторного представления текста; метрики по реляционному и атрибутивному сходству текста; метрики ассоциации для измерения в коллакациях; способы кластеризации текста; инструмент TF-IDF для анализа главной темы; основные методы латентно-семантического анализа текста; инструменты машинного обучения (наивный байесовский классификатор) для классификации текста; основные метрики оценки качества классификации;

уметь: проводить качественную чистку текста, избавляться от стоп-слов и проводить лемматизацию и стеминг текста; создавать терм-документную матрицу двумя способами; использовать мешочек слов для анализа текста; применять процедуру TF-IDF для анализа главной темы; проводить классификацию текста (например спам/неспам) с помощью наивного байесовского классификатора; определять меру сходства текста и меру ассоциации в коллакациях; применять латентно-семантический анализ текста.

владеть: навыками предподготовки к проведению анализа текста, используя средства среды RStudio; навыками тематического моделирования, используя инструменты алгоритма TF-IDF.

а также иметь опыт применения современных методов и подходов интеллектуального анализа текста на базовом уровне средствами машинного обучения.

Для успешного прохождения курса слушатели должны на продвинутом уровне пользоваться компьютером, иметь зовые навыки в моделирование, знать основы теории вероятности и

математической статистики, иметь представление о программировании на языках высокого уровня (на пороговом уровне). Для слушателей курсов предусмотрены входные контрольные задания по теории вероятности (теорема Байеса) и математической статистике (описательные дескриптивные статистики и проверка гипотез – ошибки первого и второго рода).

Компетенция, сформированная в рамках прохождения курса, позволит развиться в профессиональной деятельности ИТ-специалистам, сменить род деятельности в рамках одной области (Информационные технологии).

СОГЛАСОВАНО

Директор ИНО _____ Т.Б.Великжанина «__5__»__10__2020г

I. ОБРАЗОВАТЕЛЬНАЯ ПРОГРАММА

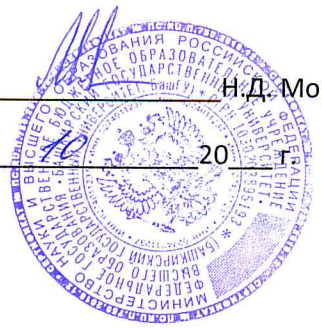
Федеральное государственное бюджетное образовательное учреждение высшего образования
«БАШКИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

УТВЕРЖДАЮ

Ректор

_____Н.Д. Морозкин

«__5__» ____10_____20___г.



Дополнительная профессиональная программа повышения квалификации

«Интеллектуальный анализ текста на основе машинного обучения»

72 час.

СОГЛАСОВАНО

Директор ИНО _____ Т.Б.Великжанина

«__5__» ____10_____2020г

УФА 2020

ОБЩАЯ ХАРАКТЕРИСТИКА ПРОГРАММЫ

1. **Цель программы** дать систематизированное представление о современных подходах к интеллектуальному анализу текста средствами машинного обучения, познакомить с основными принципами лингвостатистики, научить навыкам применения технологий интеллектуального анализа текста, относящегося к неструктурированной информации, в современных информационных средах (RStudio), в конечном итоге сформировать на базовом уровне компетенцию компетенции цифровой экономики: Способность управлять неструктурированной информацией и данными.

   Задачи:

   - дать систематизированное представление о современных подходах к интеллектуальному анализу текста;
   - формирование у слушателей профессиональных компетенций, связанных с использованием теоретических знаний в области интеллектуального анализа текста; сформировать навыки работы в среде RStudio для применения инструментов машинного обучения для интеллектуального анализа текста.

2. **Планируемые результаты обучения:**

2.1. Знание (осведомленность в областях)

2.1.1 основные метрики лингвостатистики;
2.1.2 основные законы лингвостатистики; (Хипса, Ципфа)
2.1.3 основные принципы разметки текста;
2.1.4 способы векторного представления текста;
2.1.5 метрики по реляционному и атрибутивному сходству текста;
2.1.6 метрики ассоциации для измерения в коллакациях.
2.1.7 способы кластеризации текста;
2.1.8 инструмент TF-IDF для анализа главной темы
2.1.9 основные методы латентно-семантического анализа текста.
2.1.10 инструменты машинного обучения (наивный байесовский классификатор) для классификации текста;
2.1.11 основные метрики оценки качества классификации.

2.2 Умение (способность к деятельности):

2.2.1 проводить качественную чистку текста, избавляться от стоп-слов и проводить лемматизацию и стеминг текста;
2.2.2 создавать терм-документную матрицу двумя способами;
2.2.3 использовать мешочек слов для анализа текста;
2.2.4 применять процедуру TF-IDF для анализа главной темы;
2.2.5 проводить классификацию текста (например спам/неспам) с помощью наивного байесовского классификатора;
2.2.6 определять меру сходства текста и меру ассоциации в коллакациях.
2.2.7 применять латентно-семантический анализ текста.

2.3 Навык (использование конкретных инструментов):

2.3.1 навыками предподготовки к проведению анализа текста, используя средства среды RStudio;

2.3.2 навыками тематического моделирования, используя инструменты алгоритма TF-IDF.

3 **Требования к слушателям** (возможно заполнение не всех полей)

3.1 Образование: высшее, среднее профессиональное

3.2 Квалификация: инженер, математик, филолог

3.3 Наличию опыта профессиональной деятельности: работа в Excel.

3.4 Предварительное освоение иных дисциплин/курсов /модулей: высшая математика, теория вероятностей и математическая статистика, общая теория статистики, основы лингвистики

4.Учебный план программы «Интеллектуальный анализ текста на основе машинного обучения»

№ п/п	Модуль	Всего, час	Виды учебных занятий		
			лекции	практические занятия	самостоятельная работа
0	**Входное тестирование**	1			**1**
2	Модуль 1 – Введение в лингвостатистику, основные законы, предподготовка анализа текста лингвостатистики	10	4	3	3
3	Модуль 2 – Тематическое моделирование	9	3	3	3
4	Модуль 3 – Латентно-семантический анализ	13	4	4	5
5	Модуль 4 – Кластерный анализ	13	4	5	5
6	Модуль 5 – Методы классификации размеченного текста. Оценка качества классификации	14	4	5	5
7	**Модуль 6. Комплексное задание (проект)**	8		1	7
Итоговая аттестация			**Указывается вид (экзамен, зачёт, реферат и т.д.)**		
Итоговый тест		4	Зачет - Тест		

5.Календарный план-график реализации образовательной программы

(дата начала обучения – дата завершения обучения) в текущем календарном году, указания на периодичность набора групп (не менее 1 группы в месяц)

№ п/п	Наименование учебных модулей	Трудоёмкость (час)	Сроки обучения
0	**Входное тестирование**	**1**	**1.11.2020**

1	Модуль 1 − Введение в лингвостатистику, основные законы, предподготовка анализа текста лингвостатистики	10	2.11.2020- 3.11.2020-
2	Модуль 2 − Тематическое моделирование	9	4.11.2020- 5.11.2020-
3	Модуль 3 − Латентно-семантический анализ	13	6.11.2020- 9.11.2020
4	Модуль 4 – Кластерный анализ	13	10.11.2020- 13.11.2020
5	Модуль 5 − Методы классификации размеченного текста. Оценка качества классификации	14	14.11.2020- 17.11.2020
6	**Модуль 6. Комплексное задание (проект)**	8	18.11.2020- 20.11.2020
	Итоговое тестирование	4	21.11.2020
Всего:		72	1.11.2020- 21.11.2020

6.Учебно-тематический план программы « Интеллектуальный анализ текста на основе машинного обучения»

№ п/п	Модуль / Тема	Всего, час	Виды учебных занятий			Формы контроля
			лекции	практические занятия	самостоятельная работа	
0	**Входное тестирование**	**1**			**1**	**Тест**
1	Модуль 1 − Введение в лингвостатистику, основные законы, предподготовка анализа текста лингвостатистики	10	4	3	3	Кейс 1
2	Модуль 2 − Тематическое моделирование	9	3	3	3	Кейс 2 **Тест**
3	Модуль 3 − Латентно-семантический анализ	13	4	4	5	Кейс 3 **Тест**
4	Модуль 4 – Кластерный анализ	13	4	5	5	Кейс 4 **Тест**
5	Модуль 5 − Методы классификации размеченного текста. Оценка качества классификации	14	4	5	5	Кейс 5 **Тест**
6.	**Модуль 6. Комплексное задание (проект)**	8		1	7	Комплексное задание - Кейс

8	Итоговое тестирование	4			4	Тест

7. Учебная (рабочая) программа повышения квалификации « Интеллектуальный анализ текста на основе машинного обучения»

7.1 Модуль 1 «су, основные законы, предподготовка анализа текста лингвостатистики» (10 ак.часов)

Темы

Задачи, решаемые с применением интеллектуального анализа текста. Определение функции частотности слов. Статистическая мера связи в коллакациях: метод MI. Статистическая мера связи в коллакациях: логарифм правдоподобия. Синтагматическая связь между элементами словосочетаний. Основные законы лингвостатистики: Ципфа, Хипса, Ципфа с поправкой Мандельброта. Определение корпуса текста, разметка текста. Векторное представление текста. Избавление от стоп-слов в корпусе текста. Стэминг и лемматизация. Создание терм-документной матрицы. Формирование мешочка слов.

Задания в виде кейса 1 Установка R Studio, определение настроек**.**

7.2 Модуль 2 «Тематическое моделирование» (9 ак. часов)

Темы

Латентно-семантический анализ: сравнение двух термов между собой. Латентно-семантический анализ: сравнение двух документов между собой. Латентно-семантический анализ: сравнение терма и документа. Инструмент Word2Vec: алгоритма обучения : CBOW (Continuous Bag of Words). Инструмент Word2Vec: алгоритма обучения:Skip-gram. Инструмент Global2Vec.

Задания в виде кейса 2 и тестирования

7.3 Модуль 3 Латентно-семантический анализ (13 ак. часов)

Темы

Латентно-семантический анализ: сравнение двух термов между собой. Латентно-семантический анализ: сравнение двух документов между собой. Латентно-семантический анализ: сравнение терма и документа. Инструмент Word2Vec: алгоритма обучения : CBOW (Continuous Bag of Words). Инструмент Word2Vec: алгоритма обучения:Skip-gram. Инструмент Global2Vec.

Задания в виде кейса 3 и тестирования

7.4 Модуль 4 Кластерный анализ (13 ак. часов)

Темы

Методы кластеризации. Критерий качества кластеризации. Кластеризация методом Custom Search Folders. Кластеризация текста методом Suffix Tree. Кластеризация текста методом k-средних.

Задания в виде кейса 4 и тестирования

7.5 Модуль 5 Методы классификации размеченного текста. Оценка качества классификации (14 ак. часов)

Темы

Темы

Наивный байесовский классификатор при классификации текста. Принципы валидации данных для обучения моделей классификации текста. Метрики качества классификации текста. ROC-анализ для оценки качества классификации текста.

Задания в виде кейса 5 и тестирования

7.6 Модуль 6 **Комплексное задание (проект)** (8 ак. часов)

Темы

Загрузка и анализ корпуса текста. Предобработка текста. Реализация механизмов стемминга и лемматизации текста. Построение терм-документной матрицы. Реализация механизма частотного анализа текста, построение облака слов для нескольких статей. Выявление коллокации. Вычисление расстояния между статьями в одной и разных категориях, сравнение результатов. Кластеризация текста статей из 5-6 категорий. Реализация бинарной и многоклассовой классификации. Оценка качества полученных моделей.

Задания в виде комплексного проекта.

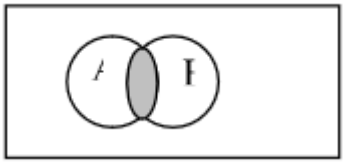
Описание практико-ориентированных заданий и кейсов

	Номер темы/модуля	**Наименование практического занятия**	**Описание**
1.1	1	Кейс-задание 1. **Установка R Studio**	Цель практического задания: формирование навыков работы с R Studio, установка соответствующих настроек
1.2.	2.	Кейс-задание 2. **Частотный анализ текста, построение облака слов**	Цели практического задания: научится проводить предобработку текста, проводить частотный анализ, строить облачко слов
1.3	3	Кейс-задание 3. **Мера TF-IDF, определение коллокаций в тексте**	Цели практического задания: выделение главной темы из текста с помощью меры TF-ITF
1.4	4	Кейс-задание 4 (модуль 4). **Кластеризация текста (обучение без учителя)**	Цели практического задания: проведение тематической кластеризации текста различными способами
1.5	5	**Кейс-задание 5. (Модуль 5). Фильтрация на основе наивного байесовского классификатора**	Цели практического задания: проведение классификации предобработанного текста на основе наивного байесовского классификатора
1.6.	6	**Кейс-задание 6. (модуль 6). Комплексное задание (проект)**	Цели практического задания: дать комплексную оценку сформированности компетенции цифровой экономики: способность управлять неструткрированной информацией и знаниями.

8.Оценочные материалы по образовательной программе

8.1. Вопросы тестирования по модулям

№ модуля	Вопросы входного тестирования	Вопросы промежуточного тестирования	Вопросы итогового тестирования
0	1. Сумма двух событий – это событие, состоящее A. в одновременном их появлении B. в появлении по крайней мере одного из них C. в их последовательном появлении D. в не появлении ни одного из них 2. Множество результатов, отобранных из генеральной совокупности, называют A. Выборкой B. Вариационным рядом C. Статистикой критерия D. Точечными оценками 3. Статистическая гипотеза – это утверждение о свойствах A. генеральной совокупности B. выборки C. конкретного объекта 4. Значение признака, находящееся в середине вариационного ряда наблюдений, a. мода; b. средняя арифметическая; c. медиана; d. частота; e. частость. 5. Названия гипотезы, противоположной проверяемой: A. нулевая B. простая C. конкурирующая D. альтернативная E. Сложная 6. Интервал возможных значений парного коэффициента корреляции при наличии между величинами X и Y отрицательной, но не функциональной связи: A. (–1; 0) B. (0; 1) C. (–1; -0,5) D. (–0,5; 0) E. [–1; 0] 7. Алгебраическая квадратная матрица является вырожденной, если: A. Определитель матрицы равен нулю		1. Отличие корпуса от коллекции текстов 1. Наличие разметки 2. Прагматическая ориентированность 3. Электронный вид 4. Представление в виде абзацев 2. Стеминг это: 1. процесс нахождения основы слова 2. процесс нахождения единых словоформ 3. процесс нахождения флексии 3. Терм-документная матрица – это матрица в которой строки это документы коллекции, столбцы термы (слова) 1. Матрица из 0 и 1 в зависимости от того встретилось слово в тексте или нет 2. Матрица, в которой строки это словосочетания, столбцы - слова 3. Матрица, в которой строки это документы коллекции, столбцы термы (слова) 4. Матрица, в которой текстам (абзацам) присвоены значения от -1 до 1 с учетом коннотации 4. Метрика TF-IDF больший вес предать: 1. словам с низкой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах 2. словам с высокой частотой в пределах конкретного

	B. Ранг матрицы равен размерности матрицы C. У нее имеется обратная матрица D. Определитель матрицы равен единице 8. Заданы множества A={2,3,4,5} и D={3,4,5}. Верным для них будет утверждение: A. Множество A - подмножество множества D B. Множество D - подмножество множества A C. Множество A и множество D равны D. Множество A - множество-степень множества D 9. На рисунке показано:  A. A∪ B B. A ∩B C. A∈ B D. A/ B 10. Для того, чтобы два вектора были ортогональны необходимо и достаточно, чтобы A. их скалярное произведение равнялось нулю B. их векторное произведение равнялось нулевом вектору C. их векторное произведение равнялось нулю **D.** их скалярное произведение равнялось нулевому вектору		документа и с высокой частотой употреблений в других документах 3. словам с низкой частотой в пределах конкретного документа и с высокой частотой употреблений в других документах 4. словам с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах 5. Для оценки связности текстов используется: 1. Мягкая косинусная мера 2. Расстояние Махалонобиса 3. Метрика Жакарда 4. Евклидовое расстояние 6. Если два слова часто встречаются в тексте рядом, то их называют: 1. парадигматически параллельными 2. имеющим атрибутивное сходство 3. имеющим реляционное сходство 4. синтагматически ассоциированными 7. Кластерный анализ предназначен для: 1. разбиения множества объектов на классы (пучки) в соответствии с их мерой сходства 2. исследования влияния одной или нескольких независимых переменных на зависимую переменную 3. анализа моделей зависимости среднего значения некоторой случайной величины

			одновременно от набора (основных) качественных факторов и (сопутствующих) количественных факторов 4. поиска зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях 8. Какой метод кластеризации производит классификацию документов с использованием самонастраивающейся нейронной сети? 1. k-средних 2. Scatter/Gather 3. Latent Semantic Analysis 4. Self-Organizing Maps 9. В основании наивного байесовского классификатора 1. Условие зависимости гипотез между собой 2. принципы вычисления апостериорной вероятности 3. принципы вычисления априорной вероятности 4. принципы определения максимума правдоподобия 10. Суть метода кросс-валидации заключается: 1. проведение кросс-проверки качества модели на различных выборках 2. В разбиении исходной выборки на пять частей и обучении модели на четырех из них, а на одной тестирование результатов 3. В разбиении обучающей выборки на несколько частей и обучении модели

			на части из них, а на одной тестирование результатов **4.** В делении выборки на обучающую и тестовую
2		9. В лингвостатистике закон, описывающий обратно пропорциональную зависимость частоты встречаемости слова от его порядкового номера в упорядоченном ряду: E. Закон Хипса F. Закон Ципфа G. Закон нормального распределения H. Закон Дальтона 10. Отличие корпуса от коллекции текстов E. Наличие разметки F. Прагматическая ориентированность G. Электронный вид H. Представление в виде абзацев 11. Стеминг это: A. процесс нахождения основы слова B. процесс нахождения единых словоформ C. процесс нахождения флексии 12. Технология One-hot encoding это:, f. Получение взвешенной матрицы признкаов; g. Процедура присвоения каждому слова аннотации; h. Получение разряженной матрицы из 0 и 1; i. Получении матрицы признаков с присвоением значений от -1 до 1 с учетом коннотации выражения. 13. Терм-документная матрица – это матрица в которой строки это документы	

		коллекции, столбцы термы (слова) F. Матрица из 0 и 1 в зависимости от того встретилось слово в тексте или нет G. Матрица, в которой строки это словосочетания, столбцы - слова H. Матрица, в которой строки это документы коллекции, столбцы термы (слова) I. Матрица, в которой текстам (абзацам) присвоены значения от -1 до 1 с учетом коннотации 14. Частотность как термин лексикостатистики это: F. Отношение частоты встречаемости слова по отношению к мощности словаря G. Отношение частоты встречаемости слова в тексте по отношению к размерности текста H. Отношение размерности текста по отношению к частоте встречаемости слова в тексте **I.** Отношение размерности текста к мощности словаря	
3		1. Если два слова часто встречаются в тексте рядом, то их называют: E. парадигматически параллельными F. имеющим атрибутивное сходство G. имеющим реляционное сходство H. синтагматически ассоциированными 2. Для оценки связности текстов используется: E. Мягкая косинусная мера F. Расстояние Махалонобиса G. Метрика Жакарда	

		H. Евклидовое расстояние 3. В чем отличие атрибутивного и реляционного сходства? A. Реляционное сходство учитывает связность слов в отличии от атрибутивного сходства B. Реляционное сходство учитывает отношение совместного употребления слов в отличии от атрибутивного сходства C. Атрибутивное сходство учитывает связность слов по принципу отношения к одному гипериониму в отличии от реляционного сходства D. Реляционное сходство учитывает связность слов в отличии от атрибутивного сходства 4. Метрика Резника используется: A. Для оценки связности текстов B. Для оценки сходства слов C. Для оценки связи в коллокациях D. Для оценки связности в биграммах 5. Для оценки синтагматической связи между элементами словосочетаний используется: A. Мера *t-score* B. Мера MI C. Мера VIF D. Метрика Жакарда 6. Под тематическим моделированием понимают: A. Статистический анализ текста для выявления латентных терминов в коллекции текстовых документов B. Статистический анализ текста для выявления латентных тем в	

		коллекции текстовых документов C. Статистическую меру согласованности текстов между собой D. Анализ вероятностных условных распределений тем над терминами 7. Вес терма в документе определяется: A. Как относительная частота встречаемости темы в документе B. Как булевая метрика C. Как функция от количества вхождений терма в документе D. Как абсолютная частота встречаемости темы в документе 8. Основная гипотеза распределения в лингвистике: A. Слова, встречающиеся в схожих контекстах, стремятся иметь близкий смысл B. Темы, встречающиеся в схожих документах, стремятся иметь близкий смысл C. Слова, встречающиеся в схожих коллокациях, стремятся иметь близкий смысл D. Слова, встречающиеся в схожих N-граммах, стремятся иметь близкий смысл 9. Метрика TF-IDF больший вес предать: A. словам с низкой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах B. словам с высокой частотой в пределах конкретного документа и с высокой частотой	

		употреблений в других документах C. словам с низкой частотой в пределах конкретного документа и с высокой частотой употреблений в других документах D. словам с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах 10. SVD -разложение при анализе текст используют: A. Для извлечения ключевой темы за счет уменьшения размеров B. Для избавления от разрежённости терм-доументной матрицы C. Для измерения метрики TF **D.** Для измерения метрики IDF	
4		1. Кластерный анализ предназначен для: A. разбиения множества объектов на классы (пучки) в соответствии с их мерой сходства B. исследования влияния одной или нескольких независимых переменных на зависимую переменную C. анализа моделей зависимости среднего значения некоторой случайной величины одновременно от набора (основных) качественных факторов и (сопутствующих) количественных факторов D. поиска зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях 2. Отличительной особенностью какого	

		метода кластеризации является то, что он позволяет сузить результаты поиска путём распределения их по «папкам»? A. Suffix Tree Clustering B. k-средних C. Custom Search Folders D. Self-Organizing Maps 3. Отличительной особенностью какого метода кластеризации является то, что он позволяет выявить латентные (скрытые) факторы, которые в дальнейшем будут основой для образования кластеров документов? A. Latent Semantic Analysis B. Single Link C. Scatter/Gather D. Concept Indexing (CI) 4. При применении какого метода кластерного анализа кластеры образуются в узлах специального вида дерева, которое строится из слов и фраз входных документов? A. Custom Search Folders B. Suffix Tree Clustering C. Suffix Tree Clustering D. Latent Semantic Analysis 5. При применении какого метода кластерного анализа множество документов разбивают на кластеры, расположенные в древовидной	

		структуре, получаемой с помощью иерархической агломеративной кластеризации? A. Custom Search Folders B. Suffix Tree Clustering C. k-средних D. Complete Link 6. Какой метод кластеризации представляет собой итеративный процесс, разбивающий множество документов на группы и представляющий затем эти группы пользователю для дальнейшего анализа. A. Scatter/Gather B. k-средних C. Suffix Tree Clustering D. Latent Semantic Analysis 7. В каком методе кластеризации кластеры представлены в виде центроидов, являющихся «центром массы» всех документов, входящих в кластер? A. Group Average B. Self-Organizing Maps C. k-средних D. Scatter/Gather 8. Отличительной особенностью какого метода кластеризации является то, что он разделяет множество документов на две части на каждом шаге рекурсии? A. Group Average B. Concept Indexing (CI) C. Single Link D. Complete Link	

		9. Какой метод кластеризации производит классификацию документов с использованием самонастраивающейся нейронной сети? A. k-средних B. Scatter/Gather C. Latent Semantic Analysis D. Self-Organizing Maps	
5		1. Суть метода кросс-валидации заключается: 5. проведение кросс-проверки качества модели на различных выборках 6. В разбиении исходной выборки на пять частей и обучении модели на четырех из них, а на одной тестирование результатов 7. В разбиении обучающей выборки на несколько частей и обучении модели на части из них, а на одной тестирование результатов 8. В делении выборки на обучающую и тестовую 2. Под чувствительностью модели понимают: 1. Доля истинно положительных примеров 2. Число верно классифицированных положительных примеров	

		3. Число положительных примеров, классифицированных как отрицательные 4. Доля положительных примеров, классифицированных как отрицательные 3. Метрика классификации, учитывающая специфичность и чувствительность 1. Точность (Accuracy) 2. Джини 3. AUC 4. F-мера 4. Основой для принятия решения о качестве классификации является: 1. Матрица сопряженности 2. ROC-кривая 3. Специфичность 4. Чувствительность 5. На основе ROC-кривой можно сделать вывод: 1. О качестве классификатора 2. О влиянии критерия отсечения на качество модели классификации 3. О точности предсказания класса 4. О наличии пропусков в исходных данных 6. В основании наивного байесовского классификатора 5. Условие зависимости	

		гипотез между собой 6. принципы вычисления апостериорной вероятности 7. принципы вычисления априорной вероятности 8. принципы определения максимума правдоподобия	

8.2. Минимальным проходным баллом теста считается 60% верных ответов по результатам суммарно 2 попыток

Кейс в 10 баллов: Максимум 10 баллов - выставляется при выполнении всех требований к отчету, подробном описании всех этапов и представлении выводов, увязывающих выполненный кейс с встречающимися в практике задачами.
8-10 баллов выставляется при выполнении всех требований к отчету, подробном описании всех этапов или представлении выводов, увязывающих выполненный кейс с встречающимися в практике задачами
6-7 баллов выставляется при выполнении всех требований к отчету, подробном описании отдельных этапов или кратких выводов.
Минимально допустимый балл 5 баллов -выставляется при выполнении минимального требования к отчету кейса

Кейс в 40 баллов: Максимум 40 баллов - выставляется при выполнении всех требований к отчету, подробном описании всех этапов и представлении выводов, увязывающих выполненный кейс с встречающимися в практике задачами.
30-40 баллов выставляется при выполнении всех требований к отчету, подробном описании всех этапов или представлении выводов, увязывающих выполненный кейс с встречающимися в практике задачами
21-29 баллов выставляется при выполнении всех требований к отчету, подробном описании отдельных этапов или кратких выводов.
Минимально допустимый балл 20 баллов -выставляется при выполнении минимального требования к отчёту кейса**.**

8.3. итоговое задание по всей образовательной программе

Цели задания: оценка сформированности компетенции по анализу данных и машинному обучению (способность управлять неструктурированной информацией и знаниями) на базовом уровне.

Оценка сформированности компетенции происходит в два этапа. В рамках шестого модуля слушатели выполняют итоговое комплексное задание по интеллектальному анализу текста, а «знаниевый» результат обучения подтверждают посредством выполнения теста (вопросы приведены в разделе 8.1.

8.4. Задания-кейсы

Кейс-задание 1 (по модулю 1). Установка R Studio

1. Скачать необходимые программы для своей операционной системы 1.1. R • Для win – https://cran.r-project.org/bin/windows/base/
   - Для macos – https://cran.r-project.org/bin/macosx/

   1.2. RStudio • https://rstudio.com/products/rstudio/download/#download

   1.3. Rtools (только для win) • https://cran.r-project.org/bin/windows/Rtools/
2. Установить скаченные программы. 2.1. Порядок установки R -> RStudio -> Rtools

   2.2. Для пользователей Windows: • Важно, чтобы путь установки НЕ содержал русских букв

   • Нежелательно устанавливать программы в стандартную папку «Program Files», поскольку в дальнейшем возникнут проблемы при установке дополнительных библиотек

   • Лучше всего установить R в «C:/R», RStudio в «C:/RStudio», а Rtools в «C:/Rtools»
3. RStudio – это оболочка для языка R. В данном курсе вы будете использовать именно эту программу. После установки запустите RStudio и убедитесь, что она правильно установлена. Также необходимо изменить некоторые стандартные настройки. Для этого откройте раздел Tools – Global options
   - 3.1. В разделе General установите настройки на UTF8.:

Кейс-задание 2 (модуль 2). Частотный анализ текста, построение облака слов

1. Ознакомиться с теоретическим материалом по теме (видео-лекциями и дополнительными материалами)
2. Скачать файл с исходным кодом скрипта на языке R (lab2_freq_analysis.R) и исходный текст для анализа (rus-news.txt)
3. Открыть файл скрипта в RStudio. Если комментарии на русском языке отображаются некорректно, то сменить кодировку и перезапустить RStudio (см. видео практики № 2)
4. Выполнить команды для подключения пакетов
5. Выполнить команду для просмотра пути к текущей рабочей директории – getwd()
6. Убедиться, что файл с исходными текстами (rus-news.txt) находится в рабочей директории. При необходимости скопировать файл в эту папку
7. Выполнить считывание файла с исходными текстами в переменную с помощью функции readLines()
8. Построить облако слов на основе исходного текста без предварительной обработки:
   8.1. Сформировать корпус текстов
   8.2. Сформировать терм-документную матрицу (ТДМ)
   8.3. Сформировать дата-фрейм из 2х столбцов – слов и их частот
   8.4. Построить облако слов
9. Выполнить предварительную обработку и построить облако слов:
   9.1. Перевести текст в нижний регистр символов
   9.2. Удалить числа
   9.3. Удалить знаки препинания
   9.4. Удалить стоп-слова
   9.5. Построить облако слов
10. Выполнить стемминг текста и построить облако слов:
    10.1. Выполнить стемминг с помощью стандартной функции text_tokens из пакета corpus
    10.2. Построить облако слов
11. Сравнить полученные результаты. Сделать выводы.
12. Составить отчёт, содержащий снимки экрана с результатами выполнения практической работы. Загрузить отчёт в личный кабинет.

Выполнять кейс-задание рекомендуется с использованием среды RStudio. Результатом выполнения кейс-задания является отчет №2.

Кейс-задание 3 (модуль 3). Мера TF-IDF, определение коллокаций в тексте

Цель: провести качественную чистку текста и оценить важность слов в тексте с помощью лингвостатистической меры TF-IDF.

Задание

1. Ознакомиться с теоретическим материалом по теме (видео-лекциями и дополнительными материалами)
2. Скачать файл с исходным кодом скрипта на языке R (lab3_freq_analysis.R), записанный в скрипте для второго кейса, и исходный текст для анализа (rus-news.txt)
3. Открыть файл скрипта в RStudio. Если комментарии на русском языке отображаются некорректно, то сменить кодировку и перезапустить RStudio (см. видео практики № 2)
4. Выполнить команды для подключения пакетов

```
install.packages("quanteda")

install.packages("knitr")
```

5. Подключить все библиотеки установленных пакетов:

```
library(tm)              # пакет для анализа текста (TextMining)

library(wordcloud)       # пакет для построения облака слов (wordcloud)

library(corpus)          # пакет для анализа корпусов текстов

library(RColorBrewer)    # пакет, содержащий наборы цветов

library(ggplot2)         # пакет для построения графиков и визуализации
данных

library(quanteda)        # пакет для количественного анализа текстовых
данных

library(knitr)           # пакет для построения отчётов в r
```

6. Выполнить команду для просмотра пути к текущей рабочей директории – getwd()
7. Убедиться, что файл с исходными текстами (rus-news.txt) находится в рабочей директории. При необходимости скопировать файл в эту папку
8. Выполнить считывание файла с исходными текстами в переменную с помощью функции readLines()
9. Выполнить предварительную обработку текста:
   9.1. Перевести текст в нижний регистр символов
   9.2. Удалить числа
   9.3. Удалить знаки препинания
   9.4. Удалить стоп-слова
10. Сформировать корпус текстов с помощью команды Corpus:
11. Сформировать терм-документную матрицу с помощью команды `TermDocumentMatrix`
12. Построить data-фрейм для удобного отображения значений меры TF-IDF.
13. Построить график значений меры TF-IDF для термов текста, таким образом проиллюстрировать закон Ципфа
14. Определить 10 наиболее часто встречающихся термов
15. Построить диаграмму 10 наиболее часто встречающихся термов, сделать выводы
16. Выявить коллокации в тексте с помощью команды textstat_collocations

17. Рассчитать меру сходства между текстами с помощью косинусной меры, меры Жаккарда, меры Дайса.

17* Рассчитать меру сходства между текстами с помощью меры Хаманна (`hamman`), простого мэтчинга (`simple matching`, поиска точного совпадения по каждому хешу).

18. Составить отчёт, содержащий снимки экрана с результатами выполнения практической работы. Загрузить отчёт в личный кабинет.

* не обязательно к выполнению (дополнительное задание)

Результатом выполнения кейс-задания является отчет №3.

Кейс-задание 4 (модуль 4). **Кластеризация текста (обучение без учителя)**

Задание

1. Ознакомиться с теоретическим материалом по теме (видео-лекциями и дополнительными материалами)
2. Скачать файл с исходным кодом скрипта на языке R (lab4_text-clustering.R), записанный в скрипте для второго кейса, и исходные тексты для анализа (rus-news.txt)
3. Открыть файл скрипта в RStudio. Если комментарии на русском языке отображаются некорректно, то сменить кодировку и перезапустить RStudio (см. видео практики № 2)
4. Выполнить команды для подключения пакетов
   install.packages("proxy")

   install.packages("dbscan")

   Установить пакеты tm и ggplot2

5. Подключить все библиотеки установленных пакетов:
   library(tm) # пакет для анализа текста (TextMining)

   library(ggplot2) # пакет для построения графиков и визуализации данных

   library(proxy) # пакет для вычисления мер близости между векторами и матрицами

   library(dbscan) # пакет для кластеризации методом dbscan

6. Выполнить команду для просмотра пути к текущей рабочей директории – getwd()
7. Убедиться, что файл с исходными текстами (rus-news.txt) находится в рабочей директории. При необходимости скопировать файл в эту папку
8. Выполнить считывание файла с исходными текстами в переменную с помощью функции readLines()
9. Выполнить предварительную обработку текста:
   9.1. Перевести текст в нижний регистр символов
   9.2. Удалить числа
   9.3. Удалить знаки препинания
   9.4. Удалить стоп-слова
10. Сформировать корпус текстов с помощью команды Corpus:
11. Сформировать терм-документную матрицу с помощью команды TermDocumentMatrix
12. Построить data-фрейм для удобного отображения значений меры TF-IDF.
13. В качестве весов термов в матрице использовать метрику TF-IDF с помощью команды weightTfIdf из библиотеки tm
14. Отобразить матрицу в качестве весов, определенных по мере TF-IDF, Определить количество строк и столбцов в полученной матрице, выразить матрицу в виде дата-фрейма

15. Отбросить редкие термы с коэффициентом разряженности выше 0.999 с помощью команды removeSparseTerms
16. Построить матрицу косинусных расстояний между текстами для последующей кластеризации, выразить матрицу в виде дата-фрейма
17. Определить число кластеров для метода k-средних (в примере 4)
18. По методу k-средних провести кластеризацию с помощью команды kmeans над матрицей tfidf.matrix.

17 По методу Уорда (суффиксных деревьев) провести кластеризацию с помощью команды hclust над матрицей dist.matrix

* 18 По методу dbscan провести кластеризацию с помощью команды dbscan над матрицей dist.matrix

19 Провести визуализацию полученных результатов кластеризации

20 Составить отчёт, содержащий снимки экрана с результатами выполнения практической работы, сделать выводы. Загрузить отчёт в личный кабинет.

* дополнительное задание

Результатом выполнения кейс-задания является отчет №4.

Кейс-задание 5. (Модуль 5). Фильтрация на основе наивного байесовского классификатора

Классификация в анализе текста

Цель: провести анализ текста, отделив спам от неспама, используя наивный байесовский классификатор.

Задание

1. Ознакомиться с теоретическим материалом по теме (видео-лекциями и дополнительными материалами)
2. Скачать файл в формату .csv, в котором хранится база размеченных данных спам/неспам.
3. Скачать файл с исходным кодом скрипта на языке R (naive-bayes-spam-detection.R), записанный в скрипте для второго кейса.
4. Открыть файл скрипта в RStudio. Если комментарии на русском языке отображаются некорректно, то сменить кодировку и перезапустить RStudio (см. видео практики № 2)
5. Выполнить команды для подключения пакетов

   `install.packages("quanteda")` пакет для работы с текстом

   `install.packages("RColorBrewer")` пакет для работы с цветом для построения облачков слов

   `install.packages("ggplot2")` пакет для поддержки графики

   `install.packages("caret")` пакет для расчета матрицы сопряженности и определения метрик качества классификатора

   Возможно пакеты `quanteda` и `caret` установятся не полностью, в этом случае дополнительно установить

```
install.packages("spacyr")

install.packages("e1071")
```

6. Подключить все библиотеки установленных пакетов:
```
require(quanteda)

?quanteda

require(RColorBrewer)

require(ggplot2)
```
В случае неполного подключения пакетов `quanteda` и `caret` выполнить команды
```
library(caret)

library(e1071)
```

7. Выполнить команду для просмотра пути к текущей рабочей директории – getwd()
8. Убедиться, что файл с исходными текстами (spam.csv) находится в рабочей директории. При необходимости скопировать файл в эту папку
9. Выполнить считывание файла с исходными размеченными данными в файл с помощью функций
fileName = "spam.csv"

10. Разделить базу сообщений на две побазы, только со спамом и только с неспамом:
spam=read.csv(fileName,header=TRUE, sep=",", quote='\"\"', stringsAsFactors=FALSE)

11. Просмотреть статистики по сформированным базам с помощью команды table:
12. Для удобства сменить названия столбцов таблицы
13. Настроить генератор случайных чисел (установка начального номера последовательности) с помощью команды set.seed(), необходимый для перемешивания случайным образом данных с помощью команды [sample(nrow()),]
14. Сформировать корпус размеченных сообщений только из спамам с помощью команды Corpus:
15. Сформировать матрицу "документы-признаки" с помощью функции dfm()
16. Преобразовать полученную матрицу в DataFrame.
17. Сформировать облако слов для СПАМ-сообщений
18. Сформировать облако слов для НЕСПАМ-сообщений
19. Сделать выводы (промежуточные) о разнице сообщений спам и неспам
20. Разделить общую выборку сообщений на обучающую и тестовую подвыборки
21. Сформировать матрицу "документы-признаки" с помощью функции dfm()
22. Отсечь редко встречающиеся слова
23. Преобразовать полученную матрицу в DataFrame.
24. Разделить полученный DataFrame на обучающую и тестовую выборки
25. Обучить наивный байесовский классификатор на обучающей выборке с помощью команды textmodel_nb
26. Выполнить прогнозирование на тестовой выборке
27. Сформировать таблицу на тестовой выборке из спрогнозированных и реальных типов (спам или «неспам») сообщений table(predicted=pred_df$pred, actual=spam.test[,1])
28. Рассчитать матрицу сопряженности (неточности) с помощью команды table
29. Определить показатели качества классификации с помощью команды confusionMatrix
30. Сделать выводы о качестве классификации на основе наивного байесовского классификатора.
31. Сделать отчет по работе, загрузить в личный кабинет.

Результатом выполнения кейс-задания является отчет № 5.

Кейс-задание 6. (модуль 6). Комплексное задание (проект)

Итоговое комплексное Задание имеет цель оценить сформированность компетенции: способность управлять нестуктурированной информацией и знаниями

1. Загрузить исходный корпус текста – новости с сайта lenta.ru за 1999-2019 гг (340 Мб в сжатом виде, > 800 тыс. новостей из > 20 категорий) - https://github.com/yutkin/Lenta.Ru-News-Dataset/releases/download/v1.0/lenta-ru-news.csv.gz
2. Выполнить загрузку корпуса. Проанализировать его количественные параметры (кол-во статей, категорий и тегов, даты публикаций и др.)
3. Реализовать механизм очистки и предварительной обработки текста, в т.ч. удаления стоп-слов. Проверить на одной или нескольких статьях
4. Реализовать механизм стемминга и лемматизации текста
5. Построить терм-документную матрицу (ТДМ)
6. Реализовать механизм частотного анализа текста, построить облако слов для нескольких статей
7. Выявить коллокации. Построить ТДМ с учётом n-грамм (биграмм и триграмм)
8. Вычислить расстояние между статьями из одной категории, а также из разных категорий. Сравнить результаты. Использовать различные метрики сходства (косинусную меру, меру Жакарда и др.)
9. Произвести кластеризацию текстов статей из 5-6 категорий. Сравнить результаты кластеризации с метками исходных категорий статей.
10. Выбрать 2 категории. Реализовать бинарную классификацию статей из 2х категорий. Оценить качество полученной модели.
11. Выбрать 10 категорий. Реализовать многоклассовую классификацию статей из 10 категорий. Оценить качество полученной модели.

Результатов выполнения задания является отчет 6, загружаемый в систему.

8.5.

Наименовании модуля	Задание	Балл	Критерии оценки
Входное тестирование	ТЕСТ	10	Проходной балл -5
Модуль 1 − Введение в лингвостатистику, основные законы, предподготовка анализа текста лингвостатистики	Кейс 1	10	Система R должна быть загружена (минимально допустимый балл - 8 баллов)
Модуль 2 − Тематическое моделирование	Кейс 2	10	Для минимального балла (5) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов анализа данных
	Тест	10	Проходной балл - 6
Модуль 3 − Латентно-семантический анализ	Кейс 3	10	Для минимального балла (5) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов латентно-семантического анализа
	Тест	10	Проходной балл - 6
Модуль 4 – Кластерный анализ	Кейс 4	10	Для минимального балла (5) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов кластеризации
	Тест	9	Проходной балл - 5

Модуль 5 – Методы классификации размеченного текста. Оценка качества классификации	Кейс 5	12	Для минимального балла (6) должны быть выполнены все задания, но могут быть допущены неточности в интерпретации результатов классификации текста
	Тест	6	Проходной балл – 3
Модуль 6 – **Комплексное задание (проект)**	Задание на проект	40	Итоговое задание считается выполненным на минимально допустимый балл 20, если есть полностью сформированное задание, но отсутствует интерпретация полученных результатов
Итоговая аттестация	Тест	10	Проходной балл – 6
	Минимальный балл для получения зачета по КПК - 80, максимальный - 147		

9.Организационно-педагогические условия реализации программы

9.1. Кадровое обеспечение программы

№ п/п	**Фамилия, имя, отчество (при наличии)**	**Место основной работы и должность, ученая степень и ученое звание (при наличии)**	**Ссылки на веб-страницы с портфолио (при наличии)**	**Фото в формате jpeg**	**Отметка о полученном согласии на обработку персональных данных**
1	Лакман Ирина Александровна	ФГБОУ ВО «Башкирский государственный университет", заведующая лабораторией исследования социально-экономических проблем регионов, к.т.н., доцент		**Загружено на платформу**	Да
2	Галямов Айрат Фаритович	ФГБОУ ВО «Башкирский государственный университет", доцент кафедры цифровой экономики и коммуникаций, к.т.н., доцент		**Загружено на платформу**	Да

9.2.Учебно-методическое обеспечение и информационное сопровождение

Учебно-методические материалы	
Методы, формы и технологии	Методические разработки, материалы курса, учебная литература
Методы организации учебно-познавательной деятельности: практический;	1. Анализ данных : учебник для академического бакалавриата / ГУ - Высшая школа экономики; под

Форма: дистанционная; Технологии: Информационно – коммуникационная технология; Кейс технология	ред. В. С. Мхитаряна .— Москва : Юрайт, 2016 .— 490 (13 экз.) 2. **Наследов, Андрей Дмитриевич**. Математические методы психологического исследования. Анализ и интерпретация данных : учеб. пособие / А. Д. Наследов .— 2-е изд., испр. и доп. — СПб. : Речь, 2006 .— 392 с. (1 экз.) 3. **Тюрин , Ю. Н.** Анализ данных на компьютере : учеб. пособ. по напр. "Математика", "Математика. Прикладная математика" / Ю. Н. Тюрин , А. А. Макаров .— 4-е изд., перераб. — М. : Форум, 2010 .— 367 с. (21 экз.) 4. Латентно-семантический анализ в задаче автоматического аннотирования [[Текст]] / И. В. Машечкин [и др.] // Программирование. — 2011 .— N 6 .— С. 67-77 . 5. Лингвостатистика и вычислительная лингвистика : труды по лингвостатистике / [отв. ред. Я. Соонтак] .— Тарту, 1982 .— 168 с. (1 экз.) **Дополнительная литература** 1. **Корнилина, Е. Д.** Латентно-семантический анализ предвыборных партийных программ на выборах в Государственную Думу 2007 и 2011 годов [[Текст]] / Е. Д. Корнилина, А. П. Петров // Вестник Московского университета. Сер. 12. Политические науки. — 2013 .— № 2 .— С. 80-88 :. 2. **Игнатьев, Н. А. (доктор физико-математических наук)** . Вычисление обобщенных показателей и интеллектуальный анализ данных [[Текст]] / Н. А. Игнатьев // Автоматика и телемеханика. — 2011 .— N 5 .— С. 183-190 :. 3. **Овсяницкая, Лариса Юрьевна (кандидат технических наук)** . Интеллектуальный анализ данных как составляющая педагогического управления [[Текст]] / Л. Ю. Овсяницкая // Образование и наука. — 2013 .— № 10 .— С. 80-90. 4. **Винстон, Уэйн**. Бизнес-моделирование и анализ данных. Решение актуальных задач с помощью Microsoft EXCEL : пер. с англ. яз. / У. Винстон ; перевод Ю. Бочиной .— 5-е изд. — Санкт-Петербург : Питер, 2018 .— 864 с. (8 экз.) 5. Бонцанини М. Анализ социальных медиа на Phyton / пер. с англ. А.В. Логунова.— М : ДМК Пресс, 2018.— 288 с.: ил. . [Электронный ресурс] URL=https://e.lanbook.com/reader/book/108129/#4 . Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А.А. Слинкина. – М.: ДМК Пресс, 2015. – 400 с.: ил. [Электронный ресурс] URL=https://e.lanbook.com/reader/book/69955/#4

Информационное сопровождение	
Электронные образовательные ресурсы	Электронные информационные ресурсы
http://sdo.bashedu.ru/course/view.php?id=2254	https://www.kaggle.com/

9.3.Материально-технические условия реализации программы

Вид занятий	Наименование оборудования, программного обеспечения
Лекции, практические занятия	**Аппаратные требования** Intel Pentium или аналогичный процессор с тактовой частотой 300MHz и выше. SVGA монитор, с разрешением экрана, как минимум, 800x600 точек и глубиной цвета 16 bit (рекомендуемое разрешение экрана — 1024x768). Звуковая карта, акустическая система или наушники. Доступ в Интернет со скоростью 56 кбит/с и выше. **Программное обеспечение** Операционная система: Windows 7 или более продвинутая, Macintosh, Linux Браузер: Internet Explorer 7 или более продвинутый, Mozilla Firefox (скачать бесплатно: http://www.mozilla.org/download.html) и т.п. **Для просмотра электронных версий учебных курсов необходимо наличие установленных программ:** Microsoft Internet Explorer 7.0 и выше (Загрузить с сайта www.microsoft.com) Adobe Flash Player версии 7.0 и выше (Загрузить с сайта http://www.adobe.com/)

7. Паспорт компетенций

1	Наименование компетенции	Пояснения	Способность управлять неструктурированной информацией и данными
2	Указание типа компетенции	Профессиональная	Профессиональная
3	Определение, содержание и основные сущностные характеристики компетенции		– Под компетенцией понимается способность управлять неструктурированной информацией и данными, в виде размеченного текста, используя современные алгоритмы машинного обучения Слушатель должен: знать: – основные метрики лингвостатистики; – основные законы лингвостатистики; (Хипса, Ципфа) – основные принципы разметки текста; – способы векторного представления текста; – метрики по реляционному и атрибутивному сходству текста; – метрики ассоциации для измерения в коллакациях. – способы кластеризации текста; – инструмент TF-IDF для анализа главной темы – основные методы латентно-семантического анализа текста. – инструменты машинного обучения (наивный байесовский классификатор) для классификации текста; – основные метрики оценки качества классификации; уметь: – проводить качественную чистку текста, избавляться от стоп-слов и проводить лемматизацию и стеминг текста; – создавать терм-документную матрицу двумя способами; – использовать мешочек слов для анализа текста; – применять процедуру TF-IDF для анализа главной темы; – проводить классификацию текста (например спам/неспам) с помощью наивного байесовского классификатора; – определять меру сходства текста и меру ассоциации в коллакациях. – применять латентно-семантический анализ текста владеть:

			— методами предподготовки к проведению анализа текста, используя средства среды RStudio; — навыками тематического моделирования, используя инструменты алгоритма TF-IDF. а также иметь опыт — иметь опыт применения современных методов и подходов интеллектуального анализа текста на базовом уровне средствами машинного обучения.
4	Дескриптор знаний, умений и навыков по уровням	Уровни сформированности компетенции обучающегося начальный/базовый	индикаторы сформированности компетенции (знать, уметь, владеть) обучающегося в зависимости от уровня начальный/базовый
		Начальный уровень (Компетенция недостаточно развита. Частично проявляет навыки, входящие в состав компетенции. Пытается, стремится проявлять нужные навыки, понимает их необходимость, но у него не всегда получается)	знать: — основные метрики лингвостатистики; — основные законы лингвостатистики; (Хипса, Ципфа) — основные принципы разметки текста; — способы векторного представления текста; — метрики по реляционному и атрибутивному сходству текста; — метрики ассоциации для измерения в коллакациях. — способы кластеризации текста **уметь:** — проводить качественную чистку данных, проводить качественную чистку текста, избавляться от стоп-слов и проводить лемматизацию и стеминг текста; — создавать терм-документную матрицу двумя способами; — использовать мешочек слов для анализа текста владеть: — методами предподготовки к проведению анализа текста, используя средства среды RStudio;
		Базовый уровень (Уверенно владеет навыками, способен, проявлять соответствующие навыки в ситуациях с элементами неопределённости сложности)	знать: — инструмент TF-IDF для анализа главной темы — основные методы латентно-семантического анализа текста. — инструменты машинного обучения (наивный байесовский классификатор) для классификации текста; — основные метрики оценки качества классификации; **уметь:** — применять процедуру TF-IDF для анализа главной темы;

			– проводить классификацию текста (например спам/неспам) с помощью наивного байесовского классификатора; – определять меру сходства текста и меру ассоциации в коллакациях. – применять латентно-семантический анализ текста владеть: – навыками тематического моделирования, используя инструменты алгоритма TF-IDF.
5	Характеристика взаимосвязи данной компетенции с другими компетенциями/ необходимость владения другими компетенциями для формирования данной компетенции		Компетенции цифровой грамотности Компетенции в базовой лингвистики (лингвистический анализ)
6	Средства и технологии оценки		Кейсы-задания, комплексное итоговое задание (проект), тесты

VI.Иная информация о качестве и востребованности образовательной программы (результаты профессионально-общественной аккредитации образовательной программы, включение в системы рейтингования, призовые места по результатам проведения конкурсов образовательных программ и др.) (при наличии)

Общественная аккредитация программы не проводилась

V.Рекомендаций к программе от работодателей: наличие не менее двух писем и/или подтверждения на цифровой платформе Государственной системы предоставления ПЦС от работодателей о рекомендации образовательной программы для реализации в рамках Государственной системы предоставления ПЦС на формирование у трудоспособного населения компетенций цифровой экономики с указанием востребованности результатов освоения программы в сфере деятельности соответствующих компаний и готовности к рассмотрению заявок наиболее успешно освоивших образовательную программу граждан на прохождение стажировки и (или) собеседования на предмет трудоустройства путем проставления отметки в профиле программы

Документы загружены на платформу

Указание на возможные сценарии профессиональной траектории граждан по итогам освоения образовательной программы (в соответствии с приложением)

Цели получения персонального цифрового сертификата	
текущий статус	**цель**
Развитие компетенций в текущей сфере занятости	
работающий по найму в организации, на предприятии	развитие профессиональных качеств
работающий по найму в организации, на предприятии	повышение заработной платы

работающий по найму в организации, на предприятии	смена работы без изменения сферы профессиональной деятельности
Переход в новую сферу занятости	
8.	
освоение смежных профессиональных областей	повышение уровня дохода, расширение профессиональной деятельности

VII.Дополнительная информация

VIII.Приложенные Скан-копии

Утверждённая рабочая программа (подпись, печать, в формате pdf) загружена на платформу