МИНОБРНАУКИ РОССИИ
ФГБОУ ВО «БАШКИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
ИНСТИТУТ ЭКОНОМИКИ, ФИНАНСОВ И БИЗНЕСА

ИНСТИТУТ ЭКОНОМИКИ, ФИНАНСОВ И БИЗНЕСА

Утверждено:
на заседании кафедры
протокол от «22» мая 2017 г. № 9
Зав. кафедрой
/Р.Х.Бахитова

Согласовано:
Председатель УМК института
/ Н.Г. Вишневская

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Анализ неструктурированной информации

Вариативная часть

Программа магистратуры

Направление подготовки
38.04.05 Бизнес-информатика

Направленность (профиль) подготовки
«Информационная бизнес-аналитика»

Квалификация
магистр

Разработчик (составитель) РПД:
доцент, к-т техн..наук Лакман И.А.

Для приема 2017 г.

Уфа 2017 г.

Составитель / составители: к-т техн. наук, доцент Лакман И.А.

Рабочая программа дисциплины утверждена на заседании кафедры Математические методы в экономике протокол от «22» мая 2017 г. № 9.

Дополнения и изменения, внесенные в рабочую программу дисциплины: обновлен список рекомендованной литературы, вопросы к экзамену, профессиональные базы данных и информационные системы, утверждены на заседании кафедры математических методов в экономике протокол от «18» июня 2018 г. № 13.

Заведующий кафедрой / Бахитова Р.Х./

Дополнения и изменения, внесенные в рабочую программу дисциплины, утверждены на заседании кафедры _____,
протокол № ____ от «____» _____ 20 _ г.

Заведующий кафедрой _____/ _____Ф.И.О/

Дополнения и изменения, внесенные в рабочую программу дисциплины, утверждены на заседании кафедры _____,
протокол № ____ от «____» _____ 20 _ г.

Заведующий кафедрой _____/ _____Ф.И.О/

Дополнения и изменения, внесенные в рабочую программу дисциплины, утверждены на заседании кафедры _____,
протокол № ____ от «____» _____ 20 _ г.

Заведующий кафедрой _____/ _____Ф.И.О/

Список документов и материалов

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы 4

2. Цель и место дисциплины в структуре образовательной программы 6

3. Содержание рабочей программы (объем дисциплины, типы и виды учебных занятий, учебно-методическое обеспечение самостоятельной работы обучающихся) 6

4. Фонд оценочных средств по дисциплине 11

4.1 Перечень компетенций с указанием этапов их формирования в процессе освоения образовательной программы. Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания 11

4.2. Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и опыта деятельности, характеризующих этапы формирования компетенций 14

5. Учебно-методическое и информационное обеспечение дисциплины 31

5.1. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины 31

5.2. Перечень ресурсов информационно-телекоммуникационной сети «Интернет» и программного обеспечения, необходимых для освоения дисциплины 34

6. Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине 35

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

В результате освоения образовательной программы обучающийся должен овладеть следующими результатами обучения по дисциплине:

Результаты обучения		Формируемые компетенции (с указанием кода)	Примечание
Знания	Знать: основные принципы анализа неструктурированной информации (текстовой); метрики лингвостатистики; основные законы лингвостатистики; основные принципы разметки текста; способы векторного представления текста; способы кластеризации текста; основные методы латентно-семантического анализа текста.	ПК-1: способностью готовить аналитические материалы для оценки мероприятий и выработки стратегических решений в области ИКТ	
	Знать: основные методы латентно-семантического анализа текста; основные алгоритмы интеллектуального анализа текста	ПК-12 - способностью проводить научные исследования для выработки стратегических решений в области ИКТ	
Умения	Уметь: проводить качественную чистку текста, избавляться от стоп-слов и проводить лемматизацию; создавать терм-документную матрицу двумя способами; использовать мешочек слов для анализа текста;	ПК-1: способностью готовить аналитические материалы для оценки мероприятий и выработки стратегических решений в области ИКТ	
	Уметь: применять процедуру TF-IDF для анализа главной темы; применять латентно-семантический анализ текста. Применять наивный байесовский классификатор для анализа текста	ПК-12 - способностью проводить научные исследования для выработки стратегических решений в области ИКТ	

Владения (навыки/ опыт деятельности)	Владеть: методами предподготовки к проведению анализа текста, используя средства среды R Studio; современными методами и подходами интеллектуального анализа текста на базовом уровне	ПК-1: способностью готовить аналитические материалы для оценки мероприятий и выработки стратегических решений в области ИКТ	
	Владеть:навыками тематического моделирования, используя инструменты алгоритма TF-IDF, и метода Word2Vec, Global2Vec	ПК-12 - способностью проводить научные исследования для выработки стратегических решений в области ИКТ	

2. Цель и место дисциплины в структуре образовательной программы

Дисциплина «Анализ неструктурированной информации» является дисциплиной по выбору вариативной части образовательной программы.

Целью изучения дисциплины «Анализ неструктурированной информации» является формирование знаний, навыков и умений, необходимых для понимания и практической реализации современных концепций интеллектуального анализа текста, относящегося к неструктурированной информации.

Дисциплина изучается на 2 курсе магистратуры в 1 и 2 семестрах.

Для успешного освоения курса необходимы компетенции, сформированные в рамках курса бакалавриата теории вероятностей и математической статистики, а также магистерского курса Анализа данных.

Дисциплина является базовой для освоения профильных дисциплин, таких как «Системы искусственного интеллекта», успешного прохождения практики и государственной итоговой аттестации.

3. Содержание рабочей программы (объем дисциплины, типы и виды учебных занятий, учебно-методическое обеспечение самостоятельной работы обучающихся)

МИНОБРНАУКИ РОССИИ
ФГБОУ ВО «БАШКИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
ИНСТИТУТ ЭКОНОМИКИ, ФИНАНСОВ И БИЗНЕСА

СОДЕРЖАНИЕ РАБОЧЕЙ ПРОГРАММЫ

дисциплины «Анализ неструктурированной информации»
на 3, 4 семестр
очно-заочной формы обучения

Вид работы	**Объем дисциплины**	
Общая трудоемкость дисциплины (ЗЕТ / часов)	6/216	
	3 семестр	4 семестр
	2/72	4/144
Учебных часов на контактную работу с преподавателем:		
лекций	4	4
практических/ семинарских	16	16
лабораторных		
других (групповая, индивидуальная консультация и иные виды учебной деятельности, предусматривающие работу обучающихся с преподавателем) (ФКР)	0,2	1,2
Учебных часов на самостоятельную работу обучающихся (СР)	51,8	86,8
Учебных часов на подготовку к экзамену/зачету/дифференцированному зачету (Контроль)		36

Форма(ы) контроля:
Зачет 3 семестр
Экзамен 4 семестр

№ п/п	Тема и содержание	Форма изучения материалов: лекции, практические занятия, семинарские занятия, лабораторные работы, самостоятельная работа и трудоемкость (в часах)					Основная и дополнительная литература, рекомендуемая магистрантам (номера из списка)	Задания по самостоятельной работе магистрантов	Форма текущего контроля успеваемости (коллоквиумы, контрольные работы, компьютерные тесты и т.п.)
		Всего	ЛК	ПР/СЕМ	ЛР	СРС			
1	2	3	4	5	6	7	8	9	10
Модуль 1. Введение в лингвостатистику, основные законы лингвостатистики									
1.	Функция частотности слов Статистическая мера связи в коллакациях Синтагматическая связь между элементами словосочетаний	24	1	3		20	№№ 4, 5, из основного списка, №№ 1 из дополнительного списка	Работа с литературой и другими рекомендуемыми источниками. Подготовка доклада по теме лингвостатистики	Оценка подготовленных докладов
2.	Закон Ципфа Закон Хипса	24	1	3		20	№№ 4, 5, из основного списка, №№ 1 из дополнительного списка	Работа с литературой и другими рекомендуемыми источниками. Подготовка доклада по теме лингвостатистики	Оценка подготовленных докладов
3.	Итоговый контроль по Модулю 1	4				4		Подготовка к устному опросу	Устный опрос
Модуль 2. Предподготовка к проведению анализа текста									
4.	Основные понятия интеллектуального анализа текста • Корпус текста, разметка текста • Векторное представление текста	24	1	3		20	№№ 1,2 из основного списка, №№ 3, 4 из дополнительного списка	Работа с литературой и другими рекомендуемыми источниками. Выполнение кейса 1	Проверка выполнения кейса 1

5.	Основные этапы предподготовки анализа текста: • Избавление от стоп-слов, • Стэминг и лемматизация • Создание терм-документной матрицы • Мешочек слов	24	1	3		20	№№ 1,2 из основного списка, №№ 3, 4 из дополнительного списка	Работа с литературой и другими рекомендуемыми источниками. Выполнение кейса 1	Проверка выполнения кейса 1
6.	Итоговый контроль по Модулю 2	3,8				3,8		Подготовка к устному опросу	Устный опросу
7.	ФКР	0,2				0,2			
8.	Зачет								
Модуль 3. Тематическое моделирование									
9.	Мера сходства и ассоциации слов: • Дистрибутивный анализ: контекстные вектора • Косинусная мера сходства • Мягкая косинусная мера сходства	24	1	3		20	№№ 2, 3 из основного списка, №№ 1, 4 из дополнительного списка	Работа с литературой и другими рекомендуемыми источниками. Выполнение кейса 2	Проверка выполнения кейса 2
10.	Выделение главной темы: • Векторное представление слов • Выделение главной темы с помощью алгоритма TF-IDF • Кластеризация текста	24	1	3		20	№№ 2, 3 из основного списка, №№ 1, 4 из дополнительного списка	Работа с литературой и другими рекомендуемыми источниками. Выполнение кейса 2	Проверка выполнения кейса2
11.	Итоговый контроль по модулю 3	4				4		Подготовка к устному опросу	Устный опросу
Модуль 4. Латентно-семантический анализ									
12.	Латентно-семантический анализ: • сравнение двух термов между собой; • сравнение двух документов между собой;	24	1	3		20	№№ 2, 4 из основного списка, №№ 1, 4 из дополнительного списка	Работа с литературой и другими рекомендуемыми источниками. Выполнение кейса 3	Проверка выполнения кейса 3

	• сравнение терма и документа. • Применение наивного байесовского классификатора для анализа текста								
13.	Инструмент Word2Vec: • алгоритма обучения : CBOW (Continuous Bag of Words) • алгоритма обучения :Skip-gram Инструмент Global2Vec:	24	1	3		20	№№ 2,3 из основного списка, №№ 4 из дополнительного списка	Работа с литературой и другими рекомендуемыми источниками. Выполнение кейса 4	Проверка выполнения кейса4
14.	Итоговый контроль по Модулю 4	2,8				2,8		Подготовка к устному опросу	Устный опрос
15.	ФКР	3,2				3,2			
16.	Экзамен	36				36			
17.	Всего часов:	216	8	32	-	176			

4. Фонд оценочных средств по дисциплине

4.1 Перечень компетенций с указанием этапов их формирования в процессе освоения образовательной программы. Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания

ПК-1: способностью готовить аналитические материалы для оценки мероприятий и выработки стратегических решений в области ИКТ

Этап (уровень) освоения компетенции	**Планируемые результаты обучения** (показатели достижения заданного уровня освоения компетенций)	**Критерии оценивания результатов обучения**			
		2	**3**	**4**	**5**
Первый этап (уровень)	Знать: основные принципы анализа неструктурированной информации (текстовой); метрики лингвостатистики; основные законы лингвостатистики; основные принципы разметки текста; способы векторного представления текста; способы кластеризации текста; основные методы латентно-семантического анализа текста	Фрагментарные представления об основных принципах анализа неструктурированной информации (текстовой); метриках лингвостатистики; основных законах лингвостатистики; основных принципах разметки текста; способах векторного представления текста; способах кластеризации текста; основных методах латентно-	Неполные представления об основных принципах анализа неструктурированной информации (текстовой); метриках лингвостатистики; основных законах лингвостатистики; основных принципах разметки текста; способах векторного представления текста; способах кластеризации текста; основных методах латентно-	Сформированные, но содержащие отдельные пробелы представления об основных принципах анализа неструктурированной информации (текстовой); метриках лингвостатистики; основных законах лингвостатистики; основных принципах разметки текста; способах векторного представления текста; способах кластеризации текста; основных методах латентно-	Сформированные систематические представления об основных принципах анализа неструктурированной информации (текстовой); метриках лингвостатистики; основных законах лингвостатистики; основных принципах разметки текста; способах векторного представления текста; способах кластеризации текста; основных

		семантического анализа текста	семантического анализа текста	семантического анализа текста	методах латентно-семантического анализа текста
Второй этап (уровень)	Уметь: проводить качественную чистку текста, избавляться от стоп-слов и проводить лемматизацию; создавать терм-документную матрицу двумя способами; использовать мешочек слов для анализа текста;	Фрагментарные умения проводить качественную чистку текста, избавляться от стоп-слов и проводить лемматизацию; создавать терм-документную матрицу двумя способами; использовать мешочек слов для анализа текста;	В целом успешное, но не систематическое умение проводить качественную чистку текста, избавляться от стоп-слов и проводить лемматизацию; создавать терм-документную матрицу двумя способами; использовать мешочек слов для анализа текста;	В целом успешное, но содержащее отдельные пробелы в умении проводить качественную чистку текста, избавляться от стоп-слов и проводить лемматизацию; создавать терм-документную матрицу двумя способами; использовать мешочек слов для анализа текста;	Сформированное умение проводить качественную чистку текста, избавляться от стоп-слов и проводить лемматизацию; создавать терм-документную матрицу двумя способами; использовать мешочек слов для анализа текста;
Третий этап (уровень)	Владеть: методами предподготовки к проведению анализа текста, используя средства среды R Studio; современными методами и подходами интеллектуального анализа текста на базовом уровне	Фрагментарное владение методами предподготовки к проведению анализа текста, используя средства среды R Studio; современными методами и подходами интеллектуального анализа текста на базовом уровне	В целом успешное, но не систематическое применение методов предподготовки к проведению анализа текста, используя средства среды R Studio; Современных методаов и подходов интеллектуального анализа текста на базовом уровне	В целом успешное, но содержащее отдельные пробелы применение методов предподготовки к проведению анализа текста, используя средства среды R Studio; Современных методаов и подходов интеллектуального анализа текста на базовом уровне	Успешное и систематическое применение методов предподготовки к проведению анализа текста, используя средства среды R Studio; Современных методаов и подходов интеллектуального анализа текста на базовом уровне

ПК-12 - способностью проводить научные исследования для выработки стратегических решений в области ИКТ

Этап (уровень) освоения компетенции	**Планируемые результаты обучения** (показатели достижения заданного уровня освоения компетенций)	**Критерии оценивания результатов обучения**			
		2	**3**	**4**	**5**
Первый этап (уровень)	Знать: основные методы латентно-семантического анализа текста; основные алгоритмы интеллектуального анализа текста	Фрагментарные представления об основных методах латентно-семантического анализа текста; основных алгоритмах интеллектуального анализа текста	Неполные представления об основных методах латентно-семантического анализа текста; основных алгоритмах интеллектуального анализа текста	Сформированные, но содержащие отдельные пробелы представления об основных методах латентно-семантического анализа текста; основных алгоритмах интеллектуального анализа текста	Сформированные систематические представления об основных методах латентно-семантического анализа текста; основных алгоритмах интеллектуального анализа текста
Второй этап (уровень)	Уметь: применять процедуру TF-IDF для анализа главной темы; применять латентно-семантический анализ текста; применять наивный	Фрагментарные умения применять процедуру TF-IDF для анализа главной темы; применять латентно-семантический анализ текста, применять наивный байесовский классификатор для анализа текста	В целом успешное, но не систематическое умение применять процедуру TF-IDF для анализа главной темы; применять латентно-семантический анализ текста; применять наивный байесовский	В целом успешное, но содержащее отдельные пробелы в умении применять процедуру TF-IDF для анализа главной темы; применять латентно-семантический анализ текста;	Сформированное умение применять процедуру TF-IDF для анализа главной темы; применять латентно-семантический анализ текста; применять наивный байесовский классификатор для анализа текста

	байесовский классификатор для анализа текста		классификатор для анализа текста	применять наивный байесовский классификатор для анализа текста	
Третий этап (уровень)	Владеть: навыками тематического моделирования, используя инструменты алгоритма TF-IDF, и методов Word2Vec, Global2Vec	Фрагментарное владение навыками тематического моделирования, используя инструменты алгоритма TF-IDF, и методов Word2Vec, Global2Vec	В целом успешное, но не систематическое применение навыков тематического моделирования, используя инструменты алгоритма TF-IDF, и методов Word2Vec, Global2Vec	В целом успешное, но содержащее отдельные пробелы применение навыков тематического моделирования, используя инструменты алгоритма TF-IDF, и методов Word2Vec, Global2Vec	Успешное и систематическое применение навыков тематического моделирования, используя инструменты алгоритма TF-IDF, и методов Word2Vec, Global2Vec

4.2.Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и опыта деятельности, характеризующих этапы формирования компетенций

Этапы освоения	Результаты обучения	Компетенция	Оценочное средство
1-й этап Знания	основные принципы анализа неструктурированной информации (текстовой); метрики лингвостатистики; основные законы лингвостатистики; основные принципы разметки текста; способы векторного представления текста; способы кластеризации текста; основные методы латентно-семантического анализа текста	ПК-1	индивидуальное задание: доклад, кейс-задание, устный опрос
	основных методах латентно-семантического анализа текста; основных алгоритмах интеллектуального анализа текста	ПК-12	индивидуальное задание: доклад, кейс-задание, устный опрос
2-й этап Умения	проводить качественную чистку текста, избавляться от стоп-слов и проводить лемматизацию; создавать терм-документную матрицу двумя способами; использовать мешочек слов для анализа текста;	ПК-1	индивидуальное задание: кейс-задание, устный опрос
	применять процедуру TF-IDF для анализа главной темы; применять латентно-семантический анализ текста	ПК-12	индивидуальное задание: кейс-задание, устный опрос
3-й этап Владения (навыки / опыт деятельности)	Владение методами предподготовки к проведению анализа текста, используя средства среды R Studio; современными методами и подходами интеллектуального анализа текста на базовом уровне	ПК - 1	индивидуальное задание: кейс-задание,
	Владение навыками тематического моделирования, используя инструменты алгоритма TF-IDF, и методов Word2Vec, Global2Vec	ПК-12	индивидуальное задание: кейс-задание,

Оценка уровня освоения дисциплины осуществляется в виде текущего и промежуточного контроля успеваемости магистрантов, и на основе критериев оценки уровня освоения дисциплины.

Контроль представляет собой набор заданий и проводится в форме контрольных мероприятий по оцениванию фактических результатов обучения магистрантов и осуществляется ведущим преподавателем.

Формы и содержание текущего/рубежного контроля:

- контроль посещаемости занятий;
- своевременное выполнение кейс-задач;
- выборочная проверка ответов на вопросы самоконтроля;
- оценка уровня развития компетенций в ходе решения кейс-задач на реальных данных.

Магистранты допускаются к экзамену по дисциплине при условии сдачи всех рубежей и заданий, предусмотренных программами текущего контроля. Экзамен проводится в виде устного собеседования по учебному материалу дисциплины. Результат сдачи зачета оцениваются в ведомостях отметкой «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Критерии оценки зачета:

отметка	Описание
«зачтено»	Выставляется магистранту на основании анализа текущей успеваемости, если магистрант выполнил кейс-задания и подготовил отчет об их выполнении. Магистрант без затруднений ответил на все вопросы, связанные с выполнением кейс-задания.
«не зачтено»	Выставляется магистранту на основании анализа текущей успеваемости, если магистрант не выполнил кейс-задания и не подготовил отчет об их выполнении. Магистрант не может ответить на вопросы, связанные с выполнением кейс-задания.

Критерии оценки экзамена:

отметка	Описание
«отлично»	выставляется магистранту, если магистрант дал полные, развернутые ответы на все теоретические вопросы билета, продемонстрировал знание терминологии, основных элементов, умение применять теоретические знания при выполнении практических заданий. Магистрант без затруднений ответил на все дополнительные вопросы
«хорошо»	выставляется магистранту, если магистрант раскрыл в основном теоретические вопросы, однако допущены неточности в определении основных понятий. При ответе на дополнительные вопросы допущены небольшие неточности.
«удовлетворительно»	выставляется магистранту, если при ответе на теоретические вопросы магистрантом допущено несколько существенных ошибок в толковании основных понятий. Логика и полнота ответа страдают заметными изъянами. Заметны пробелы в знании основных методов.

	Теоретические вопросы в целом изложены достаточно, но с пропусками материала. Имеются принципиальные ошибки в логике построения ответа на вопрос.
«неудовлетворительно»	выставляется магистранту, если ответ на теоретические вопросы свидетельствует о непонимании и крайне неполном знании основных понятий и методов. Обнаруживается отсутствие навыков применения теоретических знаний при выполнении практических заданий.

Экзаменационные билеты

Экзамен является оценочным средством для всех этапов освоения компетенций.
В экзаменационном билете – 2 вопроса .

Образец экзаменационного билета:

Башкирский государственный университет

Институт экономики, финансов и бизнеса

Кафедра математических методов в экономики

Направление подготовки 38.04.05 «Бизнес-информатика»

Программа «Информационная бизнес-аналитика»

Дисциплина «Анализ неструктурированной информации»

Билет к экзамену № 1

1. Определение корпуса текста, разметка текста
2. Выделение главной темы с помощью алгоритма TF-IDF

Заведующий кафедрой Р.Х. Бахитова

Примерные вопросы для экзамена:

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

13 Дистрибутивный анализ: контекстные вектора
14 Косинусная мера сходства слов и словосочетаний
15 Мягкая косинусная мера сходства слов и словосочетаний
16 Векторное представление слов
17 Выделение главной темы с помощью алгоритма TF-IDF
18 Кластеризация текста методом k-средних
19 Латентно-семантический анализ: сравнение двух термов между собой
20 Латентно-семантический анализ: сравнение двух документов между собой
21 Латентно-семантический анализ: сравнение терма и документа
22 Инструмент Word2Vec: алгоритма обучения : CBOW (Continuous Bag of Words)
23 Инструмент Word2Vec: алгоритма обучения :Skip-gram
24 Инструмент Global2Vec:

Кейс-задания

Кейс-задача 1. Предподготовка текста, создание мешочка слов, кластеризация текста.

Цель: провести качественную чистку текста с последующей его кластеризацией.

Задание

1) сформировать корпус текста (не менее 1000 слов);
2) подготовить данные, удалить стоп-слова: все предлоги, местоимения, числа и т.п. с помощью функций removeWords и removeNumbers;
3) при анализе текста сформировать словарь дополнительных стоп-слов;
4) провести лемматизацию текста (с помощью пакета Mystem), приведя слова к единой словоформе;
5) создать текстовый корпус и матрицу TermDocument;
6) удалить пустые строки в тексте;
7) оценить проведённый анализ с помощью создания мешочка слов.
5) провести нормализацию векторов предобратонного текста.
6) провести кластеризацию предобратонного текста методом k-средних;
7) сделать вывод.

Выполнять кейс-задание рекомендуется с использованием среды R, используя библиотеки tm и wordcloud, RTextTools, Textcat.

Результатом выполнения кейс-задания является отчет №1.

Кейс-задача 2. Выделение главной темы с помощью алгоритма TF-IDF

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

Задание

1) сформировать корпус текста (на русском или английском языке не менее 1000 слов);
2) подготовить данные, а именно удалить все предлоги, стоп слова, местоимения, числа и т.д. с помощью функций removeWords и removeNumbers;
3) создать текстовый корпус и матрицу TermDocument;
4) провести подсчет частоты слов с помощью функции “Wordcloud”;

5) построить облако наиболее встречающихся слов;
6) подсчитать устойчивые выражения и провести процедуру коллокации;
7) провести TF – IDF оценку и отобразить важные слова в виде облако слов;
8) сделать вывод.

Результатом выполнения кейс-задания является отчет №2.

Кейс-задача 3. Применение наивного байесовского классификатора для анализа текста

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

Задание

1) Провести предварительную обработку с вводом текстового предложения
2) Создать список с индексами обучающих трейнов, проверочных и тестовых наборов
3) Сформировать терм-документную матрицу
4) Создать наивный байесовский классификатор спама на основе обучающей выборки
5) Вычислить класс входного предложения на основе полученной модели классификации на тестовой выборке
6) Построить матрицу сопряженности
7) Вычислить процент ошибки классификации
8) сделать вывод.

Результатом выполнения кейс-задания является отчет №3.

Примечание: для выполнения кейса использовать библиотеку в R quanteda-tools.

Кейс-задача 4. Проведение LSA-анализа, реализация инструмента Word2Vec

Цель: провести латентно-семантический анализ текста, используя инструмент Word2Vec с различной архитектурой.

Задание

1) сформировать корпус текста (на русском или английском языке не менее 1000 слов);
2) подготовить данные, а именно удалить все предлоги, стоп слова, местоимения, числа и т.д.;
3) создать текстовый корпус;
4) выделить слова по их окружению в режиме «skipgrams», используя инструмент Word2Vec
5) выделить слова вместе с их окружением в режиме CBOW, используя инструмент Word2Vec;
6) провести кластеризацию слов, используя инструмент Word2Vec;
7) найти семантически близкие слова, используя инструмент Word2Vec;
8) сделать вывод.

Результатом выполнения кейс-задания является отчет №4.

К отчетам по выполненным кейс-заданиям предъявляются следующие требования:

1. Четкое формулирование поставленной цели исследования
2. Формулирование задач, решение которых необходимо для достижения поставленной цели.
3. Описание в виде пунктов, тех действий, которые требуются для решения поставленных задач. Все рисунки и таблицы последовательно нумеруются и описываются. Каждый

пункт решения поставленных задач сопровождается анализом принятого решения. При проведении статистических тестов, обязательно выписывается нулевая и альтернативная гипотеза, а также уравнение, обосновывающее тест, формулируется принятие решения на обосновано выбранном уровне значимости, указывается критическая область отказа от нулевой гипотезы в пользу альтернативной.

4. В заключении приводятся результаты проведенного анализа неструктурированной информации либо в виде кластеров текста, либо в виде списка семантически близких слов или выражений. В зависимости от цели выполнения кейс-задания..

Критерии оценки кейс заданий:

- оценка «зачтено» выставляется магистранту, при выполнении задания без существенных замечаний.
- оценка «не зачтено» выставляется магистранту, при наличие существенных замечаний.

Темы докладов:

1. Применение инструментов TextMining для поиска информации
2. Применение инструментов TextMining для предварительной обработки информации
3. Применение инструментов TextMining для извлечения информации
4. Средства анализа текстовой информации Oracle — Oracle Text2
5. Средства анализа текстовой информации от IBM Intelligent Miner for Text
6. Средства анализа текстовой информации SAS Institute — Text Miner
7. Метод анализа текстовой информации, основанный на терминах (Term Based Method (TBM)).
8. Метод анализа текстовой информации, основанный на фразах (Phrase Based Method (PBM)).
9. Метод анализа текстовой информации, основанный на концепциях или понятиях (Concept Based Method (CBM)).
10. Метод анализа текстовой информации шаблонной систематики (Pattern Taxonomy Method (PTM)).
11. Метод анализа текстовой информации Term Based Method

Требование к докладу

1. Доклад должен отражать только актуальные тенденции интеллектуального анализа текста
2. При подготовке к докладу должно быть использовано не менее 5 источников информации
3. Доклад должен занимать по объемы от 6-8 минут, и 2 минуты отводится на дискуссию и вопросы
4. Доклад должен сопровождаться презентацией 7-10 слайдов

Критерии оценки докладов:

- оценка «зачтено» выставляется магистранту, при выполнении всех требований к докладу, а также умение аргументировано отвечать на задаваемые по теме доклада вопросы.
- оценка «не зачтено» выставляется магистранту, при невыполнении всех требований к докладу, а также неумение аргументировано отвечать на задаваемые по теме доклада вопросы.

–

5. Учебно-методическое и информационное обеспечение дисциплины

5.1. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины

Основная литература

1. Анализ данных : учебник для академического бакалавриата / ГУ - Высшая школа экономики; под 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.)

5.2. Перечень ресурсов информационно-телекоммуникационной сети «Интернет» и программного обеспечения, необходимых для освоения дисциплины

Пользователям библиотеки БашГУ предоставляется возможность использования следующих электронных информационных ресурсов:

№	Наименование Интернет-ресурса	Ссылка (URL) на Интернет ресурс
1.	Федеральная служба государственной статистики	www.gks.ru
2.	Министерство финансов РФ	www.minfin.ru
3.	Международный валютный фонд	www.imf.org
4.	Центр макроэкономического анализа и краткосрочного прогнозирования	www.forecast.ru
5.	Территориальный орган Федеральной службы государственной статистики по РБ	www.bashstat.ru
6.	Информационно-издательский центр «Статистика России»	www.infostat.ru
7.	Информационно-аналитический сайт в области информационных технологий	citforum.ru
8.	Издание о высоких технологиях	cnews.ru
9.	Библиотека Г. Верникова – все о менеджменте и IT - подборка аналитических материалов по вопросам экономики, менеджмента и информационных технологий.	vernikov.ru
10.	Официальный портал ИТ-директоров (Реестр ИТ-поставщиков)	globalcio.ru
11.	Журнал CIO – руководитель информационной службы	cio-world.ru
12.	Единый архив экономических и социологических данных ВШЭ	http://sophist.hse.ru/

1. База данных периодических изданий на платформе EastView: «Вестники Московского университета», «Издания по общественным и гуманитарным наукам» - https://dlib.eastview.com/
2. Информационная система «Единое окно доступа к образовательным ресурсам» - http://window.edu.ru
3. Научная электронная библиотека eLibrary.ru - http://elibrary.ru/defaultx.asp
4. Справочно-правовая система Консультант Плюс - http://www.consultant.ru/
5. Электронная библиотечная система «Университетская библиотека онлайн» – https://biblioclub.ru/
6. Электронная библиотечная система «ЭБ БашГУ» – https://elib.bashedu.ru/
7. Электронная библиотечная система издательства «Лань» – https://e.lanbook.com/
8. Электронный каталог Библиотеки БашГУ – http://www.bashlib.ru/catalogi.

9. Архивы научных журналов на платформе НЭИКОН (Cambridge University Press, SAGE Publications, Oxford University Press) - https://archive.neicon.ru/xmlui/

10. Издательство «Annual Reviews» - https://www.annualreviews.org/

11. Издательство «Taylor&Francis» - https://www.tandfonline.com/

12. Windows 8 Russian. Windows Professional 8 Russian Upgrade Договор №104 от 17.06.2013 г. Лицензии бессрочные.

13. Microsoft Office Standard 2013 Russian. Договор №114 от 12.11.2014 г. Лицензии бессрочные.

14. Windows 8 Russian. Windows Professional 8 Russian Upgrade. Договор № 104 от 17.06.2013 г. Лицензии бессрочные.

6. Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине

Наименование специализированных аудиторий, кабинетов, лабораторий	Вид занятий	Наименование оборудования, программного обеспечения
учебная аудитория для проведения занятий лекционного типа: лаборатория социально-экономического моделирования № 107 (помещение, ул.Карла Маркса, д.3, корп.4), лаборатория анализа данных № 108 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 110 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 111 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 114 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 122 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 204 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 207 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 208 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 209 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 210 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 212 (гуманитарный корпус), аудитория № 213 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 218 (гуманитарный корпус), аудитория № 220 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 221 (гуманитарный корпус), аудитория № 222 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 301 (гуманитарный корпус), аудитория № 305 (гуманитарный корпус), аудитория № 307 (гуманитарный корпус), аудитория № 308 (гуманитарный корпус), аудитория № 309 (гуманитарный корпус), аудитория № 110 (гуманитарный корпус), лаборатория исследования процессов в экономике и управлении № 311а (гуманитарный корпус), лаборатория информационных технологий в экономике и управлении № 311в (гуманитарный корпус).).	Лекции	**лаборатория социально-экономического моделирования № 107:** учебная мебель, доска, проекционный экран с электродиодом lumien master control, проектор casio, персональный компьютер пэвм кламас в комплекте – 18 шт. **лаборатория анализа данных № 108:** учебная мебель, доска, персональный компьютер пэвм кламас в комплекте – 17 шт. **аудитория № 110:** учебная мебель, доска, телевизор led. **аудитория № 111:** учебная мебель, доска, телевизор led. **аудитория № 114:** учебная мебель, доска. **аудитория № 115:** учебная мебель, колонки (2 шт.), динамики, dvd плеер tochiba, магнитола sony (4 шт.) **аудитортия №118:** учебная мебель, проектор benq, колонки (2 шт.), музыкальный центр lg, флипчарт магнитно-маркерный на треноге **аудитория № 122:** учебная мебель, доска. **аудитория № 204:** учебная мебель, доска, проекционный экран с электродиодом lumien master control, проектор casio. **аудитория № 207:** учебная мебель, доска, телевизор led tcl. **аудитория № 208:** учебная мебель, доска, телевизор led tcl. **аудитория № 209:** учебная мебель, доска. **аудитория № 210:** учебная мебель, доска. **аудитория № 212:** учебная мебель, доска, проектор infocus. **аудитория № 213:** учебная мебель, доска, проекционный экран с электродиодом lumien master control, проектор casio. **аудитория № 218:** учебная мебель, доска, мультимедиа-проектор infocus. **аудитория № 220:**
учебная аудитория для проведения занятий семинарского типа: лаборатория социально-экономического моделирования № 107 (помещение, ул.Карла Маркса, д.3, корп.4), лаборатория анализа данных № 108 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 110 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 111 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 114 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 122 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 204 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 207	Практические/семинарские занятия	

(помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 208 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 209 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 210 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 212 (гуманитарный корпус), аудитория № 213 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 218 (гуманитарный корпус), аудитория № 220 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 221 (гуманитарный корпус), аудитория № 222 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 301 (гуманитарный корпус), аудитория № 305 (гуманитарный корпус), аудитория № 307 (гуманитарный корпус), аудитория № 308 (гуманитарный корпус), аудитория № 309 (гуманитарный корпус), аудитория № 110 (гуманитарный корпус), лаборатория исследования процессов в экономике и управлении № 311а (гуманитарный корпус), лаборатория информационных технологий в экономике и управлении № 311в (гуманитарный корпус).		учебная мебель, доска. **аудитория № 221** учебная мебель, доска. **аудитория № 222** учебная мебель, доска. **аудитория № 301** учебная мебель, экран на штативе, проектор aser. **аудитория № 302** учебная мебель, персональный компьютер в комплекте hp, моноблок, персональный компьютер в комплекте моноблок iru. **аудитория № 305** учебная мебель, доска, проектор infocus. **аудитория № 307** учебная мебель, доска. **аудитория № 308** учебная мебель, доска. **аудитория № 309** учебная мебель, доска. **лаборатория исследования процессов в экономике и управлении № 311а** учебная мебель, доска, персональный компьютер lenovo thinkcentre – 16 шт. **лаборатория информационных технологий в экономике и управлении № 311в** учебная мебель, доска, персональный компьютер в комплекте № 1 iru corp 510 – 14 шт. **аудитория № 312** учебная мебель, доска. 1. Windows 8 Russian. Windows Professional 8 Russian Upgrade. Договор № 104 от 17.06.2013 г. Лицензии – бессрочные. Microsoft Office Standard 2013 Russian. Договор № 114 от 12.11.2014 г. Лицензии – бессрочные.
учебная аудитория для проведения групповых и индивидуальных консультаций: лаборатория социально-экономического моделирования № 107 (помещение, ул.Карла Маркса, д.3, корп.4), лаборатория анализа данных № 108 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 110 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 111 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 114 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 122 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 204 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 207 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 208 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 209 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 210 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 212 (гуманитарный корпус), аудитория № 213 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 218 (гуманитарный корпус), аудитория № 220 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 221 (гуманитарный корпус), аудитория № 222 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 301 (гуманитарный корпус), аудитория № 305 (гуманитарный корпус), аудитория № 307 (гуманитарный корпус), аудитория № 308 (гуманитарный корпус), аудитория № 309 (гуманитарный корпус), аудитория № 110 (гуманитарный корпус), лаборатория исследования процессов в экономике и управлении № 311а	Групповые и индивидуальные консультаций	

(гуманитарный корпус), лаборатория информационных технологий в экономике и управлении № 311в (гуманитарный корпус).		
учебная аудитория для текущего контроля и промежуточной аттестации: лаборатория социально-экономического моделирования № 107 (помещение, ул.Карла Маркса, д.3, корп.4), лаборатория анализа данных № 108 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 110 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 111 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 114 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 122 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 204 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 207 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 208 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 209 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 210 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 212 (гуманитарный корпус), аудитория № 213 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 218 (гуманитарный корпус), аудитория № 220 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 221 (гуманитарный корпус), аудитория № 222 (помещение, ул.Карла Маркса, д.3, корп.4), аудитория № 301 (гуманитарный корпус), аудитория № 305 (гуманитарный корпус), аудитория № 307 (гуманитарный корпус), аудитория № 308 (гуманитарный корпус), аудитория № 309 (гуманитарный корпус), аудитория № 110 (гуманитарный корпус), лаборатория исследования процессов в экономике и управлении № 311а (гуманитарный корпус), лаборатория информационных технологий в экономике и управлении № 311в (гуманитарный корпус).	Текущий контроль и промежуточная аттестация	
помещения для самостоятельной работы: аудитория № 302 читальный зал (гуманитарный корпус).	Самостоятельная работа	