

МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «ВГУ»)

УТВЕРЖДАЮ
Заведующий кафедрой

Математических методов исследования операций



Азарнова Т.В.

18 апреля 2025 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Б1.В.ДВ.01.02 Автоматическая обработка и анализ текстовых данных

1. Код и наименование направления подготовки/специальности:

09.04.03 Прикладная информатика

2. Профиль подготовки / специализация/магистерская программа:

Прикладная информатика в социальных и медицинских системах

3. Квалификация (степень) выпускника: магистр

4. Форма образования: очная

5. Кафедра, отвечающая за реализацию дисциплины:

Математических методов исследования операций

6. Составители программы: Замятин Игорь Викторович, к. ф.-м. наук

7. Рекомендована:

Научно-методическим советом факультета прикладной математики, информатики и механики

Протокол о рекомендации:

№6 от 17.03.2025

8. Учебный год: 2026/2027

Семестр(-ы): 4

9. Цели и задачи учебной дисциплины:

Целью курса является формирование у студентов представления о типах задач, возникающих в области обработки естественного языка (Natural Language Processing – NLP) и методах их решения, которые помогут студентам выявлять, формализовать и успешно решать практические задачи анализа текстовых данных, возникающие в процессе их профессиональной деятельности.

Задачи курса:

1. Познакомиться с типами задач обработки естественного языка (NLP).
2. Ознакомиться с основными моделями машинного обучения, используемыми для решения задач NLP.
3. Научиться выявлять, формализовать и решать практические задачи NLP.
4. Ознакомиться с основными фреймворками для решения задач NLP и научиться применять на практике.

10. Место учебной дисциплины в структуре ООП:

Дисциплина относится к вариативной части дисциплин по выбору. Для освоения дисциплины необходимы знания, умения и навыки, сформированные в ходе освоения дисциплины «Машинное обучение», «Современные нейросетевые технологии».

11. Планируемые результаты обучения по дисциплине/модулю (знания, умения, навыки), соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями) и индикаторами их достижения:

Код	Название компетенции	Код(ы)	Индикатор(ы)	Планируемые результаты обучения
ПК-1	Управление работами по сопровождению и проектами создания (модификации) ИС, автоматизирующих задачи организационного управления и бизнес-процессы.	ПК-1.3	Организационное и технологическое обеспечение проектирования и дизайна ИС	Знать: - современные подходы к разработке прикладных решений в области искусственного интеллекта. Уметь: - применять инструментальные среды и программно-технические платформы для решения задач в области искусственного интеллекта; - руководить проектами по созданию, внедрению и поддержке систем искусственного интеллекта. Владеть: - навыками построения и проверки качества моделей машинного обучения; - навыками планирования и постановки задач в области искусственного интеллекта.
ПК-2	Непосредственное руководство процессами разработки программного обеспечения	ПК-2.3	Применяет методы прикладной информатики для решения профессиональных задач в области обработки различных типов информации	Знать: - принципы построения систем обработки естественного языка, методы и подходы к планированию и реализации проектов по их созданию. Уметь: - применять методы машинного обучения при решении задач обработки естественного языка. Владеть: - методами анализа текстовых данных.

12. Объем дисциплины в зачетных единицах/часах в соответствии с учебным планом — 4/144

Форма промежуточной аттестации – зачет.

13. Трудоемкость по видам учебной работы

Вид учебной работы		Трудоемкость (часы)		
		Всего	В том числе в интерактивной форме	По семестрам
				4 сем.
Аудиторные занятия		32	16	32
в том числе:	лекции	16		16
	практические			
	лабораторные	16	16	16
Самостоятельная работа		112		112
Итого:		144	16	144
Форма промежуточной аттестации (зачет – 6 ч., курс. работа – 3 ч.)				

13.1 Содержание дисциплины:

№ п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
1. Лекции			
1.1	Задачи обработки естественного языка.	Классификация задач NLP. Сферы применения NLP.	Б1.В.ДВ.01.02 Автоматическая обработка и анализ текстовых данных
1.2	Анализ структуры текста. Предобработка текста.	Ключевые термины. Морфологический, синтаксический анализ. Методы предобработки текста (удаление стоп-слов, нормализация, токенизация).	
1.3	Модели представления текста.	n-граммы, skip-граммы. Bag of Words (BoW) и CBoW. Word Embeddings, методы Word2Vec и GloVe.	
1.4	Задачи классификации текстов. Анализ тональности.	Метрики оценки. Категоризация с помощью моделей кластеризации. Классификация текстов с помощью деревьев решений, SVM, Байесовского классификатора, нейронных сетей.	
1.5	Тематическое моделирование.	Моделирование тем (TM), распределенное представление тем (DTM), анализ отношений тем (TR), латентное размещение Дирихле (LDA).	
1.6	Генерация текста. Языковые модели.	Архитектуры BERT и GPT. Трансформеры. Метрики оценки.	
1.7	Чат-боты.	Основные термины. Логика обработки запросов. Генерация ответов. NLU и NLG.	
3. Лабораторные работы.			
3.1	Задачи обработки естественного языка. Предобработка текста.	Работа с текстом. Парсинг. Инструменты для морфологического, синтаксического анализа. Методы предобработки текста (удаление стоп-слов, нормализация, токенизация). Библиотеки ruMorphy, nltk, BeautifulSoup.	Б1.В.ДВ.01.02 Автоматическая обработка и анализ текстовых данных
3.2	Модели представления текста.	Инструменты векторизации текста. n-граммы, skip-граммы. Bag of Words (BoW) и CBoW. Word Embeddings, методы Word2Vec и GloVe. Библиотека scikit-learn. Библиотека Gensim.	
3.3			
3.4	Задачи классификации текстов. Анализ тональности.	Категоризация с помощью моделей кластеризации. Классификация текстов с помощью деревьев решений, SVM, Байесовского классификатора, нейронных сетей.	
3.5	Тематическое моделирование.	Моделирование тем (TM), распределенное представление тем (DTM), анализ отношений тем (TR), латентное размещение Дирихле (LDA).	

№ п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
3.6	Генерация текста. Языковые модели.	Архитектуры BERT и GPT. Трансформеры. Метрики оценки.	
3.7	Чат-боты.	Основные термины. Логика обработки запросов. Генерация ответов. NLU и NLG.	

13.2 Темы (разделы) дисциплины и виды занятий:

№ п/п	Наименование раздела дисциплины	Виды занятий (часов)				
		Лекции	Практические	Лабораторные	Самостоятельная работа	Всего
1	Задачи обработки естественного языка.	2		0	14	16
2	Анализ структуры текста. Предобработка текста.	2		3	14	19
3	Модели представления текста.	2		3	14	19
4	Задачи классификации текстов. Анализ тональности.	2		2	14	18
5	Тематическое моделирование.	2		2	14	18
6	Генерация текста. Языковые модели.	4		4	28	36
7	Чат-боты.	2		2	14	18
	Итого:	16		16	112	144

14. Методические указания для обучающихся по освоению дисциплины

Работа с конспектами занятий, презентациями, выполнение практических заданий для самостоятельной работы, выполнение лабораторных работ, использование рекомендованной литературы и методических материалов, в том числе размещенных на странице курса «Б1.В.ДВ.01.02 Автоматическая обработка и анализ текстовых данных» на портале «Электронный университет ВГУ» <https://edu.vsu.ru/course/view.php?id=25866>, автор Замятин И.В.

В рамках общего объема часов, отведенных для изучения дисциплины, предусматривается выполнение следующих видов самостоятельных работ студентов (СРС): изучение теоретического материала, написание программ по темам, изученным на лабораторных занятиях. К промежуточной аттестации (зачет) учащимися выполняются лабораторные работы.

При использовании дистанционных образовательных технологий и электронного обучения необходимо выполнять все указания преподавателей по работе на LMS-платформе, своевременно подключаться к online-занятиям, соблюдать рекомендации по организации самостоятельной работы.

15. Перечень основной и дополнительной литературы, ресурсов интернет, необходимых для освоения дисциплины:

а) основная литература:

№ п/п	Источник
1	Ганегедара, Т. Обработка естественного языка с TensorFlow : руководство / Т. Ганегедара ; перевод с английского В. С. Яценкова. — Москва : ДМК Пресс, 2020. — 382 с. — ISBN 978-5-97060-756-5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/140584 (дата обращения: 11.10.2024).
2	Bird, S. Natural Language Processing with Python / S. Bird. — O'Reilly Media, 2016. — 512 с. —

	ISBN 978-1-49191-342-0. — Текст: электронный — URL: https://www.nltk.org/book/ (дата обращения: 11.10.2024).
3	Йылдырым, С. . Осваиваем архитектуру Transformer. Разработка современных моделей с помощью передовых методов обработки естественного языка / М. . Асгари-Ченаглу; С. . Йылдырым .— пер. с англ. В. С. Яценкова. — Москва : ДМК Пресс, 2022 .— 321 с. : ил. — ISBN 978-1-80107-765-1 (англ.) .— ISBN 978-5-93700-106-1 (рус.) .— URL: https://rucont.ru/efd/810502 (дата обращения: 02.07.2025)

б) дополнительная литература:

№ п/п	Источник
4	Обработка естественного языка с использованием языка программирования Python : учебное пособие : в 2 частях / составитель А. Б. Мантусов. — Элиста : КГУ, 2022 — Часть 1 — 2022. — 56 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/360923 (дата обращения: 11.10.2024).
5	Гольдберг, Й. Нейросетевые методы в обработке естественного языка : руководство / Й. Гольдберг ; перевод с английского А. А. Слинкина. — Москва : ДМК Пресс, 2019. — 282 с. — ISBN 978-5-97060-754-1. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/131704 (дата обращения: 11.10.2024).
6	Митчелл, Р. Скрапинг веб-сайтов с помощью Python : руководство / Р. Митчелл ; перевод с английского А. В. Груздев. — Москва : ДМК Пресс, 2016. — 280 с. — ISBN 978-5-97060-223-2. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/100903 (дата обращения: 11.10.2024).
7	Jurafsky, D. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition with Language Models [Электронный ресурс] : / D. Jurafsky, J. H. Martin. — 2024. — 591 с. — Режим доступа: https://web.stanford.edu/~jurafsky/slp3/

в) информационные электронно-образовательные ресурсы (официальные ресурсы интернет):

№ п/п	Источник
8	Электронная библиотечная система «Издательства «Лань». Режим доступа: http://e.lanbook.com/
9	Автоматическая обработка и анализ текстовых данных / И.В. Замятин — Образовательный портал «Электронный университет ВГУ». — Режим доступа: https://edu.vsu.ru/course/view.php?id=25866
10	Электронная библиотечная система ВГУ. Режим доступа: http://www.lib.vsu.ru

16. Перечень учебно-методического обеспечения для самостоятельной работы

Самостоятельная работа обучающегося должна включать подготовку к лабораторным занятиям, выполнение текущих заданий по освоению соответствующих тем курса, выполнение курсовой работы и подготовку к промежуточной аттестации. Для этого рекомендуется освоить теоретический материал соответствующих тем по конспектам лекций, презентационному материалу, размещенному на ЭО ресурсах, литературу из представленного перечня, материалы с тематических ресурсов сети Интернет.

№ п/п	Источник
1	Автоматическая обработка и анализ текстовых данных / И.В. Замятин — Образовательный портал «Электронный университет ВГУ». — Режим доступа: https://edu.vsu.ru/course/view.php?id=25866
2	Электронная библиотечная система «Издательства «Лань». Режим доступа: http://e.lanbook.com/
3	Электронная библиотечная система ВГУ. Режим доступа: http://www.lib.vsu.ru

17. Информационные технологии, используемые для реализации учебной дисциплины, включая программное обеспечение и информационно-справочные системы (при необходимости)

Python 3 с подключенными библиотеками (дистрибутив Anaconda).

Дисциплина реализуется с применением электронного обучения и дистанционных образовательных технологий. Для организации самостоятельной работы обучающихся

используется онлайн-«Б1.В.ДВ.01.02 Автоматическая обработка и анализ текстовых данных», размещенный на платформе Электронного университета ВГУ (LMS moodle), а также Интернет-ресурсы, приведенные в п.15в.

18. Материально-техническое обеспечение дисциплины:

Лекционная аудитория должна быть оснащена современным компьютером с подключенным к нему проектором на настенный экран. Лабораторные занятия должны проводиться в специализированной аудитории, оснащенной современными персональными компьютерами и программным обеспечением в соответствии с тематикой изучаемого материала.

19. Оценочные средства для проведения текущей и промежуточной аттестаций

Порядок оценки освоения обучающимися учебного материала определяется содержанием следующих разделов дисциплины:

№ п/п	Наименование раздела дисциплины (модуля)	Компетенция(и)	Индикатор(ы) достижения компетенции	Оценочные средства
1.	Задачи обработки естественного языка.	ПК-2	ПК-2.3	Задание для Лабораторной работы №1
2.	Анализ структуры текста. Предобработка текста.			
3.	Модели представления текста.			
4.	Задачи классификации текстов. Анализ тональности.			
5.	Тематическое моделирование.			
6.	Генерация текста. Языковые модели.	ПК-1	ПК-1.3	Задание для Лабораторной работы №2
7.	Чат-боты.			
Промежуточная аттестация форма контроля – зачет				Темы для курсовых работ Список вопросов к экзамену

20 Типовые оценочные средства и методические материалы, определяющие процедуры оценивания

20.1 Текущий контроль успеваемости

Контроль успеваемости по дисциплине осуществляется с помощью следующих оценочных средств:

Лабораторная работа № 1.

Построить и обучить модель, которая (на выбор):

- сможет определять принадлежит ли предложение из текста определенному автору (или НЕ принадлежит). В качестве исходных данных использовать текст любого произведения (прозы) любого автора с сайта www.lib.ru (либо из другого источника)

ЛИБО

- сможет определить тему сообщения из новостной ленты (можно использовать любой новостной RSS-канал)

Последовательность решения:

1. Выбрать любого писателя, авторство которого будет проверяться с помощью модели. Все студенты выбирают разных авторов.

ЛИБО

Выбрать любой RSS-канал новостей.

2. Загрузить любое произведение выбранного автора (ЛИБО новости по выбранной теме), а также два произведения двух других авторов, как пример ДРУГОГО авторства (ЛИБО новости по двум другим темам). В качестве выборки использовать массив предложений из выбранных произведений (ЛИБО новостные сообщения целиком).

3. Построить модель, способную определить авторство по отдельному предложению (ЛИБО теме новостного сообщения).

Лабораторная работа № 2.

Разработать и реализовать последовательность шагов по созданию эффективной стратегии формирования запросов-промpts для LLM с интеграцией RAG.

1. Подготовка корпуса данных

Создать небольшой набор документов (примерно 5–7 статей или фрагментов текста), связанных с выбранной темой (например, технологический прогресс в области искусственного интеллекта, новые открытия в медицине и т.п.). Это будут внешние источники, из которых система должна извлекать дополнительную информацию.

2. Формирование структуры запроса

Для каждого подготовленного документа разработать два-три уникальных промpta-запроса разной сложности. Они должны различаться степенью детализации и объемом запрашиваемой информации. Например:

- Простой запрос: «Краткое резюме статьи о достижениях AI».
- Сложный запрос: «Проанализируй преимущества и недостатки использования искусственного интеллекта в медицинской диагностике, приведя конкретные примеры из текста».

При составлении запросов учитывать особенности формата, необходимого для RAG-модели:

- Включить название или ключевое слово документа для помощи в поиске.
- Предоставлять необходимые ориентиры и критерии отбора информации.

3. Проверка и оценка эффективности запросов

Используя подготовленные документы и разработанные запросы, проведите тестирование, отправив промpts в систему и получив ответы. Проверьте следующие аспекты:

- Насколько точно выбраны соответствующие фрагменты из исходных документов?
- Соответствует ли полученный ответ поставленной задаче?
- Есть ли проблемы с интерпретацией заданных критериев отбора информации?

Оцените работу системы по трем параметрам:

1. Релевантность извлеченной информации.
2. Полнота и информативность ответа.
3. Скорость и удобство взаимодействия.

4. Оптимизация запросов

Проведите сравнительный анализ работы модели с каждым запросом и предложите способы улучшения изначальных промpts. Например, добавьте уточняющую инструкцию по фильтрации или введите дополнительные ключевые слова для повышения точности выбора фрагмента.

5. Итоговая проверка и оформление отчета

Подготовьте отчет, содержащий:

- Исходные тексты документов.
- Подготовленные промpts и итоговые ответы модели.

- Таблицу сравнения характеристик различных версий запросов.
- Заключение о достоинствах и недостатках RAG-подхода и выводы по дальнейшему совершенствованию запросов.

20.2 Промежуточная аттестация

Промежуточная аттестация по дисциплине осуществляется с помощью следующих оценочных средств: Собеседование по вопросам к зачету. Контрольно-измерительные материалы промежуточной аттестации включают в себя теоретические вопросы, позволяющие оценить уровень полученных знаний и практические задания, позволяющие оценить степень сформированности умений и навыков.

Промежуточная аттестация проводится в соответствии с Положением о промежуточной аттестации обучающихся по программам высшего образования.

Для оценивания результатов обучения на зачете используются следующие показатели:

- 1) знание учебного материала и владение понятийным аппаратом теории NLP;
- 2) умение анализировать текстовые данные;
- 3) умение применять методы машинного обучения при решении задач классификации текстов;
- 4) владение навыками использования библиотек языка Python для построения систем обработки естественного языка;
- 6) владение навыками работы с LLM и RAG.

По учебному плану предусмотрен зачет.

Критерии оценки «зачтено» — продемонстрировано знание теоретического материала, положительные результаты решения лабораторных работ.

Соотношение показателей, критериев и шкалы оценивания результатов обучения.

Критерии оценивания компетенций	Уровень сформированности компетенций	Шкала оценок
Обучающийся владеет понятийным аппаратом данной области науки (теоретическими основами дисциплины), сдал все лабораторные работы.	Базовый уровень	Зачтено
Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки, не сдал хотя бы одну лабораторную работу.	—	Не зачтено

20.3 Фонд оценочных средств сформированности компетенций студентов, рекомендуемый для проведения диагностических работ

Какой (или какие) из перечисленных методов сохраняют семантику слов в процессе их векторизации?

#	Ответы	Оценка
---	--------	--------

#	Ответы	Оценка
A.	tf-idf	-50
B.	BoW (мешок слов)	-50
C.	Word2Vec	50
D.	GloVe	50

Какой (или какие) из перечисленных методов используют нейронные сети для кодирования слов?

#	Ответы	Оценка
A.	tf-idf	-50
B.	BoW (мешок слов)	-50
C.	Word2Vec	50
D.	GloVe	-50
E.	FastText	50

Какой (или какие) из перечисленных методов используют частоты вхождения слов в корпус при кодировании слов?

#	Ответы	Оценка
A.	tf-idf	33.3
B.	BoW (мешок слов)	33.3
C.	Word2Vec	-33.3
D.	GloVe	33.3
E.	FastText	-33.3

Какой (или какие) из перечисленных методов могут использовать при кодировании символьные n-граммы?

#	Ответы	Оценка
A.	tf-idf	33.3
B.	BoW (мешок слов)	33.3
C.	Word2Vec	-33.3
D.	GloVe	-33.3
E.	FastText	33.3

Для получения каких векторных представлений используется процесс обучения с учителем?

#	Ответы	Оценка
A.	tf-idf	0
B.	BoW (мешок слов)	0
C.	Word2Vec	50
D.	GloVe	0
E.	FastText	50

Tf-idf - это

#	Ответы	Оценка
---	--------	--------

#	Ответы	Оценка
A.	мера оценки важности слова в контексте документа, являющегося частью текстового корпуса	100
B.	алгоритм классификации текстов	0
C.	метрика ошибки в задачах классификации текстов	0

какая метрика используется в Word2Vec для определения близости векторных представлений слов?

#	Ответы	Оценка
A.	Косинусное расстояние	100
B.	MSE	0
C.	RMSE	0
D.	MAE	0

Какой из данных методов не используется для классификации текстов

#	Ответы	Оценка
A.	Наивный Байес	0
B.	Линейная регрессия	100
C.	Логистическая регрессия	0
D.	Градиентный бустинг	0

Что такое тематическое моделирование?

#	Ответы	Оценка
A.	Выделение в тексте кластеров слов, объединенных общей темой и формирование списка тем для данного текста	100
B.	Метод кодирования текстовой информации	0
C.	Процедура разбиения текста на отдельные слова	0
D.	Определение того, к какой единственной теме относится весь анализируемый текст	0

Выделите отличия между технологиями W2V и FastText

#	Ответы	Оценка
A.	W2V использует контекстное кодирование, а FastText - частотное	0
B.	W2V использует для кодирования архитектуру skip-gram, а FastText - CBOW	0
C.	fastText использует символьные n-граммы, а W2V - нет	100
D.	astText использует негативное сэмплирование, а W2V - нет	0

Что такое негативное семплирование?

#	Ответы	Оценка
---	--------	--------

#	Ответы	Оценка
A.	Это метод, добавляющий в выборку для обучения векторной модели отрицательные примеры, то есть пары слов, не являющихся соседями по контексту	100
B.	Это метод, при котором компоненты векторной модели после кодирования принимают отрицательные значения	0
C.	Это метод, который кодирует близкими векторами слова, не похожие по контексту	0
D.	Это метод сокращения обучающей выборки	0

Выделяют такие этапы построения модели “мешок слов”:

#	Ответы	Оценка
A.	токенизация	33.3
B.	построение словаря	33.3
C.	моделирование тем	0
D.	формирование разреженной матрицы	33.3

Для улучшения качества моделей обработки естественного языка принято:

#	Ответы	Оценка
A.	Искать последовательности символов, которые состоят как минимум из двух букв или цифр и отделены друг от друга границами слов.	33.3
B.	Использовать только те токены, которые встречаются по крайней мере в двух документах	33.3
C.	Исключать слова, которые встречаются слишком часто	33.3

На этапе предобработки текста применяют методы:

#	Ответы	Оценка
A.	Преобразование всех слов к строчному регистру	33.3
B.	Подключение списка стоп-слов	33.3
C.	Лемматизация	33.3
D.	TF-IDF	-33.3

Метод, в результате которого производится масштабирование признаков в зависимости от степени их информативности, называется:

#	Ответы	Оценка
A.	tf-idf	100
B.	стемминг	0
C.	Doc2Vec	0

Что такое n-граммы в задачах анализа текстов?

#	Ответы	Оценка
---	--------	--------

#	Ответы	Оценка
A.	последовательности токенов, добавляемые в словарь	100
B.	объединение в словаре двух слов с одной и той же основой	0
C.	последовательность символов, стоящих рядом в тексте	0

Что такое биграмма в задачах анализа текстов?

#	Ответы	Оценка
A.	устойчивое сочетание из двух соседних слов, добавляемое в словарь	100
B.	объединение в словаре двух слов с одной и той же основой	0
C.	два символа, стоящих рядом в тексте	0

Что такое лемматизация?

#	Ответы	Оценка
A.	мера оценки важности слова в контексте документа, являющегося частью текстового корпуса	0
B.	процесс приведения слова к нормальной словарной форме	100
C.	процесс выделения основы слова	0

Что такое мешок слов?

#	Ответы	Оценка
A.	представление текстовой информации в виде набора слов, без учета структуры исходного текста	0
B.	Метод векторизации, который кодирует текстовый корпус в виде разреженной матрицы, каждая строка в которой соответствует отдельному документу, а каждый столбец — определенному слову	100
C.	способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов	0

Что такое стемминг?

#	Ответы	Оценка
A.	мера оценки важности слова в контексте документа, являющегося частью текстового корпуса	0
B.	процесс приведения слова к нормальной словарной форме	0
C.	процесс выделения основы слова	100

Что такое токенизация в машинном обучении?

#	Ответы	Оценка
A.	разбиение анализируемого текста на отдельные слова (или устойчивые сочетания слов)	100
B.	формирование словаря на основе текстового корпуса	0
C.	алгоритм кодирования текстовых данных	0

Задания раздела 20.3 рекомендуются к использованию при проведении диагностических работ с целью оценки остаточных результатов освоения данной дисциплины (знаний, умений, навыков).