МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»
(РУТ (МИИТ)

Рабочая программа дисциплины (модуля),
как компонент образовательной программы
высшего образования - программы магистратуры
по направлению подготовки
09.04.01 Информатика и вычислительная техника,
утвержденной первым проректором РУТ (МИИТ)
Тимониным В.С.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Информационный поиск и анализ текстов

Направление подготовки: 09.04.01 Информатика и вычислительная техника

Направленность (профиль): Искусственный интеллект и предиктивная аналитика в транспортных системах

Форма обучения: Очная

1. Общие сведения о дисциплине (модуле).

Цель дисциплины «Информационный поиск и анализ текстов» заключается в освоении принципов интеллектуального анализа текста в задачах информационного поиска и извлечения информации.

В рамках дисциплины формируются знания о компьютерной лингвистике, способах предварительной обработки и преобразования языка, определении связи между словами, построении моделей классификации текста и структуре и особенностях построения поисковых систем.

На лабораторных занятиях у обучающихся формируются навыки работы с современными библиотеками интеллектуального анализа и обработки естественного языка Apache OpenNLP и FreeLing языка программирования Java и поисковых систем Elasticsearch, Apache Solr или Apache Lucene.

2. Планируемые результаты обучения по дисциплине (модулю).

Перечень формируемых результатов освоения образовательной программы (компетенций) в результате обучения по дисциплине (модулю):

ПК-3 - Способен спроектировать, разработать, обучить, оценить и развернуть модели искусственного интеллекта в соответствии с методологией MLOps.

Обучение по дисциплине (модулю) предполагает, что по его результатам обучающийся будет:

Знать:

- принципы текстовой классификации;
- отличительные особенности полнотекстового, релевантного и нечетного поиска;
- отличительные особенности задач классификации для информационного поиска и извлечения информации.

Уметь:

- проводить автомазитизрованную обработку естественного языка;
- представлять текстовую информацию в векторной плоскости представления слов и связанных выражений;
- оценивать релевантность результатов поиска на ранжированных и неранжированных данных;

Владеть:

- навыком предварительной обработки и построения семантических связей между словами;

- навыком построения и настройки поисковых систем;

- навыком интегрирования и развертывания среды разработки и обучения модели искусственного интеллекта для информационного поиска.

3. Объем дисциплины (модуля).

3.1. Общая трудоемкость дисциплины (модуля).

Общая трудоемкость дисциплины (модуля) составляет 5 з.е. (180 академических часа(ов).

3.2. Объем дисциплины (модуля) в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении учебных занятий:

Тип учебных занятий	Количество часов	
	Всего	Сем. №2
Контактная работа при проведении учебных занятий (всего):	32	32
В том числе:		
Занятия лекционного типа	16	16
Занятия семинарского типа	16	16

3.3. Объем дисциплины (модуля) в форме самостоятельной работы обучающихся, а также в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении промежуточной аттестации составляет 148 академических часа (ов).

3.4. При обучении по индивидуальному учебному плану, в том числе при ускоренном обучении, объем дисциплины (модуля) может быть реализован полностью в форме самостоятельной работы обучающихся, а также в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении промежуточной аттестации.

4. Содержание дисциплины (модуля).

4.1. Занятия лекционного типа.

№ п/п	Тематика лекционных занятий / краткое содержание
1	Анализ текста. Рассматриваемые вопросы: - представления языка – фонетика, морфология, синтаксис, семантика; - морфологический и синтаксический анализ; - обзор направлений анализа текста: классификация, машинный перевод, , извлечение ключевых слов (NER), информационный поиск, извлечение информации.
2	Задачи информационного поиска. Рассматриваемые вопросы: - классификация и фильтрация документов; - кластеризация полнотекстовых документов; - контекстно-зависимый анализ; - этапы информационного поиска.
3	Информационный поиск. Рассматриваемые вопросы: - полнотекстовый поиск и словари (тезариусы); - булевский поиск, релевантный и нечеткий поиск.
4	Методы нечеткого поиска. Рассматриваемые вопросы: - расширение множества признаков; - фрагментное индексирование (N-граммы); - хэширование по сигнатуре; - метрические деревья: аннотированное суффиксное дерево (АСД).
5	Индексирование и ранжирование текстов. Рассматриваемые вопросы: - инвертированный индекс; - латентно-семантическим индексирование (LSI); - вероятностный латентно-семантический анализ: латентное раамещение Дирихле.
6	Векторизация текста. Рассматриваемые вопросы: - построение векторного набора слов: быстрое кодирование (One-Hot Encoding), вещественные векторы; - методы «мешка слов»: Bag of Words(BoW), Continuous Bag of Words (CBow), Deep CBow; - модель скип-грамм (Skip Gram); - статистические меры: частота слова (TF), TF/IDF; - семантические вложения (эмбеддинги);
7	Оценка эффективности информационного поиска. Рассматриваемые вопросы: - релевантность и оценка релевантности ; - критерии оценки эффективности; - оценка неранжированных и ранжированных результатов поиска.
8	Поисковые системы. Рассматриваемые вопросы: - типы информационных систем; - устройство и принцип работы поисковых систем; - понятие семантического ядра.

4.2. Занятия семинарского типа.

Лабораторные работы

№ п/п	Наименование лабораторных работ / краткое содержание
1	Предварительная обработка текста. В результате выполнения лабораторных работ студент производит настройку среды разработки и осваивает механизмы обнаружения и выделения предложений в наботе текстовых данных с помощью библитек OpenNLP и FreeLing языка программирования Java.
2	Сегментация слов. В результате выполнения лабораторных работ студент осваивает механизмы выделения связанных слов (токенов) на наботе текстовых данных с помощью библитек OpenNLP и FreeLing языка программирования Java.
3	Маркировка частей речи. В результате выполнения лабораторных работ студент осваивает механизмы выделения частей речи на наботе текстовых данных с помощью библитек OpenNLP и FreeLing языка программирования Java.
4	Лемматизация слов. В результате выполнения лабораторных работ студент осваивает механизмы объединения словоформ на наботе текстовых данных с помощью библитек OpenNLP и FreeLing языка программирования Java.
5	Кластеризация частей речи. В результате выполнения лабораторных работ студент осваивает механизмы разделения частей речи на наботе текстовых данных с помощью библитек OpenNLP и FreeLing языка программирования Java.
6	Распознование именнованных сущностей. В результате выполнения лабораторных работ студент осваивает механизмы поиска и выделения именнованных объектов на наботе текстовых данных с помощью библитек OpenNLP и FreeLing языка программирования Java.
7	Полнотекстовый поиск В результате выполнения лабораторных работ студент осваивает механизмы полнотекстового поиска на наботе текстовых данных с помощью библитек Elasticsearch, Apache Solr или Apache Lucene языка программирования Java.
8	Поиск по нескольким полям В результате выполнения лабораторных работ студент осваивает механизмы информационного поиска по нескольким полями (терминам) на наботе текстовых данных с помощью библитек Elasticsearch, Apache Solr или Apache Lucene языка программирования Java.

4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	Изучение рекомендованной литературы.
2	Подготовка к лабораторным работам.
3	Выполнение курсового проекта.
4	Подготовка к промежуточной аттестации.
5	Подготовка к текущему контролю.

4.4. Примерный перечень тем курсовых проектов

1. Разработать систему релевантного информационного поиска для предметной области «Новостная лента»

2. Разработать систему релевантного информационного поиска для

предметной области «Библиотека»

3. Разработать систему релевантного информационного поиска для предметной области «Аренда жилья»

4. Разработать систему релевантного информационного поиска для предметной области «Онлайн магазин»

5. Разработать систему релевантного информационного поиска для предметной области «Электронная почта»

6. Разработать систему релевантного информационного поиска для предметной области «Веб-учебник»

7. Разработать систему релевантного информационного поиска для предметной области «Онлайн энциклопедия»

8. Разработать систему релевантного информационного поиска для предметной области «Сайт отзывов»

9. Разработать систему релевантного информационного поиска для предметной области «Поиск вакансий»

10. Разработать систему релевантного информационного поиска для предметной области «Одноклассники»

5. Перечень изданий, которые рекомендуется использовать при освоении дисциплины (модуля).

№ п/п	Библиографическое описание	Место доступа
1	Бенджамин Бенгфорт. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. – Санкт-Петербург: Питер, 2021. – 368 с. – ISBN 978-5-4461-1153-4.	https://ibooks.ru/bookshelf/365298/reading (дата обращения: 1.11.2022). – Текст : электронный
2	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.	https://e.lanbook.com/book/140584 (дата обращения: 1.11.2022). – Текст : электронный
3	Даг, Т. Релевантный поиск с использованием Elasticsearch и Solr / Т. Даг, Б. Джон ; перевод с английского А. Н. Киселев. — Москва : ДМК Пресс, 2018. — 408 с. — ISBN 978-5-97060-592-9.	https://e.lanbook.com/book/111439 (дата обращения: 1.11.2022). – Текст : электронный
4	Бринк Х. Машинное обучение / Х. Бринк, Д. Ричардс, М. Феверолф. – Санкт-Петербург :	https://ibooks.ru/bookshelf/355472/reading (дата обращения: 1.11.2022). – Текст :

	Питер, 2017. – 336 с. – ISBN 978-5-496-02989-6	электронный
5	Ингерсолл, Г. С. Обработка неструктурированных текстов. Поиск, организация и манипулирование / Г. С. Ингерсолл, Т. С. Мортон, Э. Л. Фэррис. — Москва : ДМК Пресс, 2015. — 414 с. — ISBN 978-5-97060-144-0.	https://e.lanbook.com/book/73069 (дата обращения: 1.11.2022). – Текст : электронный

6. Перечень современных профессиональных баз данных и информационных справочных систем, которые могут использоваться при освоении дисциплины (модуля).

Электронно-библиотечная система Научно-технической библиотеки РУТ(МИИТ) (http://library.miit.ru/)

Электронно-библиотечная система издательства «Лань» (http://e.lanbook.com/)

Электронно-библиотечная система ibooks.ru (http://ibooks.ru/)

7. Перечень лицензионного и свободно распространяемого программного обеспечения, в том числе отечественного производства, необходимого для освоения дисциплины (модуля).

Прикладное программное обеспечение
Браузер Microsoft Internet Explorer или его аналоги
Пакет офисных программ Microsoft Office или его аналоги
Среда разработки PyCharm Community Edition

8. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю).

Учебные аудитории для проведения учебных занятий, оснащенные компьютерной техникой и наборами демонстрационного оборудования.

Для практических занятий – наличие персональных компьютеров вычислительного класса.

9. Форма промежуточной аттестации:

Курсовой проект во 2 семестре.
Экзамен во 2 семестре.

10. Оценочные материалы.

Оценочные материалы, применяемые при проведении промежуточной аттестации, разрабатываются в соответствии с локальным нормативным актом РУТ (МИИТ).

Авторы:

заведующий кафедрой, доцент, к.н. кафедры «Цифровые технологии управления транспортными процессами» В.Е. Нутович

старший преподаватель кафедры «Цифровые технологии управления транспортными процессами» Е.А. Заманов

Согласовано:

Заведующий кафедрой ЦТУТП В.Е. Нутович

Председатель учебно-методической комиссии Н.А. Андриянова