17.02.2025

ПРАКТИЧЕСКИЕ ЗАДАНИЯ

1. МОДУЛЬ 1 (17.02.2025-17.03.2025)

Разработка парсер-менеджера

A. Проектно-практическое задание модуля

i. Разработка и тестирование синтаксического анализатора для обработки html страниц.

ii. Разработка и тестирование синтаксического анализатора для автоматизированной обработки документов форматов .pdf, .doc, .docx, djvu.

B. Темы семинаров модуля:

i. Современные методы ранжирования результатов информационного поиска в Вебе.

ii. Mining query logs to improve web search engines' operations.

iii. Модели языка (Bag of words, Word2Vec, Word embedding и др.) и методы обработки текстовой информации (WordNet, лемматизация, стемминг).

iv. Кодирование текстовых данных на основе модели GPT.

v. Выявление именованных сущностей (NER).

C. Deadline модуля – 17.03.2023

2. МОДУЛЬ 2 (18.03.2025-21.04.2025)

Разработка поискового робота для сбора и обработки данных с ресурсов Web 1.0/Web 2.0

A. Проектно-практическое задание модуля

i. Разработка простейшей модели поискового робота с классическим алгоритмом сбора и обработки данных в сети Веб 1.0/Веб 2.0.

ii. Автоматизированный сбор данных с помощью простейшей модели поискового робота на основе специализированного алгоритма обхода на примере сайтов СПбГУ и МГУ – для Веб 1.0 / Автоматизированный сбор данных на основе API социальной сети VKontakte или мессенджера Телеграм об упоминаниях в пользовательских публикациях СПбГУ и МГУ – для Веб 2.0.

iii. Сбор статистики обработанных страниц для Веб 1.0: общее количество страниц и всех ссылок, количество внутренних страниц, количество неработающих страниц, количество внутренних поддоменов, общее

количество ссылок на внешние ресурсы, количество уникальных внешних ресурсов, количество уникальных ссылок на файлы doc/docx/pdf. Статистика для Веб 2.0: количество публикаций об упоминании университета, количество публикующих контент пользователей, количество лайков/просмотров/комментариев/репостов, график количество публикаций в день за собираемый период.

B. Темы семинаров модуля:

i. Методы обновления данных в индексе с помощью поисковых роботов. Стратегии равномерного и пропорционального обновления.
ii. Алгоритм PageRank и его модификации для вычисления весов Веб-страниц сайтов.
iii. Меры центральности графов, используемые при анализе данных социальных сетей.
iv. Обучение с подкреплением (Reinforcement Learning).
v. Методы сокращения размерности данных.

C. Deadline модуля – 21.04.2025

3. МОДУЛЬ 3 (22.04.2025-22.05.2025)

Разработка простейшей модели инвертированного индекса

A. Проектно-практическое задание модуля

i. Реализация и тестирование индексной структуры на основе инвертированного индекса.
ii. Применение метода сжатия инвертированного индекса с использованием дельта и гамма-кодирования Элиаса.
iii. Тестирование процесса индексирования собранных на 2м этапе веб-страниц сайта (упоминаний в социальной сети VK) СПбГУ или МГУ (скорость процесса индексирования по количеству текстовых документов около 40 тыс., проверить на сколько эффективно индексирование с использованием алгоритма сжатия уменьшает объемы хранимой информации по сравнению с классической ситуацией, не предусматривающей использование алгоритма сжатия). Проверить скорость поиска по запросу «Ректор СПбГУ/МГУ».

B. Темы семинаров модуля:

i. Распределенное (MapReduce) и динамическое индексирование.
ii. Алгоритмы сжатия: непараметрические алгоритмы дельта- и гамма-кодирования Элиаса. Алгоритмы сжатия: параметрический алгоритм кодирования Голомба.
iii. Индексная структура суффиксные деревья, их принципы структуризации информации и архитектурные особенности.

iv. Задача дедупликации данных

C. Deadline модуля – 26.05.2025