И.М. Никольский
практикум по ВПиРОД-2025

Задание 1
Загрузка и секционирование данных

Крайний срок — 7 октября (первый вторник октября).

Необходимо создать распределённую систему, состоящую из менеджера, клиента (он же — управляющая консоль) и процессов-хранителей.

Система работает с корпусом текстов. Необходимо выполнить предобработку (в каждом файле заменить первые три столбца на знак „-“) и разложить файлы по узлам. Реализовать консоль, которая может выполнять команды:

- start – запускает менеджера и хранителей
- load folder mode - загрузить корпус текстов (mode - вариант загрузки, даны ниже) из папки folder,
- purge - очистить узлы,
- find word - найти в загруженных текстах предложения со словом word; эти предложения должны быть распечатаны в окне консоли с указанием (для каждого предложения) названия соответствующего файла и номера узла.

Корпус лежит здесь:
https://github.com/thuhcsi/english-conversation-corpus/tree/master/conversations

Варианты загрузки файлов корпуса:
1. равномерная загрузка (mode = ‘e’) — корпус делится на равные части, каждая часть загружается на свой хранитель
2. неравномерная загрузка (mode = ‘u’) — на все хранители, кроме одного, загружаются по одному тексту, остальные — на оставшийся хранитель

Архитектура системы:

- рабочие процессы-хранители данных;
- менеджер — основной процесс системы
- клиент — через него происходит взаимодействие пользователя с системой, запрос данных
- связь между процессами - **RabbitMQ**

Каждый **рабочий процесс** отвечает за свою часть корпуса. На каждом из них хранится определённое количество текста. Каждый текст должен иметь метаинформацию — название соответствующего файла.

Менеджер:

- хранит информацию о том, где какие данные хранятся;
- загружает данные в систему по команде пользователя;
- выдаёт необходимые данные по команде с консоли;

Количество хранителей k — параметры ком. строки при запуске менеджера

Клиент-консоль:

- запускает менеджера, процессы-хранители и очереди RabbitMQ (после запуска всех процессов выдаёт сообщение «System ready»);
- имеет связь только с менеджером
- умеет обрабатывать команды, описанные в начале задания
- при некорректном вводе (неизвестная команда, find или load без аргумента и т. д.) выдаёт соответствующую диагностику

для простоты считаем, что клиент-консоль всегда запускается первым, остальные процессы системы можно запустить только консольной командой start.