Министерство образования Республики Беларусь

Учреждение образования

«Брестский государственный технический университет»

Кафедра ИИТ

Лабораторная работа № 4

По дисциплине ЕЯИИС

Тема: «Текстовые корпусы и методы идентификации языка текста»

Выполнил:

Студент 4 курса

Группы ИИ-16

Журавлёв В.А.

Проверил:

Слинко Е.В.

Брест, 2021

**Цель работы:** закрепить знания в решении задач построения и работы с корпусами текстов.

**Код программы:**

package org.example.service;

import org.springframework.stereotype.Component;

import javax.annotation.PostConstruct;

import java.io.BufferedReader;

import java.io.File;

import java.io.FileReader;

import java.util.Arrays;

import java.util.HashMap;

import java.util.List;

import java.util.Map;

import java.util.stream.Collectors;

@Component

public class ResourceStore {

private Map<String, Integer> rusWordsWithPosition;

private Map<String, Integer> engWordsWithPosition;

private Map<String, Integer> readGivenFiles(final String fileName) throws Exception{

Map<String, Integer> wordsWithNumberOfFilesContainingWord = new HashMap<>();

BufferedReader br = new BufferedReader(new FileReader(new File(fileName)));

StringBuilder resultStringBuilder = new StringBuilder();

String line;

while ((line = br.readLine()) != null) {

resultStringBuilder.append(line).append(" ");

}

List<String> words = Arrays.stream(resultStringBuilder.toString().split("[ \n,.;:\"'»«()—!0-9\\[\\]]"))

.filter(word -> !word.isEmpty())

.map(String::toLowerCase)

.collect(Collectors.toList());

for (String word : words) {

wordsWithNumberOfFilesContainingWord.computeIfPresent(word, (key, value) -> value + 1);

wordsWithNumberOfFilesContainingWord.putIfAbsent(word, 1);

}

return wordsWithNumberOfFilesContainingWord;

}

private List<String> getSortedWords(Map<String, Integer> map){

return map.entrySet()

.stream()

.sorted((entry1, entry2) -> entry2.getValue().compareTo(entry1.getValue()))

.map(Map.Entry::getKey)

.collect(Collectors.toList());

}

@PostConstruct

private void fillResources() throws Exception {

Map<String, Integer> rusMap = readGivenFiles("src/main/resources/rus/text.txt");

Map<String, Integer> engMap = readGivenFiles("src/main/resources/eng/text.txt");

List<String> rusSortedWords = getSortedWords(rusMap);

List<String> engSortedWords = getSortedWords(engMap);

int minSize = Math.min(rusSortedWords.size(), engSortedWords.size());

rusWordsWithPosition = new HashMap<>();

for (int i = 0; i < minSize; i++) {

rusWordsWithPosition.put(rusSortedWords.get(i), i);

}

engWordsWithPosition = new HashMap<>();

for (int i = 0; i < minSize; i++) {

engWordsWithPosition.put(engSortedWords.get(i), i);

}

}

public String getLanguage(final String fileName) throws Exception {

Map<String, Integer> requestMap = readGivenFiles("src/main/resources/data/" + fileName + ".txt");

List<String> requestSortedWords = getSortedWords(requestMap);

Integer rusCount = 0;

Integer engCount = 0;

for (int i = 0; i < requestSortedWords.size(); i++) {

String requestWord = requestSortedWords.get(i);

if(!rusWordsWithPosition.containsKey(requestWord)){

rusCount+=rusWordsWithPosition.size();

} else {

rusCount+=Math.abs(rusWordsWithPosition.get(requestWord) - i);

}

if(!engWordsWithPosition.containsKey(requestWord)){

engCount+=rusWordsWithPosition.size();

} else {

engCount+=Math.abs(engWordsWithPosition.get(requestWord) - i);

}

}

return rusCount > engCount

? "English"

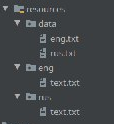
: "Русский";

}

}

**Тестирование:**

Файловая система содержит файлы:



В файлах eng/text.txt и rus.text.txt содержатся исходные тексты для составления N-грамм языков.

В файлах eng.txt и rus.txt – тексты меньшего размера для тестирования.

При запросе “eng” из текста eng.txt строятся N-граммы для сравнения с исходными N-граммами языков. В ответ приходит язык текста.  


При запросе “rus” из текста rus.txt строятся N-граммы для сравнения с исходными N-граммами языков. В ответ приходит язык текста.



**Вывод:** закрепил знания в решении задач построения и работы с корпусами текстов.