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 (2)

Пешко А.С.

Проверил:

Скарубо А.О.

Брест, 2021

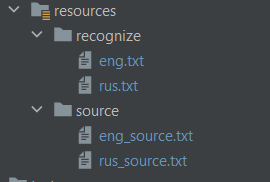
**Цель работы:** закрепить знания в решении задач построения и работы с корпусами текстов.

**Ход работы:**

**Код программы:**  
  
@Component  
public class LanguageService {  
  
 private Map<String, Integer> rusTrainingSet;  
 private Map<String, Integer> engTrainingSet;  
  
 @PostConstruct  
 private void fillResources() throws Exception {  
 Map<String, Integer> rusMap = countWordRepetitions("src/main/resources/source/rus\_source.txt");  
 Map<String, Integer> engMap = countWordRepetitions("src/main/resources/source/eng\_source.txt");  
  
 List<String> rusSortedWords = sortWords(rusMap);  
 List<String> engSortedWords = sortWords(engMap);  
 int minSize = Math.min(rusSortedWords.size(), engSortedWords.size());  
  
 rusTrainingSet = new HashMap<>();  
 for (int i = 0; i < minSize; i++) {  
 rusTrainingSet.put(rusSortedWords.get(i), i);  
 }  
  
 engTrainingSet = new HashMap<>();  
 for (int i = 0; i < minSize; i++) {  
 engTrainingSet.put(engSortedWords.get(i), i);  
 }  
 }  
  
 public String recognizeLanguage(final String fileName) throws Exception {  
 Map<String, Integer> requestMap = countWordRepetitions("src/main/resources/recognize/" + fileName + ".txt");  
 List<String> requestSortedWords = sortWords(requestMap);  
  
 int rusCount = 0;  
 int engCount = 0;  
 for (int i = 0; i < requestSortedWords.size(); i++) {  
 String requestWord = requestSortedWords.get(i);  
  
 if (!rusTrainingSet.containsKey(requestWord)) {  
 rusCount += rusTrainingSet.size();  
 } else {  
 rusCount += Math.abs(rusTrainingSet.get(requestWord) - i);  
 }  
  
 if (!engTrainingSet.containsKey(requestWord)) {  
 engCount += engTrainingSet.size();  
 } else {  
 engCount += Math.abs(engTrainingSet.get(requestWord) - i);  
 }  
 }  
 return rusCount > engCount ? "English" : "Русский";  
 }  
  
 private Map<String, Integer> countWordRepetitions(final String fileName) throws Exception {  
  
 String fileString = readFromFile(fileName);  
  
 Map<String, Integer> wordsWithNumberOfRepetitions = new HashMap<>();  
 List<String> words = Arrays.stream(fileString.split("[ ,.;:0-9\"'»«()—!\\[\\]\n]+"))  
 .filter(word -> !word.isEmpty())  
 .map(String::toLowerCase)  
 .collect(Collectors.toList());  
 for (String word : words) {  
 wordsWithNumberOfRepetitions.computeIfPresent(word, (key, value) -> value + 1);  
 wordsWithNumberOfRepetitions.putIfAbsent(word, 1);  
 }  
 return wordsWithNumberOfRepetitions;  
 }  
  
 private List<String> sortWords(Map<String, Integer> map) {  
 return map.entrySet()  
 .stream()  
 .sorted((entry1, entry2) -> entry2.getValue().compareTo(entry1.getValue()))  
 .map(Map.Entry::getKey)  
 .collect(Collectors.toList());  
 }  
  
 private String readFromFile(final String fileName) throws Exception {  
 BufferedReader br = new BufferedReader(new FileReader(fileName));  
 StringBuilder resultStringBuilder = new StringBuilder();  
 String line;  
 while ((line = br.readLine()) != null) {  
 resultStringBuilder.append(line).append(" ");  
 }  
 return resultStringBuilder.toString();  
 }  
  
}

**Тестирование:**

Файловая система содержит :

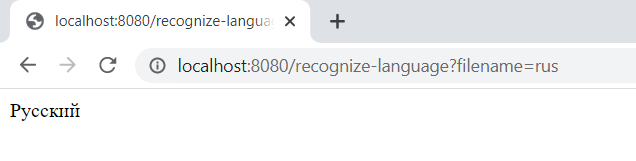


В файлах recorgnize/… содержаться тексты для тестирования (распознования языка).

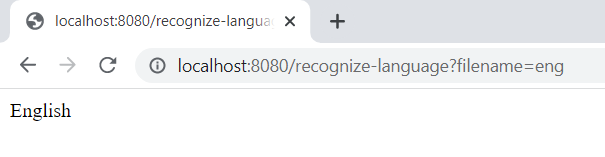
В файлах source/… содержатся исходные тексты для составления N-грамм языков.

Для тестирования передаем имя файла и как результат возвращается язык текста в переданном файле:

Rus:



Eng:



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