Name : Parth Kalbhor

BE-AIML

Practical 2 Single-pass Algorithm

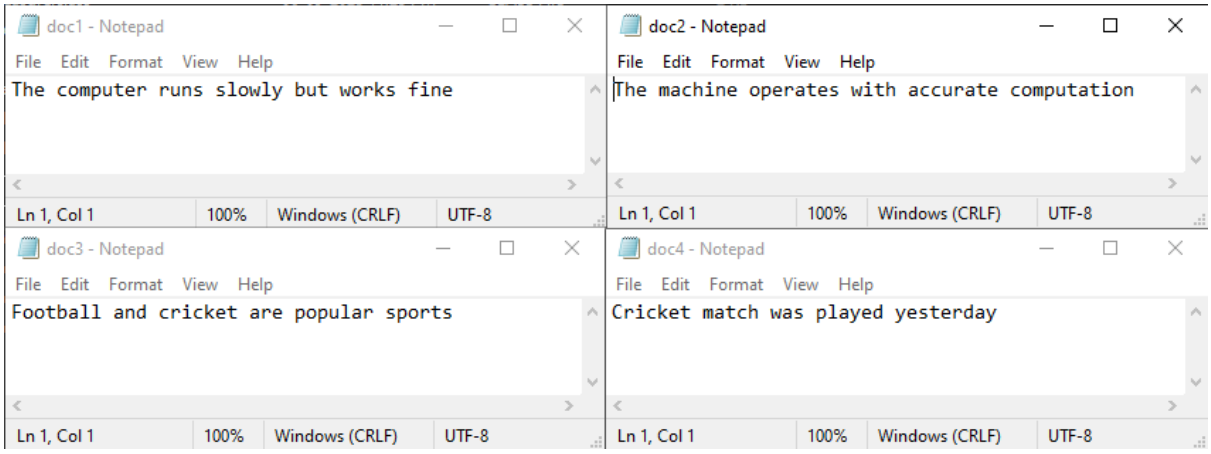
CODE:

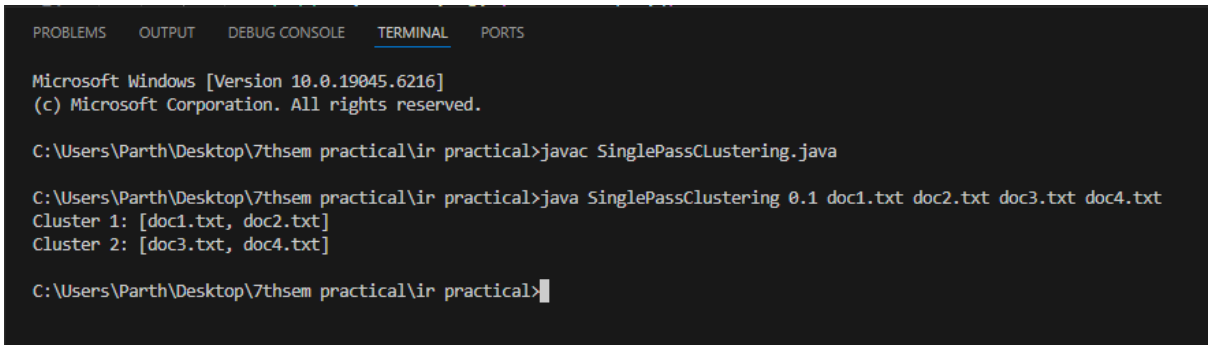
```
import java.io.*;
import java.util.*;
public class SinglePassClustering {
   static class Cluster {
      List<String> documents = new ArrayList<>();
      Map<String, Integer> centroid = new HashMap<>();
      Cluster(String doc, Map<String, Integer> vec) {
         documents.add(doc);
         centroid.putAll(vec);
      }
      // Update centroid by averaging term frequencies
      void addDocument(String doc, Map<String, Integer> vec) {
         documents.add(doc);
         for (String term : vec.keySet()) {
            centroid.put(term, centroid.getOrDefault(term, 0) + vec.get(term));
         }
      }
   }
   // Convert text into a term-frequency vector
   static Map<String, Integer> getVector(String text) {
      Map<String, Integer> vector = new HashMap<>();
      String[] words = text.toLowerCase().split("\\W+");
      for (String w : words) {
         if (w.isEmpty()) continue;
         vector.put(w, vector.getOrDefault(w, 0) + 1);
      }
      return vector;
```

```
    }
    // Cosine similarity between two vectors
    static double cosineSim(Map<String, Integer> v1, Map<String, Integer> v2) {
        double dot = 0.0, norm1 = 0.0, norm2 = 0.0;
        for (String key : v1.keySet()) {
            if (v2.containsKey(key)) dot += v1.get(key) * v2.get(key);
            norm1 += Math.pow(v1.get(key), 2);
        }
        for (int val : v2.values()) norm2 += Math.pow(val, 2);
        return (norm1 == 0 || norm2 == 0) ? 0 : dot / (Math.sqrt(norm1) * Math.sqrt(norm2));
    }
    public static void main(String[] args) throws Exception {
        if (args.length < 2) {
            System.out.println("Usage: java SinglePassClustering <similarity_threshold> <file1> <file2>
...");
            return;
        }
        double threshold = Double.parseDouble(args[0]); // e.g., 0.3
        List<String> files = Arrays.asList(Arrays.copyOfRange(args, 1, args.length));
        List<Cluster> clusters = new ArrayList<>();
        for (String file : files) {
            // Read file content
            BufferedReader br = new BufferedReader(new FileReader(file));
            StringBuilder sb = new StringBuilder();
            String line;
            while ((line = br.readLine()) != null) sb.append(line).append(" ");
            br.close();
            Map<String, Integer> vec = getVector(sb.toString());
            // Convert term frequencies to presence vector (1 if word exists)
            for (String key : vec.keySet()) vec.put(key, 1);
            // Find best matching cluster
            double bestSim = 0;
            Cluster bestCluster = null;
            for (Cluster c : clusters) {
```

```
            double sim = cosineSim(vec, c.centroid);
            if (sim > bestSim) {
                bestSim = sim;
                bestCluster = c;
            }
        }
        if (bestSim >= threshold && bestCluster != null) {
            bestCluster.addDocument(file, vec);
        } else {
            clusters.add(new Cluster(file, vec));
        }
    }
    // Print results
    int clusterId = 1;
    for (Cluster c : clusters) {
        System.out.println("Cluster " + clusterId++ + ": " + c.documents);
    }
  }
}
```

OUTPUT :


doc1 - Notepad
File Edit Format View Help
The computer runs slowly but works fine
Ln 1, Col 1 100% Windows (CRLF) UTF-8
doc2 - Notepad
File Edit Format View Help
The machine operates with accurate computation
Ln 1, Col 1 100% Windows (CRLF) UTF-8
doc3 - Notepad
File Edit Format View Help
Football and cricket are popular sports
Ln 1, Col 1 100% Windows (CRLF) UTF-8
doc4 - Notepad
File Edit Format View Help
Cricket match was played yesterday
Ln 1, Col 1 100% Windows (CRLF) UTF-8


PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
Microsoft Windows [Version 10.0.19045.6216]
(c) Microsoft Corporation. All rights reserved.
C:\Users\Parth\Desktop\7thsem practical\ir practical>javac SinglePassClustering.java
C:\Users\Parth\Desktop\7thsem practical\ir practical>java SinglePassClustering 0.1 doc1.txt doc2.txt doc3.txt doc4.txt
Cluster 1: [doc1.txt, doc2.txt]
Cluster 2: [doc3.txt, doc4.txt]
C:\Users\Parth\Desktop\7thsem practical\ir practical>