Development task

Purpose

Categorize a list of files based on their name similarity

Details

Provided a list of file names, we want to group the files based on how similar their names are.

In order to do that we want to use 2 algorithms:

1. Jaro-Winkler based algorithm for ranking “similarity” between 2 strings

2. k-means clustering, for grouping the similar strings (i.e. file names)

Reference resources:

1. Jaro-Winkler: https://github.com/TriviaMarketing/Jaro-Winkler (note this is a C++

implementation)

2. k-means clustering: https://en.wikipedia.org/wiki/K-means\_clustering (Note: this is a

very theoretical explanation, our case is bi-dimensional, (i.e. a relatively simple case)

3. Files list:

https://drive.google.com/open?id=0ByXgLChIWVLgZlNGb0s5NHg1WWVhWFBPU1

p4QlJsa2F0SjJz

The task

Write a program that accepts the “K” parameter, and the file “filelist.txt”

The program should implement the k-means clustering, and use the Jaro-Winkler algorithm to

group the file names from the list into K groups based on the “similarity-distance” from each

other.

The program should output the information into simple text or csv file(s).