Text File Similarity Measurement Report

School of Computer Science and Information Technology

COSC2468 Scripting Language Programming

Semester 1 2018 – Assignment 1 – Perl

Jenna Afarian

S3511312

The intent of this report is to explain the choices made in Part D to measure the similarity of two text files. The following interpretation of the specification was made:

1. Similarity measurement should be based on a word frequency count (or how many times words overlap with each other between files).
2. If files contain exactly the same words, they match 100%.
3. If files contain no duplicate words, they match 0%.

Using Text::Similarity::Overlaps I was able to count the number of times a word overlapped between files. I then chose to get a percentage of overlaps based on the number of words in the largest file by counting the words in both files.

To determine the percentage I used the following calculation:

```
<number of overlaps> / <number of words in largest file> * 100
```

Example of a 100% match

File 1 contents	I am a test file.
File 2 contents	I am a test file.
Number of overlaps	5
Word count of largest file	5
Calculation	5 / 5 * 100
Result	100%

Example of a 0% match

File 1 contents	I am a test file.
File 2 contents	This is not your lunch, Mr. Potter.
Number of overlaps	0
Word count of largest file	7
Calculation	0/ 7 * 100
Result	0%