CSE17040 - Document similarity

July 28, 2020

[1]:
```
import math
import string
import sys
```

[2]:
```
def read_file(filename):
    try:
        with open(filename, 'r') as f:
            data = f.read()
        return data

    except IOError:
        print("Error opening or reading input file: ", filename)
        sys.exit()
```

[3]:
```
translation_table = str.maketrans(string.punctuation+string.ascii_uppercase, "␣
↪"*len(string.punctuation)+string.ascii_lowercase)
```

[4]:
```
def get_words_from_line_list(text):
    text = text.translate(translation_table)
    word_list = text.split()
    return word_list
```

[5]:
```
def count_frequency(word_list):
    D = {}
    for new_word in word_list:
        if new_word in D:
            D[new_word] = D[new_word] + 1
        else:
            D[new_word] = 1
    return D
```

[6]:
```
def word_frequencies_for_file(filename):
    line_list = read_file(filename)
    word_list = get_words_from_line_list(line_list)
    freq_mapping = count_frequency(word_list)
    print("File", filename, ":", )
    print(len(line_list), "lines, ", )
```

```
    print(len(word_list), "words, ", )
    print(len(freq_mapping), "distinct words")
    return freq_mapping
```

[7]:
```
def dotProduct(D1, D2):
    Sum = 0.0
    for key in D1:
        if key in D2:
            Sum += (D1[key] * D2[key])
    return Sum
```

[8]:
```
def vector_angle(D1, D2):
    numerator = dotProduct(D1, D2)
    denominator = math.sqrt(dotProduct(D1, D1)*dotProduct(D2, D2))
    return math.acos(numerator / denominator)
```

[9]:
```
def documentSimilarity(filename_1, filename_2):

    sorted_word_list_1 = word_frequencies_for_file(filename_1)
    sorted_word_list_2 = word_frequencies_for_file(filename_2)
    distance = vector_angle(sorted_word_list_1, sorted_word_list_2)
    print("The distance between the documents is: % 0.4f (radians)"% distance)
```

[10]:
```
documentSimilarity('alice.txt', 'alicemodified.txt')
```

```
File alice.txt :
1326 lines,
257 words,
138 distinct words
File alicemodified.txt :
1335 lines,
259 words,
139 distinct words
The distance between the documents is:  0.0429 (radians)
```

[ ]: