Information Retrieval and Web Analytics

Semester 2 - 2022

Practical Sheet 03

Write python code to do the followings.

1). Let assume the following corpus

D1 : I am Sam.
D2 : Sam I am.
D3 : I do not like green eggs and ham.
D4 : I do not like them, Sam I am.

a).Write the code to create k-grams for the above corpus when k=1,2,3

b).Write code to find out the Jaccard coefficient between the given documents based on different k.

2) Apply levenshtein algorithm to “python” and “pythonly” and get the minimum edit distance

3) Implement the SOUNDEX algorithm.