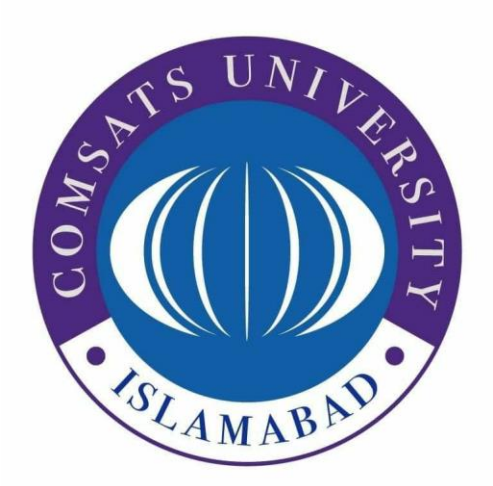
Assignment No. 4



Name: M. Sarmad Aslam

Roll No: FA21-BSE-093 - C

Subject: Data Science

COMSATS University Islamabad, Lahore Campus

11-December-2023

11-12-23

CSC461 – Assignment 4 – NLP

Muhammad Sarmad Aslam

Fa21-BSE-093

Question No. 1

i) Bag of Words (BoW):

Vocabulary: {data, science, is, one, of, the, most, important, courses, in, computer, this, best, scientists, perform, analysis}

S1: [1, 2, 1, 1, 2, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0]

S2: [1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0]

S3: [2, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1]

ii) Term Frequency (TF):

TF = (Number of times a word appears in a document) / (Total number of words in the document)

TF(S1): [1/16, 2/16, 1/16, 1/16, 2/16, 1/16, 1/16, 1/16, 1/16, 1/16, 1/16, 0, 0, 0, 0, 0]

TF(S2): [1/16, 1/16, 1/16, 1/16, 1/16, 0, 1/16, 0, 1/16, 1/16, 0, 1/16, 1/16, 0, 0, 0]

TF(S3): [2/11, 1/11, 0, 0, 1/11, 1/11, 0, 0, 0, 0, 0, 0, 0, 1/11, 1/11, 1/11]

iii) Inverse Document Frequency (IDF):

IDF = log (total number of documents / number of documents with word (term)

IDF(science) = log(3 / 3) = 0

IDF(is) = log(3 / 3) = 0

IDF(one) = log(3 / 2) = 0.1761

IDF(of) = log(3 / 3) = 0

IDF(the) = log(3 / 3) = 0

IDF(most) = log(3 / 1) = 1.0986

IDF(important) = log(3 / 1) = 1.0986

IDF(courses) = log(3 / 2) = 0.1761

IDF(in) = log(3 / 2) = 0.1761

IDF(computer) = log(3 / 1) = 1.0986

IDF(this) = log(3 / 1) = 1.0986

IDF(best) = log(3 / 1) = 1.0986

IDF(scientists) = log(3 / 1) = 1.0986

IDF(perform) = log(3 / 1) = 1.0986

IDF(analysis) = log(3 / 1) = 1.0986

TF.IDF:

TF.IDF = TF * IDF

TF.IDF(S1): [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

TF.IDF(S2): [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

TF.IDF(S3): [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1.0986/11, 1.0986/11, 1.0986/11]

Question No. 2

Cosine(S1, S2) = dot product(TF-IDF(S1), TF-IDF(S2)) / (magnitude(TF-IDF(S1)) * magnitude(TF-IDF(S2)))

Manhattan(S1, S2) = sum(abs(TF-IDF(S1) - TF-IDF(S2)))

Euclidean(S1, S2) = sqrt(sum((TF-IDF(S1) - TF-IDF(S2))^2))