11-12-23

Assignment #4

Ayesha Zahid

FA21-BSE-003

Question 1

Compute BOW, TF, IDF and TF.IDF values for each term in the given sentences.

S1: "data science is one of the most important courses in computer science."

S2: "this is one of the best data science courses."

S3: "the data scientist perform data analysis."

1- 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]

2- Term Frequency (TF):

TF = (No. of times a term appears in a sentence)/(Total no. of terms in the sentence)

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, 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

3- Inverse Document Frequency (IDF):

IDF = log (N/n), where N is the total no. of documents and n is the number of documents containing the term.

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

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

4- TF.IDF:

TF.IDF = TF x 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(S2): [0,0,0,0,0,0,0,0,0,0,0,0,0, 1.0986/11, 1.0986/11, 1.0986/11, 1.0986/11]

Question 2

Compute the similarities between S1, S2 and S3 using cosine, manhattan and euclidean distances.

Answer

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

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

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