Introduction to Data Science

Assignment 4

Name: Numan Latif

Roll No: FA21-BSE-039

Submitted to: Sir Muhammad Sharjeel

Date: December 11, 2023.

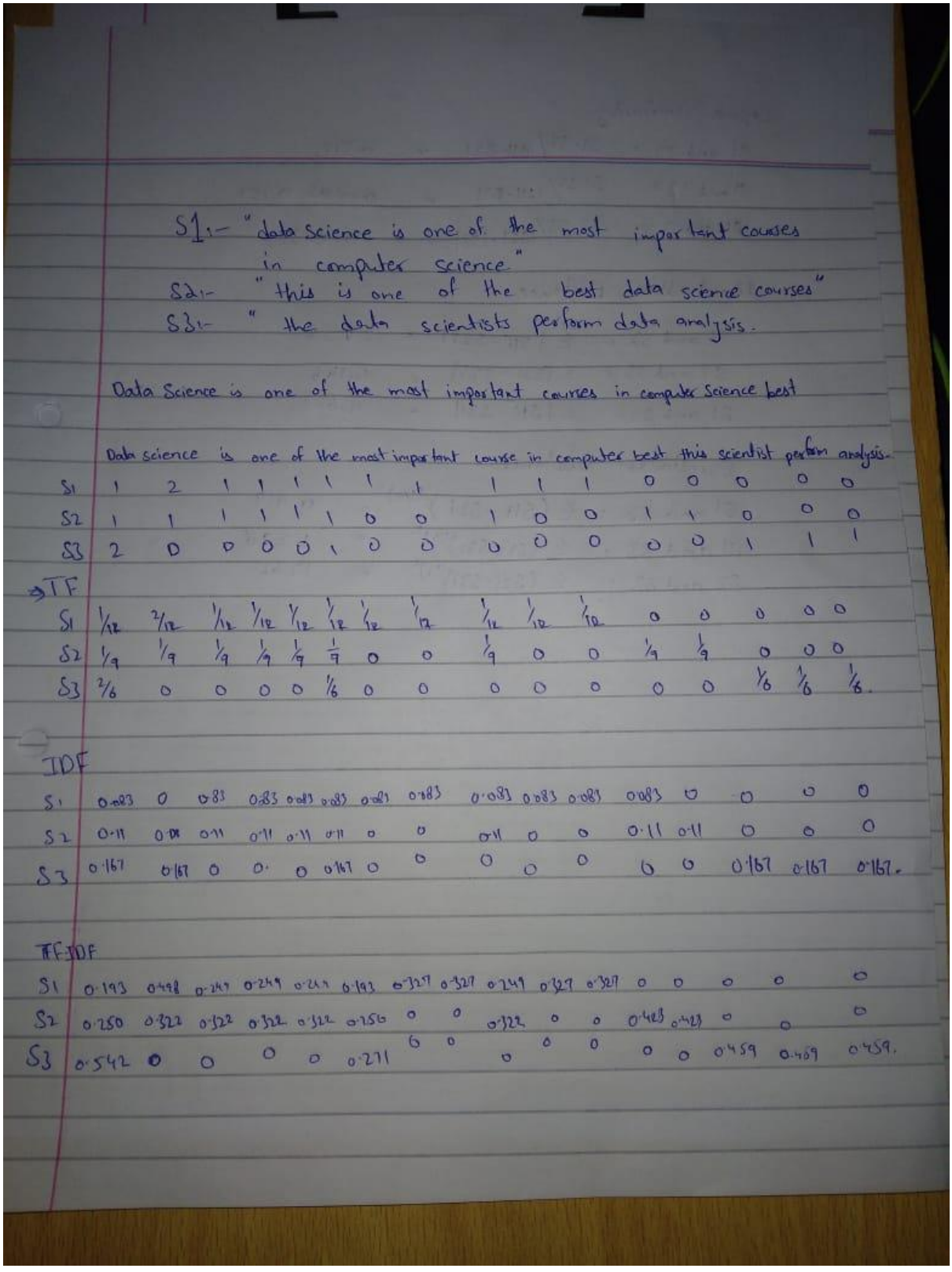
COMSATS University Islamabad, Lahore Campus

Question 1:

```
#December 11, 2023
#CSC461 – Assignment4 – NLP
#Numan Latif
#FA21-BSE-039
#In this task we calculate the BOW,TF,IDF and TF.IDF
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
import pandas as pd

sentence = [
    "data science is one of the most important courses in computer science",
    "this is one of the best data science courses",
    "the data scientists perform data analysis"
]

vectorizer_bow = CountVectorizer()
X_bow = vectorizer_bow.fit_transform(sentence)
bow_df = pd.DataFrame(X_bow.toarray(),
columns=vectorizer_bow.get_feature_names_out())
bow_df.insert(0, 'Sentence', ['S1', 'S2', 'S3'])
print("\nBoW:\n", bow_df.round(3).to_string(index=False))

vectorizer_tf = CountVectorizer()
X_tf = vectorizer_tf.fit_transform(sentence)
tf_df = pd.DataFrame(X_tf.toarray(), columns=vectorizer_tf.get_feature_names_out())
tf_df = tf_df.div(tf_df.sum(axis=1), axis=0)
tf_df.insert(0, 'Sentence', ['S1', 'S2', 'S3'])
```

```
print("\nTF:\n", tf_df.round(3).to_string(index=False))


vectorizer_idf = TfidfVectorizer(use_idf=False, norm='l1')
X_idf = vectorizer_idf.fit_transform(sentence)
idf_df = pd.DataFrame(X_idf.toarray(), columns=vectorizer_idf.get_feature_names_out())
idf_df.insert(0, 'Sentence', ['S1', 'S2', 'S3'])
print("\nIDF:\n", idf_df.round(3).to_string(index=False))


vectorizer_tfidf = TfidfVectorizer()
X_tfidf = vectorizer_tfidf.fit_transform(sentence)
tfidf_df = pd.DataFrame(X_tfidf.toarray(),
columns=vectorizer_tfidf.get_feature_names_out())
tfidf_df.insert(0, 'Sentence', ['S1', 'S2', 'S3'])
print("\nTF.IDF:\n", tfidf_df.round(3).to_string(index=False))
```

```
BoW:
 Sentence  analysis  best  computer  courses  data  important  in  is  most  of  one  perform  science  scientists  the  this
       S1         0     0         1        1     1          1   1   1     1   1    1        0        2           0    1     0
       S2         0     1         0        1     1          0   0   1     0   1    1        0        1           0    1     1
       S3         1     0         0        0     2          0   0   0     0   0    0        1        0           1    1     0

TF:
 Sentence  analysis  best  computer  courses  data  important    in    is  most    of   one  perform  science  scientists   the  this
       S1     0.000 0.000     0.083    0.083 0.083      0.083 0.083 0.083 0.083 0.083 0.083    0.000    0.167       0.000 0.083 0.000
       S2     0.000 0.111     0.000    0.111 0.111      0.000 0.000 0.111 0.000 0.111 0.111    0.000    0.111       0.000 0.111 0.111
       S3     0.167 0.000     0.000    0.000 0.333      0.000 0.000 0.000 0.000 0.000 0.000    0.167    0.000       0.167 0.167 0.000

IDF:
 Sentence  analysis  best  computer  courses  data  important    in    is  most    of   one  perform  science  scientists   the  this
       S1     0.000 0.000     0.083    0.083 0.083      0.083 0.083 0.083 0.083 0.083 0.083    0.000    0.167       0.000 0.083 0.000
       S2     0.000 0.111     0.000    0.111 0.111      0.000 0.000 0.111 0.000 0.111 0.111    0.000    0.111       0.000 0.111 0.111
       S3     0.167 0.000     0.000    0.000 0.333      0.000 0.000 0.000 0.000 0.000 0.000    0.167    0.000       0.167 0.167 0.000

TF.IDF:
 Sentence  analysis  best  computer  courses  data  important    in    is  most    of   one  perform  science  scientists   the  this
       S1     0.000 0.000     0.327    0.249 0.193      0.327 0.327 0.249 0.327 0.249 0.249    0.000    0.498       0.000 0.193 0.000
       S2     0.000 0.423     0.000    0.322 0.250      0.000 0.000 0.322 0.000 0.322 0.322    0.000    0.322       0.000 0.250 0.423
       S3     0.459 0.000     0.000    0.000 0.542      0.000 0.000 0.000 0.000 0.000 0.000    0.459    0.000       0.459 0.271 0.000
```


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 scientists perform data analysis.
Data Science is one of the most important courses in computer Science best
Data science is one of the most important course in computer best this scientist perform analysis.
S1 1 2 1 1 1 1 1 1 1 1 1 0 0 0 0 0
S2 1 1 1 1 1 1 0 0 1 0 0 1 1 0 0 0
S3 2 0 0 0 0 1 0 0 0 0 0 0 0 1 1 1
TF
S1 1/12 2/12 1/12 1/12 1/12 1/12 1/12 1/12 1/12 1/12 1/12 0 0 0 0 0
S2 1/9 1/9 1/9 1/9 1/9 1/9 0 0 1/9 0 0 1/9 1/9 0 0 0
S3 2/6 0 0 0 0 1/6 0 0 0 0 0 0 0 1/6 1/6 1/6.
IDF
S1 0.083 0 0.83 0.83 0.083 0.083 0.083 0.083 0.083 0.083 0.083 0.083 0 0 0 0
S2 0.11 0.11 0.11 0.11 0.11 0.11 0 0 0.11 0 0 0.11 0.11 0 0 0
S3 0.167 0.167 0 0 0 0.167 0 0 0 0 0 0 0 0.167 0.167 0.167.
TF-IDF
S1 0.193 0.498 0.249 0.249 0.249 0.193 0.327 0.327 0.249 0.327 0.327 0 0 0 0 0
S2 0.250 0.322 0.322 0.322 0.322 0.250 0 0 0.322 0 0 0.423 0.423 0 0 0
S3 0.542 0 0 0 0 0.271 0 0 0 0 0 0 0 0.459 0.459 0.459.

Question 2:

```
#December 11, 2023
#CSC461 – Assignment4 – NLP
#Numan Latif
#FA21-BSE-039
#In this task we calculate the Cosine, manhattan, and euclidean distances.
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity, manhattan_distances,
euclidean_distances

sentence = [
    "data science is one of the most important courses in computer science",
    "this is one of the best data science courses",
    "the data scientists perform data analysis"
]

vectorizer = TfidfVectorizer()
X_tfidf = vectorizer.fit_transform(sentence)

cosine_sim = cosine_similarity(X_tfidf)
print("\nCosine :")
print(pd.DataFrame(cosine_sim, index=['S1', 'S2', 'S3'], columns=['S1', 'S2', 'S3']).round(3))

manhattan_dist = manhattan_distances(X_tfidf)
print("\nManhattan Distance:")
print(pd.DataFrame(manhattan_dist, index=['S1', 'S2', 'S3'], columns=['S1', 'S2',
'S3']).round(3))

euclidean_dist = euclidean_distances(X_tfidf)
```

```
print("\nEuclidean Distance:")

print(pd.DataFrame(euclidean_dist, index=['S1', 'S2', 'S3'], columns=['S1', 'S2',
'S3']).round(3))
```

```
Cosine :
       S1     S2     S3
S1  1.000  0.577  0.157
S2  0.577  1.000  0.203
S3  0.157  0.203  1.000

Manhattan Distance:
       S1     S2     S3
S1  0.000  2.736  4.608
S2  2.736  0.000  4.146
S3  4.608  4.146  0.000

Euclidean Distance:
       S1     S2     S3
S1  0.000  0.919  1.298
S2  0.919  0.000  1.262
S3  1.298  1.262  0.000
```

Cosine Similarity

S1 and S2 = $S1 \cdot S2 / |S1| \cdot |S2|$ = 0.577.

S1 and S3 = $S1 \cdot S3 / |S1| \cdot |S3|$ = ~~0.203~~ 0.157

S2 and S3 = $S2 \cdot S3 / |S2| \cdot |S3|$. = 0.203.

Manhattan Distance

S1 and S2 :- $\Sigma |S1i - S2i|$ = 2.736

S2 and S3 :- $\Sigma |S2i - S3i|$ = 4.146

S1 and S3 :- $\Sigma |S1i - S3i|$ = 4.608.

Euclidean Distance

S1 and S2 : $\Sigma (S1i - S2i)^{(1/2)^2}$ = 0.919.

S1 and S3 : $\Sigma (S1i - S3i)^{(1/2)^2}$ = 1.298

S2 and S3 : $\Sigma (S2i - S3i)^{(1/2)^2}$ = 1.262.