HomeWork-7 Report Template

1. Data Preparation
    a. Storage strategy using ElasticSearch or an alternative
        i. Email data was stored as a csv file using pandas which was used for data manipulation. While ElasticSearch could offer scalability and efficient querying, pandas provides a convenient and familiar way for smaller-scale projects

2. Feature Extraction and Model Training
    a. Manual Spam Features (Part 1)
        i. Description of the process for creating n-gram lists
            1. N-gram lists such as spam_words_trial_a and spam_words_trial_b were manually curated based on domain knowledge and analysis of spam email content.
        ii. Methodology for querying ElasticSearch for feature values
            1. Feature extraction was performed directly from the email content using tokenization techniques like word tokenization from NLTK and the CountVectorizer using the curated ngram list is used.
    b. All Unigrams as Features (Part 2, MS Students)
        i. Approach for extracting all unigrams
            1. CountVectorizer is used to extract the unigrams when provided with the vocabulary of spam words
        ii. Details of sparse matrix representation
            1. Unigrams are represented using a sparse matrix format, such as the one provided by CountVectorizer from scikit-learn. This format efficiently represents large matrices with mostly zero values, saving memory and computation time.
    c. Give a description of the machine learning algorithms used for training
        i. Logistic Regression, Decision Trees, and Naive Bayes classifiers are utilized for training the models.
            1. Logistic Regression with L1 regularization is chosen for its ability to handle sparse data and feature selection.
            2. Decision Trees provide interpretability and can handle non-linear relationships between features.
            3. Naive Bayes is chosen for its simplicity and ability to handle large feature spaces efficiently.

3. Testing and Evaluation
    a. Approach taken for testing the model on the test dataset
        i. The model is tested on the test dataset using standard evaluation metrics such as ROC-AUC score and classification reports.
        ii. Additionally, the top spam documents predicted by each model are identified for further analyze if the predictions are right.
    b. Analysis of results from different algorithms (with screenshots)

```
c.  ************************************************** Part 1 : Trial A
    **************************************************
d.  Score for logistic regression is: 0.6167964839198701
e.               precision    recall  f1-score   support
f.
g.           0       0.51      0.01      0.02      5039
h.           1       0.65      1.00      0.79      9271
i.
j.     accuracy                          0.65     14310
k.    macro avg       0.58      0.50      0.40     14310
l.  weighted avg       0.60      0.65      0.51     14310
m.
n.     Top 10 spam docs for logistic regression are:  [46096, 13628,
    75309, 52106, 46854, 40876, 63965, 32640, 25145, 9552]
o.  Score for decision tree is: 0.6477638586857695
p.               precision    recall  f1-score   support
q.
r.           0       0.51      0.01      0.02      5039
s.           1       0.65      1.00      0.79      9271
t.
u.     accuracy                          0.65     14310
v.    macro avg       0.58      0.50      0.40     14310
w.  weighted avg       0.60      0.65      0.51     14310
x.
y.     Top 10 spam docs for decision tree are:  [42533, 13628, 54855,
    61837, 69894, 71745, 32444, 48829, 41902, 563]
z.  Score for naive bayes is: 0.6141348800679263
aa.              precision    recall  f1-score   support
bb.
cc.          0       0.51      0.01      0.02      5039
dd.          1       0.65      1.00      0.79      9271
ee.
ff.    accuracy                          0.65     14310
gg.   macro avg       0.58      0.50      0.40     14310
hh. weighted avg       0.60      0.65      0.51     14310
ii.
```

```
jj.    Top 10 spam docs for naive bayes are:  [46096, 13628, 75309, 52106,
   46854, 40876, 63965, 9552, 25145, 16342]
kk. ************************************************** Part 1 : Trial B
   **************************************************
ll. Score for logistic regression is: 0.7764034854528806
mm.                precision    recall  f1-score   support
nn.
oo.           0       0.64      0.26      0.37      5039
pp.           1       0.70      0.92      0.79      9271
qq.
rr.    accuracy                           0.69     14310
ss.   macro avg       0.67      0.59      0.58     14310
tt. weighted avg       0.67      0.69      0.64     14310
uu.
vv.    Top 10 spam docs for logistic regression are:  [49305, 13628,
   46096, 49254, 53923, 49993, 50806, 49695, 49769, 50450]
ww.   Score for decision tree is: 0.8480094717572257
xx.             precision    recall  f1-score   support
yy.
zz.           0       0.64      0.26      0.37      5039
aaa.              1       0.70      0.92      0.79      9271
bbb.
ccc.      accuracy                           0.69     14310
ddd.     macro avg       0.67      0.59      0.58     14310
eee.   weighted avg       0.67      0.69      0.64     14310
fff.
ggg.      Top 10 spam docs for decision tree are:  [14159, 11880, 45090,
   39522, 70822, 60923, 74013, 27623, 49912, 48350]
hhh.   Score for naive bayes is: 0.7462891056918158
iii.            precision    recall  f1-score   support
jjj.
kkk.              0       0.64      0.26      0.37      5039
lll.          1       0.70      0.92      0.79      9271
mmm.
nnn.      accuracy                           0.69     14310
ooo.     macro avg       0.67      0.59      0.58     14310
```

```
ppp.   weighted avg       0.67      0.69      0.64     14310
qqq.
rrr.    Top 10 spam docs for naive bayes are:  [44005, 38197, 19411, 25739,
   26471, 15995, 18413, 38832, 43745, 25219]
sss.   ************************************************** Part 2 :
   Unigram **************************************************
ttt. Score for logistic regression is: 0.9975632949414586
uuu.               precision    recall  f1-score   support
vvv.
www.            0       0.99      0.99      0.99      5039
xxx.            1       0.99      0.99      0.99      9271
yyy.
zzz.      accuracy                          0.99     14310
aaaa.    macro avg       0.99      0.99      0.99     14310
bbbb.  weighted avg       0.99      0.99      0.99     14310
cccc.
dddd.     Top 10 spam docs for logistic regression are:  [50660, 2612,
   51046, 58783, 15707, 69690, 16653, 43269, 31596, 44005]
eeee.  Score for decision tree is: 0.9828964216100716
ffff.            precision    recall  f1-score   support
gggg.
hhhh.           0       0.99      0.99      0.99      5039
iiii.         1       0.99      0.99      0.99      9271
jjjj.
kkkk.     accuracy                          0.99     14310
llll.   macro avg       0.99      0.99      0.99     14310
mmmm.      weighted avg       0.99      0.99      0.99     14310
nnnn.
oooo.     Top 10 spam docs for decision tree are:  [40247, 44821, 19939,
   3911, 52908, 67049, 6699, 68386, 16112, 58176]
pppp.  Score for naive bayes is: 0.9875525319507088
qqqq.              precision    recall  f1-score   support
rrrr.
ssss.           0       0.99      0.99      0.99      5039
tttt.         1       0.99      0.99      0.99      9271
uuuu.
```

```
vvvv.      accuracy                           0.99     14310
wwww.    macro avg       0.99      0.99      0.99     14310
xxxx.  weighted avg       0.99      0.99      0.99     14310
yyyy.
zzzz.      Top 10 spam docs for naive bayes are:  [14159, 60233, 60970,
   8868, 7298, 58784, 1403, 13832, 11242, 53677]
aaaaa. ************************************************** Extra Credit :
   Bigram **************************************************
bbbbb. Score for logistic regression is: 0.9969771217573791
ccccc.              precision    recall  f1-score   support
ddddd.
eeeee.            0       0.99      0.97      0.98      5039
fffff.            1       0.99      0.99      0.99      9271
ggggg.
hhhhh.     accuracy                           0.99     14310
iiiii.    macro avg       0.99      0.98      0.98     14310
jjjjj. weighted avg       0.99      0.99      0.99     14310
kkkkk.
lllll.     Top 10 spam docs for logistic regression are:  [36012, 41648,
   13904, 50907, 32640, 54667, 13628, 8550, 41912, 45141]
mmmmm.      Score for decision tree is: 0.9832611851268444
nnnnn.              precision    recall  f1-score   support
ooooo.
ppppp.            0       0.99      0.97      0.98      5039
qqqqq.            1       0.99      0.99      0.99      9271
rrrrr.
sssss.     accuracy                           0.99     14310
ttttt.    macro avg       0.99      0.98      0.98     14310
uuuuu. weighted avg       0.99      0.99      0.99     14310
vvvvv.
wwwww.           Top 10 spam docs for decision tree are:  [14159, 11143,
   17605, 70727, 61988, 25043, 43144, 70205, 64298, 748]
xxxxx. Score for naive bayes is: 0.9904238044536191
yyyyy.              precision    recall  f1-score   support
zzzzz.
aaaaaa.            0       0.99      0.97      0.98      5039
```

```
bbbbbb.                 1       0.99      0.99      0.99      9271
cccccc.
dddddd.         accuracy                            0.99     14310
eeeeee.        macro avg       0.99      0.98      0.98     14310
ffffff.   weighted avg       0.99      0.99      0.99     14310
gggggg.
hhhhhh.         Top 10 spam docs for naive bayes are:  [14159, 32012,
   19268, 42105, 30449, 62535, 3872, 28878, 60830, 70143]
iiiiii.
```

4. Results and Discussion
   a. Summary of key findings from testing the models
      i. Models trained using manual spam features achieve moderate performance, with ROC-AUC scores ranging from 0.61 to 0.77.
      ii. Models trained using all unigrams as features demonstrate significantly higher performance, with ROC-AUC scores close to 1.00, indicating near-perfect classification.
      iii. Analysis of feature importance reveals that manually curated spam features may not capture all relevant information present in the data.
   b. Feature analysis and comparison with manual spam features
      i. Comparison between manual spam features and all unigrams shows that the latter provides a more comprehensive representation of email content.
      ii. Unigrams capture nuanced patterns and variations in language usage, leading to improved model performance.

5. Extra Credit:
   a. Application to HW3 crawl data and feature expansion (if attempted)
   b. Extracting Bigrams as Features:
      i. Bigrams, or sequences of two adjacent words, can be extracted alongside unigrams to capture contextual information.
      ii. Similar to unigrams, bigrams are represented using sparse matrices, allowing for efficient storage and computation.