

You are currently looking at **version 1.1** of this notebook. To download notebooks and datafiles, as well as get help on Jupyter notebooks in the Coursera platform, visit the [Jupyter Notebook FAQ \(https://www.coursera.org/learn/python-text-mining/resources/d9pwm\)](https://www.coursera.org/learn/python-text-mining/resources/d9pwm) course resource.

Assignment 3

In this assignment you will explore text message data and create models to predict if a message is spam or not.

```
In [1]: # import libraries
import pandas as pd
import numpy as np
import re

from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer

from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import roc_auc_score

from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
```

```
In [2]: df = pd.read_csv('spam.csv')

# spam =1, not spam = 0
df['target'] = np.where(df['target']=='spam',1,0)
df.head(10)
```

```
Out[2]:
```

| | text | target |
|---|---|--------|
| 0 | Go until jurong point, crazy.. Available only ... | 0 |
| 1 | Ok lar... Joking wif u oni... | 0 |
| 2 | Free entry in 2 a wkly comp to win FA Cup fina... | 1 |
| 3 | U dun say so early hor... U c already then say... | 0 |
| 4 | Nah I don't think he goes to usf, he lives aro... | 0 |
| 5 | FreeMsg Hey there darling it's been 3 week's n... | 1 |
| 6 | Even my brother is not like to speak with me. ... | 0 |
| 7 | As per your request 'Melle Melle (Oru Minnamin... | 0 |
| 8 | WINNER!! As a valued network customer you have... | 1 |
| 9 | Had your mobile 11 months or more? U R entitle... | 1 |

```
In [3]: X_train, X_test, y_train, y_test = train_test_split(df['text'],
                                                             df['target'],
                                                             random_state=0)
```

Question 1

What percentage of the documents in spam_data are spam?

*This function should return a float, the percent value (i.e. \$ratio 100\$).**

```
In [4]: def answer_one():

    spam = len(df[df['target'] == 1])
    total = len(df)
    result = spam / total * 100

    return result
```

```
In [5]: answer_one()
```

```
Out[5]: 13.406317300789663
```

Question 2

Fit the training data `X_train` using a Count Vectorizer with default parameters.

What is the longest token in the vocabulary?

This function should return a string.

```
In [6]: def answer_two():

    vect = CountVectorizer().fit(X_train)
    result = max(vect.get_feature_names(), key = len)

    return result
```

```
In [7]: answer_two()
```

```
Out[7]: 'com1win150ppmx3age16subscription'
```

Question 3

Fit and transform the training data `X_train` using a Count Vectorizer with default parameters.

Next, fit a multinomial Naive Bayes classifier model with smoothing $\alpha=0.1$. Find the area under the curve (AUC) score using the transformed test data.

This function should return the AUC score as a float.

```
In [8]: def answer_three():

    vect = CountVectorizer().fit(X_train)
    X_train_vectorized = vect.transform(X_train)

    model = MultinomialNB(alpha=0.1).fit(X_train_vectorized, y_train)
    predictions = model.predict(vect.transform(X_test))

    result = roc_auc_score(y_test, predictions)

    return result
```

```
In [9]: answer_three()
```

```
Out[9]: 0.97208121827411165
```

Question 4

Fit and transform the training data `X_train` using a `TfidfVectorizer` with default parameters.

What 20 features have the smallest tf-idf and what 20 have the largest tf-idf?

Put these features in a two series where each series is sorted by tf-idf value and then alphabetically by feature name. The index of the series should be the feature name, and the data should be the tf-idf.

The series of 20 features with smallest tf-idfs should be sorted smallest tfidf first, the list of 20 features with largest tf-idfs should be sorted largest first.

This function should return a tuple of two series (smallest tf-idfs series, largest tf-idfs series).

```
In [10]: def answer_four():

    vect = TfidfVectorizer().fit(X_train)
    feature_names = np.array(vect.get_feature_names())
    X_train_vectorized = vect.transform(X_train)

    max_tfidf = X_train_vectorized.max(0).toarray()[0]
    index = max_tfidf.argsort()
    values = sorted_tf_idfs = max_tfidf[index]

    smallest = pd.Series(values[:20], index=feature_names[index[:20]])
    largest = pd.Series(values[-20:][::-1], index=feature_names[index[-20:][::-1]])

    return smallest, largest
```

```
In [11]: answer_four()
```

```
Out[11]: (sympathetic      0.074475
healer      0.074475
aaniye      0.074475
dependable  0.074475
companion   0.074475
listener    0.074475
athletic    0.074475
exterminator 0.074475
psychiatrist 0.074475
pest        0.074475
determined  0.074475
chef        0.074475
courageous  0.074475
stylist     0.074475
psychologist 0.074475
organizer   0.074475
pudunga     0.074475
venaam      0.074475
diwali      0.091250
mornings    0.091250
dtype: float64, 146tf150p    1.000000
havent      1.000000
home        1.000000
okie        1.000000
thanx       1.000000
er          1.000000
anything    1.000000
lei         1.000000
nite        1.000000
yup         1.000000
thank       1.000000
ok          1.000000
where       1.000000
beverage    1.000000
anytime     1.000000
too         1.000000
done        1.000000
645         1.000000
tick        0.980166
blank       0.932702
dtype: float64)
```

Question 5

Fit and transform the training data `X_train` using a `TfidfVectorizer` ignoring terms that have a document frequency strictly lower than 3.

Then fit a multinomial Naive Bayes classifier model with smoothing `alpha=0.1` and compute the area under the curve (AUC) score using the transformed test data.

This function should return the AUC score as a float.

```
In [12]: def answer_five():

    vect = TfidfVectorizer(min_df=3).fit(X_train)
    X_train_vectorized = vect.transform(X_train)

    model = MultinomialNB(alpha=0.1).fit(X_train_vectorized, y_train)

    X_test_vectorized = vect.transform(X_test)
    predictions = model.predict(X_test_vectorized)

    result = roc_auc_score(y_test, predictions)

    return result
```

```
In [13]: answer_five()
```

```
Out[13]: 0.94162436548223349
```

Question 6

What is the average length of documents (number of characters) for not spam and spam documents?

This function should return a tuple (average length not spam, average length spam).

```
In [14]: def answer_six():

    # spam =1, not spam = 0
    avg_len_not_spam = np.mean([len(x) for x in df[df['target'] == 0].text])
    avg_len_spam = np.mean([len(x) for x in df[df['target'] == 1].text])

    return (avg_len_not_spam, avg_len_spam )
```

```
In [15]: answer_six()
```

```
Out[15]: (71.023626943005183, 138.8661311914324)
```

The following function has been provided to help you combine new features into the training data:

```
In [16]: def add_feature(X, feature_to_add):
    """
    Returns sparse feature matrix with added feature.
    feature_to_add can also be a list of features.
    """
    from scipy.sparse import csr_matrix, hstack
    return hstack([X, csr_matrix(feature_to_add).T], 'csr')
```

Question 7

Fit and transform the training data `X_train` using a `TfidfVectorizer` ignoring terms that have a document frequency strictly lower than 5.

Using this document-term matrix and an additional feature, **the length of document (number of characters)**, fit a Support Vector Classification model with regularization `C=10000`. Then compute the area under the curve (AUC) score using the transformed test data.

This function should return the AUC score as a float.

```
In [17]: def answer_seven():

    vect = TfidfVectorizer(min_df=5).fit(X_train)
    X_train_vectorized = vect.transform(X_train)
    X_test_vectorized = vect.transform(X_test)

    Train = add_feature(X_train_vectorized, [len(x) for x in X_train])
    model = SVC(C=10000).fit(Train, y_train)

    Test = add_feature(X_test_vectorized, [len(x) for x in X_test])
    predictions = model.predict(Test)

    result = roc_auc_score(y_test, predictions)

    return result
```

```
In [18]: answer_seven()
```

```
Out[18]: 0.95813668234215565
```

Question 8

What is the average number of digits per document for not spam and spam documents?

This function should return a tuple (average # digits not spam, average # digits spam).

```
In [19]: def answer_eight():

    not_spam = [re.findall('[0-9]',i) for i in df['text'][df.target==0]]
    spam = [re.findall('[0-9]',i) for i in df['text'][df.target==1]]

    avg_not_spam = np.mean(list(map(len,not_spam)))
    avg_spam = np.mean(list(map(len,spam)))

    return avg_not_spam, avg_spam
```

```
In [20]: answer_eight()
```

```
Out[20]: (0.29927461139896372, 15.759036144578314)
```

Question 9

Fit and transform the training data `X_train` using a `Tfidf Vectorizer` ignoring terms that have a document frequency strictly lower than **5** and using **word n-grams from n=1 to n=3** (unigrams, bigrams, and trigrams).

Using this document-term matrix and the following additional features:

- the length of document (number of characters)
- **number of digits per document**

fit a Logistic Regression model with regularization `C=100`. Then compute the area under the curve (AUC) score using the transformed test data.

This function should return the AUC score as a float.

```
In [21]: def answer_nine():

    digit_train=[sum(char.isnumeric() for char in x) for x in X_train]
    digit_test=[sum(char.isnumeric() for char in x) for x in X_test]

    vect = TfidfVectorizer(min_df = 5, ngram_range = (1,3)).fit(X_train)

    X_train_vectorized = vect.transform(X_train)
    X_test_vectorized = vect.transform(X_test)

    Train = add_feature(X_train_vectorized, digit_train)
    model = LogisticRegression(C=100).fit(Train, y_train)

    Test = add_feature(X_test_vectorized , digit_test)
    predictions = model.predict(Test)

    result = roc_auc_score(y_test, predictions)

    return result
```

```
In [22]: answer_nine()
```

```
Out[22]: 0.96787090640544626
```

Question 10

What is the average number of non-word characters (anything other than a letter, digit or underscore) per document for not spam and spam documents?

Hint: Use `\w` and `\W` character classes

This function should return a tuple (average # non-word characters not spam, average # non-word characters spam).

```
In [23]: def answer_ten():

    not_spam = [re.findall('\W',i) for i in df['text'][df.target==0]]
    spam = [re.findall('\W',i) for i in df['text'][df.target==1]]

    avg_not_spam = np.mean(list(map(len,not_spam)))
    avg_spam = np.mean(list(map(len,spam)))

    return avg_not_spam, avg_spam
```

```
In [24]: answer_ten()
```

```
Out[24]: (17.291813471502589, 29.041499330655956)
```

Question 11

Fit and transform the training data `X_train` using a Count Vectorizer ignoring terms that have a document frequency strictly lower than **5** and using **character n-grams from n=2 to n=5**.

To tell Count Vectorizer to use character n-grams pass in `analyzer='char_wb'` which creates character n-grams only from text inside word boundaries. This should make the model more robust to spelling mistakes.

Using this document-term matrix and the following additional features:

- the length of document (number of characters)
- number of digits per document
- **number of non-word characters (anything other than a letter, digit or underscore.)**

fit a Logistic Regression model with regularization `C=100`. Then compute the area under the curve (AUC) score using the transformed test data.

Also **find the 10 smallest and 10 largest coefficients from the model** and return them along with the AUC score in a tuple.

The list of 10 smallest coefficients should be sorted smallest first, the list of 10 largest coefficients should be sorted largest first.

The three features that were added to the document term matrix should have the following names should they appear in the list of coefficients: `['length_of_doc', 'digit_count', 'non_word_char_count']`

This function should return a tuple (AUC score as a float, smallest coefs list, largest coefs list).

```
In [25]: def count_digits(str):  
        return sum(i.isdigit() for i in str)
```

```
In [26]: def answer_eleven():  
  
    vect = CountVectorizer(min_df=5, ngram_range=(2,5), analyzer='char_wb').fit(X_train)  
  
    X_train_vectorized = vect.transform(X_train)  
    X_test_vectorized = vect.transform(X_test)  
  
    train1 = np.array([len(i) for i in X_train])  
    train2 = np.array([count_digits(i) for i in X_train])  
    train3 = np.array(X_train.str.findall(r'^a-zA-Z0-9_')).str.len()  
    all_train_formats = np.stack((train1, train2, train3))  
  
    test1 = np.array([len(i) for i in X_test])  
    test2 = np.array([count_digits(i) for i in X_test])  
    test3 = np.array(X_test.str.findall(r'^a-zA-Z0-9_')).str.len()  
    all_test_formats = np.stack((test1, test2, test3))  
  
    X_train_new = add_feature(X_train_vectorized, all_train_formats)  
    X_test_new = add_feature(X_test_vectorized, all_test_formats)  
  
    model = LogisticRegression(C=100).fit(X_train_new, y_train)  
    predictions = model.predict(X_test_new)  
  
    feature_names = np.array(vect.get_feature_names())  
    feature_names = np.append(feature_names, ['A', 'B', 'C'])  
  
    index = model.coef_[0].argsort()  
  
    smallest = feature_names[index[:10]]  
    largest = feature_names[index[-10:]]  
    largest = np.flip(largest,0)  
  
    return (roc_auc_score(y_test, predictions), smallest, largest)
```

```
In [27]: answer_eleven()
```

```
Out[27]: (0.97885931107074342,  
array(['.', ' ', '...', '? ', ' i', ' y', ' go', ':'), ' h', 'go', ' m'],  
      dtype='<U5'),  
array(['B', 'ne', 'co', 'ia', 'xt', ' ch', ' x', 'mob', 'ww', 'an'],  
      dtype='<U5'))
```