Case Study 3: Email Spam

Matt Farrow

June 6, 2022

1 Introduction

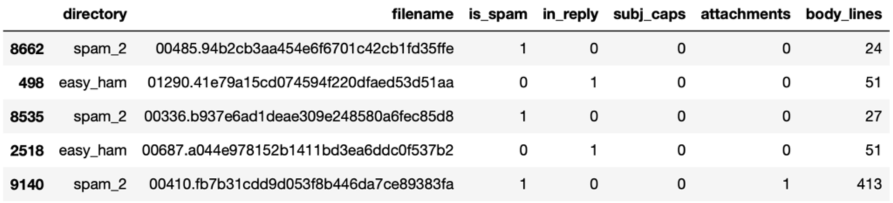
In this case study, my goal is to build a email spam classifier. The data for this analysis comes from Apache's Spam Assassin dataset, and is subset into spam messages and ham (a play on words to denote non-spam messages).

The objective of this case study is to use clustering and Naïve Bayes into order build a classifier than can separate the spam and non-spam messages. One of the challenges in this project is determining how to weight the classifier. Making it too aggressive may lead to non-spam messages being flagged as spam, too lax and the user may find themselves with unreasonable amounts of spam in their inbox.

2 Methods

2.1 Data Examination

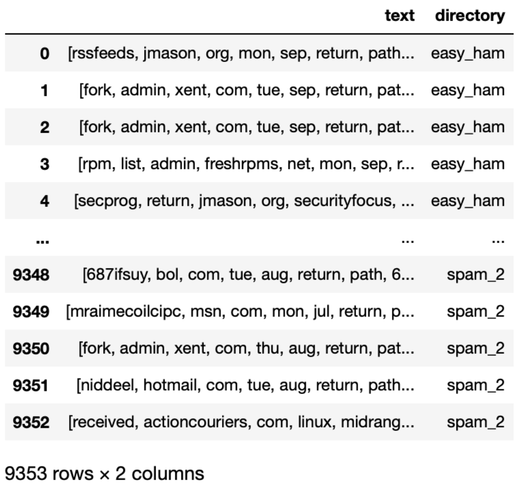
The initial data set is comprised of email messages grouped into one of five folders: `easy_ham`, `easy_ham_2`, `hard_ham`, `spam`, and `spam_2`. The first order of business was to read in the messages, identify which directory they came from, and where spam was found or not found.



	directory	filename	is_spam	in_reply	subj_caps	attachments	body_lines
8662	spam_2	00485.94b2cb3aa454e6f6701c42cb1fd35ffe	1	0	0	0	24
498	easy_ham	01290.41e79a15cd074594f220dfaed53d51aa	0	1	0	0	51
8535	spam_2	00336.b937e6ad1deae309e248580a6fec85d8	1	0	0	0	27
2518	easy_ham	00687.a044e978152b1411bd3ea6ddc0f537b2	0	1	0	0	51
9140	spam_2	00410.fb7b31cdd9d053f8b446da7ce89383fa	1	0	0	1	413

Figure 1: Results of Initial Email Load

Next, the data was normalized through a text cleanup process that removed special characters; converted text to lowercase; removed common, or stop, words; tokenized the document, and removed short tokens that might trip up the classifier.



	text	directory
0	[rssfeeds, jmason, org, mon, sep, return, path...	easy_ham
1	[fork, admin, xent, com, tue, sep, return, pat...	easy_ham
2	[fork, admin, xent, com, tue, sep, return, pat...	easy_ham
3	[rpm, list, admin, freshrpms, net, mon, sep, r...	easy_ham
4	[secprog, return, jmason, org, securityfocus, ...	easy_ham
...	...	...
9348	[687ifsuy, bol, com, tue, aug, return, path, 6...	spam_2
9349	[mraimecoilcipc, msn, com, mon, jul, return, p...	spam_2
9350	[fork, admin, xent, com, thu, aug, return, pat...	spam_2
9351	[niddeel, hotmail, com, tue, aug, return, path...	spam_2
9352	[received, actioncouriers, com, linux, midrang...	spam_2

9353 rows × 2 columns

Figure 2: Post-Text Normalization

After normalizing the text, linear discriminant analysis (LDA) clustering was performed using Mallet, a Java-based package that is designed for working with language. In my case, I used it to perform clustering and topic modeling on the emails. 15 clusters were created with word weights showing the most important terms in each of the clusters.

Word Weight	Term
0.0607	com
0.0288	net
0.0239	received
0.0195	localhost
0.0125	org
0.0115	jul
0.0114	esmtp
0.0109	netnoteinc
0.0101	content
0.0101	http
0.0097	zzzz
0.009	aug
0.0085	mail
0.0076	mon
0.0075	sep

Table 1: Cluster Example

Each topic was then examined by its frequency within each of the original directories using a heatmap.

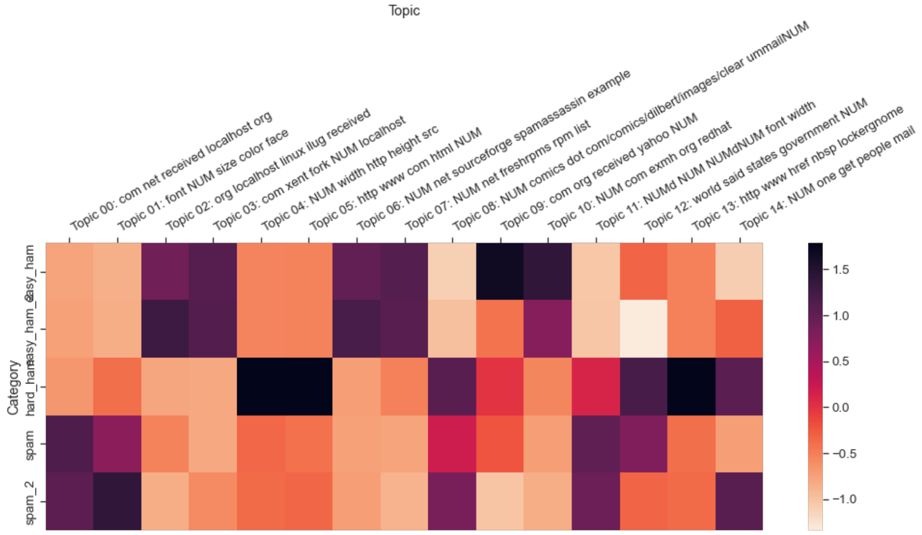


Figure 3: Clustering Heatmap

In addition to the LDA clustering that was undertaken with Mallet, an additional clustering was done using K-means clustering and an elbow plot was generated to examine the most appropriate cluster number to use.

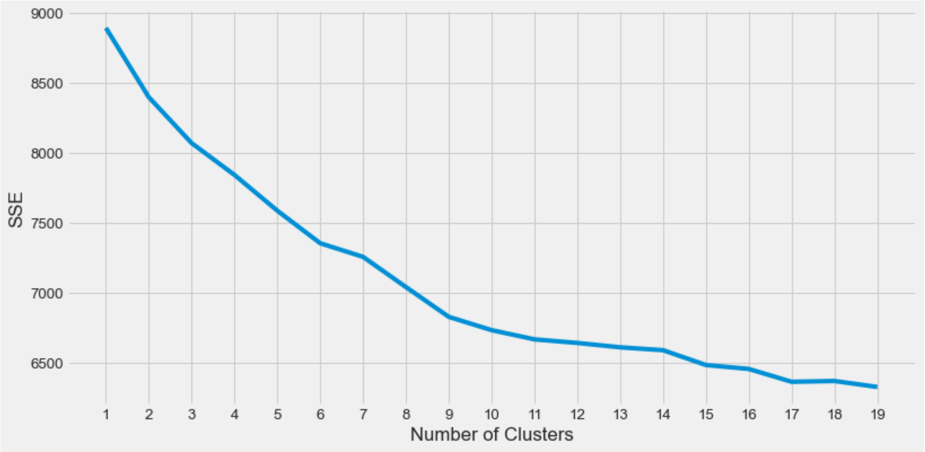


Figure 4: K-Means Clustering Elbow Plot

2.2 Model Preparation & Execution

Two approaches were taken to model building. In both, the original data was run through a Multinomial Naïve Bayes classifier. Numeric values were scaled and missing values were imputed using the median. Categorical values were one-hot encoded with missing values being replaced by the most frequent result.

3 Results

3.1 Model Results

As seen in **Table 2: Model 1 Performance MetricsTable 2**, the accuracy of this model was only about 79% with a recall of approximately 24%, indicating a poor performance when it comes to filtering spam out of user's inboxes. That fairly poor result can be further seen in the ROC curve (**Figure 5**) as well as the model's confusion matrix **(Figure 6)**.

Metric	Value
Accuracy	0.7857
Recall	0.2375
Precision	0.7651
F1	0.3625

Table 2: Model 1 Performance Metrics

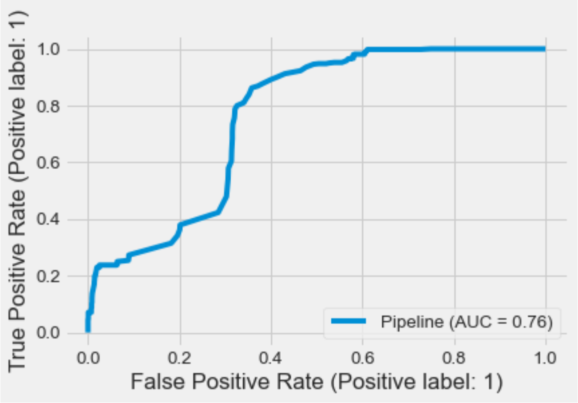


Figure 5: Model 1 ROC Curve

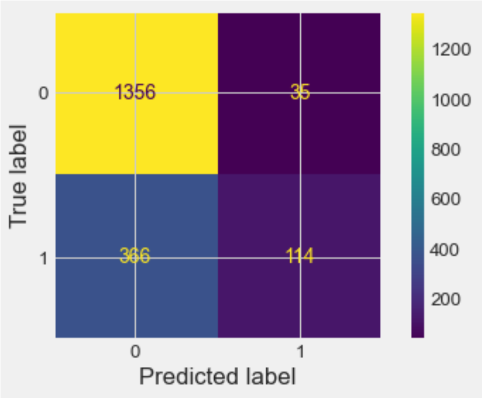


Figure 6: Model 1 Confusion Matrix

In the second model, the same pre-processing steps were taken as in the first, but with the addition of the clustering that was previously undertaken.

Metric	Value
Accuracy	0.8856
Recall	0.8979
Precision	0.7232
F1	0.8011

Table 3: Model 2 Performance Metrics

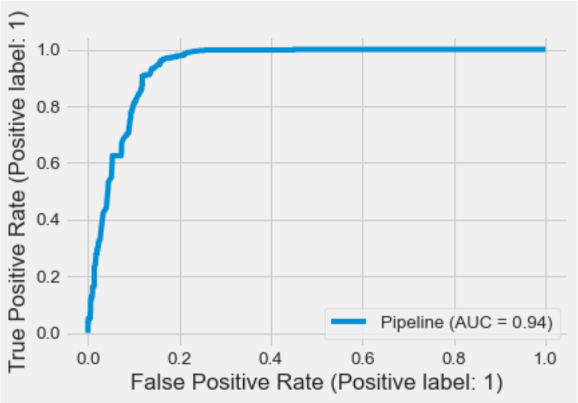


Figure 7: Model 2 ROC Curve

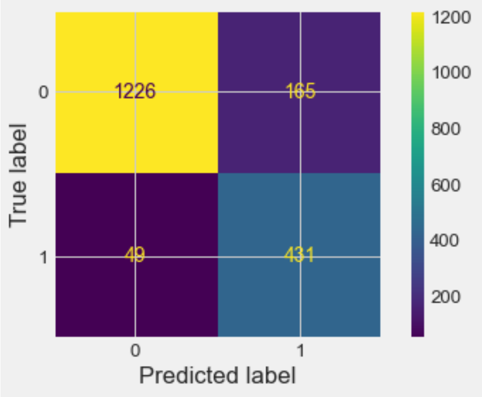


Figure 8: Model 2 Confusion Matrix

4 Conclusion

The second model that was built on the clustering results showed significantly improved recall as well as improved precision and accuracy. In addition to increasing the ability of our classifier to sift spam out of email inboxes, the inclusion of clustering means that our model can continue to adapt as new data is fed in in the form of new emails as well as user behavior in flagging messages as spam or moving them out of spam into their inboxes.

Appendix

Code

Code begins on the following page.

Case Study 3

Build a spam classifier using naive Bayes and clustering. You will have to create your own dataset from the input messages. Be sure to document how you created your dataset.

Import Libraries

In [1]:

```
import email
import os
import re
import string
import warnings
from collections import Counter
from html.parser import HTMLParser
from os import chdir, getcwd, listdir
from os.path import dirname, isfile, join, realpath

import little_mallet_wrapper as lmw
import matplotlib.pyplot as plt
import nltk
import numpy as np
import pandas as pd
import scipy
import seaborn as sns
import sklearn.cluster as cluster
from bs4 import BeautifulSoup
from gensim.corpora import Dictionary
from gensim.models.coherencemodel import CoherenceModel
from gensim.models.ldamodel import LdaModel
from imblearn.over_sampling import SMOTE
from imblearn.pipeline import Pipeline as imbpipeline
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer, WordNetLemmatizer
from scipy.sparse import csr_matrix

# Import sklearn libraries
from sklearn.calibration import CalibratedClassifierCV
from sklearn.cluster import KMeans
from sklearn.compose import ColumnTransformer
from sklearn.compose import make_column_selector as selector
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.impute import SimpleImputer
from sklearn.manifold import TSNE
from sklearn.metrics import (
    ConfusionMatrixDisplay,
    RocCurveDisplay,
    accuracy_score,
    auc,
    classification_report,
    confusion_matrix,
    f1_score,
    precision_score,
    recall_score,
    roc_curve,
    silhouette_samples,
    silhouette_score,
)
from sklearn.model_selection import (
    GridSearchCV,
    KFold,
    StratifiedKFold,
    cross_val_predict,
    cross_val_score,
    cross_validate,
    train_test_split,
)
from sklearn.naive_bayes import ComplementNB, MultinomialNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import OneHotEncoder, RobustScaler
from wordcloud import WordCloud

warnings.filterwarnings('ignore')
%matplotlib inline
```

Examine the Data

In [2]:

```
# Look at the contents of the current directory
os.listdir('.')
```

Out[2]:

```
['categories_by_topics.pdf',
 'Farrow_Matt_case_study_3.docx',
 '.DS_Store',
 '~$rrow_Matt_case_study_3.docx',
 'mallet.topic_keys.15',
 'mallet.topic_distributions.15',
 'CleanShot 2022-05-22 at 12.45.01@2x.png',
 'training.txt',
 'Farrow_Matt_case_study_3.py',
 'mallet.word_weights.15',
 'mallet.training',
 'mallet.diagnostics.15.xml',
 'mallet.model.15',
 'SpamAssassinMessages',
 '.ipynb_checkpoints',
 'Farrow_Matt_case_study_3.ipynb',
 'data']
```

In [3]:

```
# Look at the contents of the `SpamAssassinMessages` folder
os.listdir('./SpamAssassinMessages')
```

Out[3]:

```
['spam', 'hard_ham', 'spam_2', '.DS_Store', 'easy_ham', 'easy_ham_2']
```

In [4]:
```
# Glance at one of the emails to understand the raw data
os.system('cat ./SpamAssassinMessages/easy_ham/2551.3b1f94418de5bd544c977b44
```

```
From rssfeeds@jmason.org  Thu Oct 10 12:32:34 2002
Return-Path: <rssfeeds@example.com>
Delivered-To: yyyy@localhost.example.com
Received: from localhost (jalapeno [127.0.0.1])
        by jmason.org (Postfix) with ESMTP id 89EE616F03
        for <jm@localhost>; Thu, 10 Oct 2002 12:32:33 +0100 (IST)
Received: from jalapeno [127.0.0.1]
        by localhost with IMAP (fetchmail-5.9.0)
        for jm@localhost (single-drop); Thu, 10 Oct 2002 12:32:33 +0100 (IST
)
Received: from dogma.slashnull.org (localhost [127.0.0.1]) by
    dogma.slashnull.org (8.11.6/8.11.6) with ESMTP id g9A84QK14194 for
    <jm@jmason.org>; Thu, 10 Oct 2002 09:04:26 +0100
Message-Id: <200210100804.g9A84QK14194@dogma.slashnull.org>
To: yyyy@example.com
From: newscientist <rssfeeds@example.com>
Subject: Critical US satellites could be hacked
Date: Thu, 10 Oct 2002 08:04:26 -0000
Content-Type: text/plain; encoding=utf-8
X-Spam-Status: No, hits=-959.4 required=5.0
        tests=AWL,DATE_IN_PAST_03_06,T_NONSENSE_FROM_00_10,
              T_NONSENSE_FROM_10_20,T_NO
```

Out[4]:
```
0
```

```
NSENSE_FROM_20_30,
              T_NONSENSE_FROM_30_40,T_NONSENSE_FROM_40_50,
              T_NONSENSE_FROM_50_60,T_NONSENSE_FROM_60_70,
              T_NONSENSE_FROM_70_80,T_NONSENSE_FROM_80_90,
              T_NONSENSE_FROM_90_91,T_NONSENSE_FROM_91_92,
              T_NONSENSE_FROM_92_93,T_NONSENSE_FROM_93_94,
              T_NONSENSE_FROM_94_95,T_NONSENSE_FROM_95_96,
              T_NONSENSE_FROM_96_97,T_NONSENSE_FROM_97_98,
              T_NONSENSE_FROM_98_99,T_NONSENSE_FROM_99_100
        version=2.50-cvs
X-Spam-Level:

URL: http://www.newsisfree.com/click/-3,8708820,1440/
Date: Not supplied

Military communications could be jammed or intercepted and satellites thrown
off course or destroyed, a new US study warns
```

Read in Email Messages

In [5]:
```
def get_cwd():
    active_dir = getcwd()
    return active_dir

def main():
```

```
get_cwd()

directories = [
        'easy_ham',
        'easy_ham_2',
        'hard_ham',
        'spam',
        'spam_2'
        ]

res_frame = pd.DataFrame()

emails = []

bodies = []

for d in directories:
    mypath = getcwd() + '/data/' + d + '/'
    onlyfiles = [f for f in listdir(mypath) if isfile(join(mypath, f))]

    try:
        onlyfiles.remove('.DS_Store')
    except:
        pass

    for file in onlyfiles:
        with open(mypath + file, encoding='latin1') as f:
            lines = f.readlines()
            f.close()

        with open(mypath + file, encoding='latin1') as f:
            body = f.read()
            f.close()

        msg = email.message_from_string(str(body))
        tmpStr = ''

        if msg.is_multipart():
            for payload in msg.get_payload():
                tmpStr = ' '.join(str(payload.get_payload()))
            bodies.append(tmpStr)
        else:
            bodies.append(str(msg.get_payload()))

        in_reply_count = 0
        sub_line_all_caps = 0
        attachments = 0
        subject_line = []
        n_lines = 0
        blank_lines = []

        for line in lines:

            n_lines += 1
            if "Subject: Re: " in line:
               in_reply_count += 1
            if "Subject: " in line:
                   s_line = line.strip().replace('Subject: ','')
                   s_line = ''.join(e for e in s_line if e.isalnum())
                   num_upper = sum(1 for c in s_line if c.isupper())
                   ttl_chars = len(s_line)
                   if num_upper == ttl_chars:
                       sub_line_all_caps += 1
                   subject_line.append(s_line)
                if "content-type: multipart" in line.lower():
                   attachments += 1
                if line == "\n":
                   blank_lines.append(n_lines)

            temp_frame = pd.DataFrame({
                        'directory':d,
                        'filename':file,
                        'is_spam':['Y' if 'spam' in d else 'N'],
                        'in_reply': ['Y' if in_reply_count > 0 else 'N'],
                        'subj_caps': ['Y' if sub_line_all_caps > 0 else 'N']
                        'attachments': ['Y' if attachments > 0 else 'N'],
                        'body_lines': [0 if len(blank_lines) == 0 else min(b
                        }, index=[0])

            res_frame = res_frame.append(temp_frame, ignore_index=True)

            # Append body of email to collection
            text = ' '.join(lines)
            emails.append(text)

    # Add emails
    return res_frame, emails, bodies

df, emails, bodies = main()
```

Look at the Results

In [6]:
```
df.info()
df.sample(5)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9353 entries, 0 to 9352
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   directory    9353 non-null   object
 1   filename     9353 non-null   object
 2   is_spam      9353 non-null   object
 3   in_reply     9353 non-null   object
 4   subj_caps    9353 non-null   object
 5   attachments  9353 non-null   object
 6   body_lines   9353 non-null   int64 
dtypes: int64(1), object(6)
memory usage: 511.6+ KB
```

Out[6]:

	directory	filename	is_spam	in_reply	subj_caps
9258	spam_2	01226.4aaf4e328bd55191a1c46bc374069048	Y	N	N
7542	spam	00103.2eef38789b4ecce796e7e8dbe718e3d2	Y	N	N
816	easy_ham	02329.26c7e9813ea8b29dce213bc8275e2904	N	N	N
8707	spam_2	00260.49cb520f5d726da6f1ec32d0e4d2e38f	Y	N	N
4211	easy_ham	01751.bff303bb4466a91b0f88491b207e8ed8	N	N	N

In [7]:

```
df = df.replace(['Y','N'],[1,0])
df.sample(5)
```

Out[7]:

	directory	filename	is_spam	in_reply	subj_caps
8662	spam_2	00485.94b2cb3aa454e6f6701c42cb1fd35ffe	1	0	0
498	easy_ham	01290.41e79a15cd074594f220dfaed53d51aa	0	1	0
8535	spam_2	00336.b937e6ad1deae309e248580a6fec85d8	1	0	0
2518	easy_ham	00687.a044e978152b1411bd3ea6ddc0f537b2	0	1	0
9140	spam_2	00410.fb7b31cdd9d053f8b446da7ce89383fa	1	0	0

Clean up Text

Perform Cleanup Using `nltk`

In [8]:

```
stop_words = nltk.corpus.stopwords.words('english')

def normalize_document(doc):
    # lowercase and remove special characters to form a normalized document
    doc = re.sub(r'[^a-zA-Z0-9\s]', ' ', doc, re.I|re.A)
    doc = doc.lower()
    doc = doc.strip()

    # tokenize document
    tokens = nltk.word_tokenize(doc)

    # filter out stop words
    filtered_tokens = [token for token in tokens if token not in stop_words]

    # Remove numbers
    filtered_tokens = [token for token in filtered_tokens if not token.isdig

    # Remove short tokens
    filtered_tokens = [token for token in filtered_tokens if len(token) > 2]

    # re-create a normalized document
    doc = ' '.join(filtered_tokens)
    return doc

normalize_text = np.vectorize(normalize_document)
norm_text = normalize_text(emails)

print(type(norm_text),len(norm_text))
```

```
<class 'numpy.ndarray'> 9353
```

Put the Cleaned Text into a Dataframe

In [9]:

```
norm_text_2 = pd.DataFrame(norm_text)
directory = pd.DataFrame(df['directory'])
combined_df = pd.concat([norm_text_2,directory],axis=1)
combined_df.columns = ['text','directory']
combined_df['text'] = combined_df['text'].apply(lambda x: x.split())
combined_df
```

Out[9]:

	text	directory
0	[rssfeeds, jmason, org, mon, sep, return, path...	easy_ham
1	[fork, admin, xent, com, tue, sep, return, pat...	easy_ham
2	[fork, admin, xent, com, tue, sep, return, pat...	easy_ham
3	[rpm, list, admin, freshrpms, net, mon, sep, r...	easy_ham
4	[secprog, return, jmason, org, securityfocus, ...	easy_ham
...	...	...
9348	[687ifsuy, bol, com, tue, aug, return, path, 6...	spam_2
9349	[mraimecoilcipc, msn, com, mon, jul, return, p...	spam_2
9350	[fork, admin, xent, com, thu, aug, return, pat...	spam_2
9351	[niddeel, hotmail, com, tue, aug, return, path...	spam_2
9352	[received, actioncouriers, com, linux, midrang...	spam_2

9353 rows × 2 columns

Create a Tf-IDF Matrix

In [10]:

```
tf = TfidfVectorizer(ngram_range=(1,3), min_df=5, max_df=.8, stop_words=stop
tf_matrix = tf.fit_transform(norm_text)

print(tf_matrix.shape)
```

```
(9353, 166742)
```

Perform Topic Modeling

In [11]:

```
os.environ['MALLET_HOME'] = '/Users/matt/mallet-2.0.8'
path_to_mallet = '/Users/matt/mallet-2.0.8/bin/mallet'

# Define training data
combined_df['text_2'] = combined_df['text'].apply(lambda x: ' '.join(x))
training_data = [lmw.process_string(t) for t in combined_df['text_2']]
training_data = [d for d in training_data if d.strip()]

# Number of topics that user specifies
num_topics = 15

# Define output directory
output_directory_path = '/Users/matt/Documents/GitHub/7333-qtw/Case Study 3'

topic_keys, topic_distributions = lmw.quick_train_topic_model(path_to_mallet
                                                              output_directo
                                                              num_topics,
                                                              training_data)
```

```
Importing data...
Complete
Training topic model...
```

```
Mallet LDA: 15 topics, 4 topic bits, 1111 topic mask
Data loaded.
max tokens: 19655
total tokens: 4823074
<10> LL/token: -7.65818
<20> LL/token: -7.28169
<30> LL/token: -7.18082
<40> LL/token: -7.13762

0       0.33333 com net received localhost jul netnoteinc esmtp content aug
zzzz http mon smtp mail subject labs pro html version webnote
1       0.33333 font size color face nbsp arial align verdana NUM center sty
le width div NUMe sans span serif helvetica href table
2       0.33333 org localhost linux received ilug com esmtp dogma slashnull
sep jmason lugh aug ist oct tuatha date rssfeeds taint tue
3       0.33333 com xent fork NUM list localhost received freshrpms net rpm
org mailto admin esmtp http subject sep zzzlist request postfix
4       0.33333 NUM width http src img height gif www font border cnet href
bgcolor table com/b online clickthru com/click size alt
5       0.33333 http www com html radio NUM net org news weblogs com/NUM php
theregister normal blogspot article htm small weblog mediaunspun
6       0.33333 NUM net sourceforge spamassassin example list talk users raz
or lists received localhost usw aug listNUM esmtp com org subject admin
7       0.33333 NUM net lists list linux securityfocus found clean file use
system kernel files content package secprog red iNUM hat dvd
8       0.33333 NUM blockquote grants html government fool grant asp trading
improvement guide NUMarial image com/m step programs aging statements write
NUM/NUM
9       0.33333 com org received yahoo localhost NUM perl aug use sep list h
ttp yahoogroups message zzzzteana subject taint unsubscribe zzzz return
10      0.33333 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int subject
11      0.33333 NUMd NUMdNUM NUM width font size table align com http height
type content name center www ffffff input bgcolor border
12      0.33333 world said states would security also years technology unite
d government new people xent company companies first one fork many law
13      0.33333 http www href nbsp lockergnome com html class target NUM tab
le web blank type title border text name url input
14      0.33333 com one mail people get free money email make time business
received information would like want internet address send new

<50> LL/token: -7.11294
<60> LL/token: -7.09878
<70> LL/token: -7.0885
<80> LL/token: -7.08013
<90> LL/token: -7.0727

0       0.33333 com net received localhost jul netnoteinc esmtp content http
zzzz aug mon mail smtp subject labs html webnote pro click
1       0.33333 font size color face nbsp arial align NUM verdana center sty
le div width NUMe span sans helvetica serif href option
2       0.33333 org localhost linux received ilug com esmtp dogma slashnull
jmason lugh sep aug ist oct tuatha date rssfeeds tue taint
3       0.33333 com xent fork NUM list localhost received org freshrpms rpm
```

```
mailto admin esmtp subject sep http request postfix net version
4       0.33333 NUM width http src height img gif font www border cnet href
bgcolor table com/b online size clickthru com/click alt
5       0.33333 http www com html NUM radio org net weblogs news com/NUM php
theregister normal blogspot mediaunspun htm footer hover weblog
6       0.33333 NUM net sourceforge spamassassin example list talk users raz
or lists received localhost usw aug com listNUM esmtp org subject admin
7       0.33333 net NUM zzzlist list egwn lists matthias linux found clean h
ttp securityfocus reply apt mailing content date kernel version system
8       0.33333 NUM blockquote dot grants unitedmedia html fool com/comics/d
ilbert/daily asp target government grant NUM/NUM com/m NUMarial end NUMcente
r strip image write
9       0.33333 com org received yahoo localhost NUM perl use aug http list
sep yahoogroups zzzzteana message subject zzzz unsubscribe taint iiu
10      0.33333 NUM com exmh org redhat taint users spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int mailto
11      0.33333 NUMd NUMdNUM NUM width font size table http align com height
type name www content center bgcolor ffffff border color
12      0.33333 said world new states technology would people security also
united years government first companies company one could law many president
13      0.33333 http www href nbsp lockergnome com html NUM class table targ
et blank type web title border text name input value
14      0.33333 one com get people mail money email free make time business
would received information org like internet want send address

<100> LL/token: -7.06789
<110> LL/token: -7.06373
<120> LL/token: -7.05659
<130> LL/token: -7.05047
<140> LL/token: -7.04699

0       0.33333 com net received localhost jul netnoteinc esmtp content http
zzzz aug mon mail smtp subject html labs org webnote date
1       0.33333 font size face color nbsp NUM arial align verdana center sty
le div NUMe width span sans helvetica serif option aNUM
2       0.33333 org localhost linux received ilug com esmtp dogma slashnull
jmason lugh sep oct aug ist tuatha date rssfeeds jalapeno tue
3       0.33333 com xent fork NUM localhost list org received mailto sep adm
in esmtp subject request http postfix taint subscribe unsubscribe spamassass
in
4       0.33333 NUM width http src img height gif font www border cnet href
bgcolor table com/b online size clickthru com/click alt
5       0.33333 http www com html radio NUM net org weblogs news php com/NUM
theregister normal blogspot mediaunspun htm hover footer weblog
6       0.33333 NUM net sourceforge spamassassin example list talk users raz
or lists received localhost usw aug com listNUM esmtp org subject mailto
7       0.33333 net NUM freshrpms rpm list zzzlist egwn lists http received
esmtp net/mailman/listinfo/rpm oct matthias subject admin version found clea
n content
8       0.33333 NUM dot grants ummailNUM unitedmedia html NUM/click fool tar
get com/comics/dilbert/daily asp NUM/NUM end comic com/m content NUMarial sc
ript write strip
9       0.33333 com org received yahoo localhost NUM perl use aug http sep l
ist yahoogroups zzzzteana message zzzz subject unsubscribe date iiu
10      0.33333 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int mailto
11      0.33333 NUMd NUMdNUM NUM width font size http table height align com
```

```
www type content name center border bgcolor ffffff color
12      0.33333 world said new security states also would technology united
people years government one company first companies xml time law president
13      0.33333 http www href nbsp lockergnome com NUM html class table targ
et blank type border web title text name value input
14      0.33333 one get people mail money free email make com time business
would information like send received want internet address order

<150> LL/token: -7.04505
<160> LL/token: -7.04234
<170> LL/token: -7.03852
<180> LL/token: -7.0351
<190> LL/token: -7.03394

0       0.33333 com net received localhost jul netnoteinc esmtp content org
zzzz http aug mail mon smtp subject html labs date webnote
1       0.33333 font size color face NUM nbsp arial align verdana style cent
er div NUMe span sans serif width helvetica option aNUM
2       0.33333 org localhost linux ilug received com esmtp slashnull dogma
lugh jmason oct aug tuatha ist sep date rssfeeds tue http
3       0.33333 com xent fork NUM localhost list received org mailto sep adm
in esmtp subject http postfix request taint subscribe spamassassin unsubscri
be
4       0.33333 NUM width http height src img gif font www border cnet href
bgcolor com/b table online size clickthru com/click zdnet
5       0.33333 http www com html NUM radio net org weblogs news php com/NUM
theregister normal blogspot mediaunspun bgcolor htm small weblog
6       0.33333 NUM net sourceforge spamassassin example list talk users raz
or lists received localhost usw aug com esmtp listNUM org subject mailto
7       0.33333 NUM net freshrpms rpm list zzzlist egwn localhost received l
ists http esmtp subject admin oct net/mailman/listinfo/rpm version matthias
org request
8       0.33333 NUM comics dot ummailNUM unitedmedia html NUM/click target f
ool com/comics/dilbert/daily asp end content script daily com/m comic NUMcen
ter NUMverdana strip
9       0.33333 com org received yahoo NUM localhost perl use aug http list
sep yahoogroups zzzzteana unsubscribe zzzz subject message iiu date
10      0.33333 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int mailto
11      0.33333 NUMd NUM NUMdNUM width font http height table size www align
com border src img bgcolor type gif name center
12      0.33333 NUM said world new security also states would united technol
ogy people one years government company companies first xml time may
13      0.33333 http www href nbsp lockergnome com NUM html table class bord
er target blank type web title text name input value
14      0.33333 one get people mail money email make com free time business
would like information send received want address order please

<200> LL/token: -7.02983
[beta: 0.0178]
<210> LL/token: -7.01022
[beta: 0.01839]
<220> LL/token: -6.97834
[beta: 0.01856]
<230> LL/token: -6.95789
[beta: 0.01871]
<240> LL/token: -6.94395
```

```
0       0.09972 com net received localhost jul esmtp org netnoteinc content
http zzzz aug mail mon sep subject smtp html message date
1       0.03512 font NUM size face color nbsp arial align verdana style cent
er div NUMe span width helvetica sans serif option aNUM
2       0.11612 org localhost linux ilug received com esmtp slashnull dogma
lugh jmason oct aug tuatha ist http rssfeeds date sep tue
3       0.11493 com xent fork NUM localhost list received org sep mailto adm
in esmtp http subject postfix request taint spamassassin net version
4       0.01954 NUM width http height src img gif font www border cnet href
com/b bgcolor table online clickthru com/click size zdnet
5       0.0132  http www com html NUM radio net org weblogs php news com/NUM
theregister blogspot normal mediaunspun href bgcolor hover footer
6       0.05545 NUM net sourceforge spamassassin example list talk users raz
or lists received localhost usw com aug listNUM esmtp org subject admin
7       0.02977 NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject admin org oct version net/mailman/listinfo/rp
m request
8       0.03274 NUM comics dot com/comics/dilbert/images/clear blank ummailN
UM unitedmedia dilbert NUM/click target html fool com/comics/dilbert/daily a
sp end daily content script com/m NUMverdana
9       0.05426 com org received NUM yahoo localhost perl http use yahoogrou
ps aug sep list zzzzteana iiu subject message zzzz unsubscribe date
10      0.03225 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.02043 NUMd NUM NUMdNUM width font http size height table www align
com src border img bgcolor gif face color type
12      0.10907 NUM said world new security states also united technology wo
uld people government company one companies first years xml could president
13      0.04139 http www href nbsp lockergnome com NUM html table border cla
ss target blank type title web size text name input
14      0.22009 one com get mail people email money make time free business
would like send received want address order information net

[beta: 0.01883]
<250> LL/token: -6.93681
[beta: 0.01892]
<260> LL/token: -6.92967
[beta: 0.01899]
<270> LL/token: -6.92522
[beta: 0.01903]
<280> LL/token: -6.92062
[beta: 0.01908]
<290> LL/token: -6.9181

0       0.0573  com net received localhost jul org esmtp netnoteinc content
http zzzz aug mon mail sep subject smtp html date message
1       0.02391 font size NUM color face nbsp arial align verdana style cent
er div NUMe span width serif sans helvetica option aNUM
2       0.06699 org localhost linux ilug received com esmtp dogma lugh slash
null jmason oct NUM aug tuatha ist http rssfeeds date sep
3       0.05673 com xent fork NUM localhost list org received sep mailto adm
in esmtp http subject postfix request taint spamassassin net version
4       0.013   NUM width http height src img gif font www border cnet href
com/b bgcolor table online clickthru com/click size zdnet
5       0.00716 http www com html NUM radio net org weblogs php news theregi
ster com/NUM normal blogspot href mediaunspun bgcolor hover footer
```

```
6       0.03173 NUM net sourceforge spamassassin example list talk users raz
or lists received localhost usw com aug listNUM esmtp org subject admin
7       0.01823 NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject org admin com oct version request
8       0.01463 NUM comics dot com/comics/dilbert/images/clear ummailNUM uni
tedmedia blank dilbert NUM/click target html com/comics/dilbert/daily fool a
sp content com/m daily NUMverdana NUMcenter end
9       0.02997 com org received yahoo NUM localhost perl http use yahoogrou
ps sep list aug zzzzteana iiu message subject unsubscribe zzzz grp
10      0.01471 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.01472 NUMd NUM NUMdNUM width font http size height table www align
com border src img face bgcolor color gif center
12      0.06743 world said new security states also technology united govern
ment people NUM company would companies one first years xml time president
13      0.02105 http www href nbsp lockergnome com NUM font html table borde
r class size target blank title type color web text
14      0.13852 NUM one get mail people email com money make time would free
business like received send want address order information

[beta: 0.01914]
<300> LL/token: -6.91555
[beta: 0.01915]
<310> LL/token: -6.91262
[beta: 0.01921]
<320> LL/token: -6.91077
[beta: 0.01923]
<330> LL/token: -6.90894
[beta: 0.01925]
<340> LL/token: -6.90833

0       0.05058 com net received localhost org jul esmtp netnoteinc content
http zzzz aug mail mon subject sep smtp html message date
1       0.02104 font NUM size color face nbsp arial align verdana style cent
er div NUMe span width serif sans option helvetica aNUM
2       0.0579  org localhost linux ilug received com esmtp NUM lugh slashnu
ll dogma jmason oct aug tuatha ist http rssfeeds sep date
3       0.04973 com xent fork NUM localhost list org received sep mailto adm
in esmtp subject http postfix request taint spamassassin net version
4       0.01212 NUM width http height src img gif font www border cnet href
com/b bgcolor table online clickthru com/click size zdnet
5       0.00643 http www com html NUM radio net weblogs org php news theregi
ster com/NUM blogspot normal mediaunspun table bgcolor hover href
6       0.02812 NUM net sourceforge spamassassin example list talk razor use
rs lists received localhost usw com aug listNUM esmtp org subject admin
7       0.01664 NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject org admin com oct version request
8       0.00802 NUM comics dot com/comics/dilbert/images/clear ummailNUM uni
tedmedia blank dilbert NUM/click target html com/comics/dilbert/daily fool a
sp alt content daily end com/m NUMverdana
9       0.02469 com org received yahoo NUM localhost perl use http yahoogrou
ps sep aug list zzzzteana iiu subject message unsubscribe zzzz grp
10      0.01194 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.01345 NUMd NUM NUMdNUM width font http size height table www align
border com img src face bgcolor color center type
12      0.05623 world said new security states also government technology un
```

```
ited company people would one NUM first companies years xml president time
13      0.01536 http www href nbsp lockergnome com NUM font html table borde
r class size target blank title type color width web
14      0.11916 NUM one get people mail email money com make time free would
business like send want order received address list

[beta: 0.01924]
<350> LL/token: -6.90683
[beta: 0.01925]
<360> LL/token: -6.90515
[beta: 0.01923]
<370> LL/token: -6.90546
[beta: 0.01926]
<380> LL/token: -6.90474
[beta: 0.01928]
<390> LL/token: -6.90373

0       0.04935 com net received localhost org jul esmtp netnoteinc content
http zzzz aug mail mon subject sep smtp html date message
1       0.0202  font NUM size color face nbsp arial align verdana style cent
er div NUMe span width serif sans option helvetica aNUM
2       0.05325 org localhost linux ilug received com esmtp NUM lugh dogma s
lashnull jmason oct aug tuatha http ist rssfeeds date sep
3       0.04729 com xent fork NUM localhost list org received sep mailto adm
in esmtp subject http postfix request taint spamassassin net version
4       0.01098 NUM width http height src img gif font www border cnet href
com/b bgcolor table online clickthru com/click size zdnet
5       0.00569 http www com html NUM radio net org weblogs php news theregi
ster com/NUM normal blogspot bgcolor mediaunspun table hover footer
6       0.02575 NUM net sourceforge spamassassin example list talk razor use
rs lists received localhost usw com aug listNUM esmtp org subject admin
7       0.01597 NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject org admin com oct version request
8       0.00674 NUM comics dot com/comics/dilbert/images/clear ummailNUM bla
nk unitedmedia dilbert NUM/click target html com/comics/dilbert/daily asp fo
ol alt content daily com/m NUMverdana NUMcenter
9       0.02293 com org received yahoo NUM localhost perl http use yahoogrou
ps sep aug list zzzzteana iiu message subject zzzz unsubscribe grp
10      0.01092 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.01278 NUMd NUM NUMdNUM width font http size height table align www
com src border img face color bgcolor center type
12      0.0529  said world security new also states government company NUM t
echnology united people would companies one xml years first president may
13      0.01373 http www href nbsp lockergnome com NUM font html table borde
r class size target blank title type color width web
14      0.11228 NUM one get people mail email make money com time would free
like send business order received address want work

[beta: 0.01927]
<400> LL/token: -6.90389
[beta: 0.01928]
<410> LL/token: -6.90261
[beta: 0.01927]
<420> LL/token: -6.90274
[beta: 0.01928]
<430> LL/token: -6.90177
```

```
[beta: 0.01926]
<440> LL/token: -6.90165

0       0.04857 com net received localhost org jul esmtp netnoteinc content
http zzzz aug mail mon sep subject smtp html date message
1       0.01996 font NUM size color face nbsp arial align verdana style cent
er div NUMe span width serif sans option helvetica aNUM
2       0.05367 org localhost linux ilug received com esmtp NUM lugh slashnu
ll dogma jmason oct aug tuatha http ist rssfeeds sep date
3       0.04611 com xent fork NUM localhost list org received sep mailto adm
in esmtp http subject postfix request taint spamassassin net version
4       0.01088 NUM width http height src img gif font www border cnet href
com/b bgcolor table online clickthru com/click size zdnet
5       0.00551 http www com html NUM radio net org weblogs php news theregi
ster com/NUM blogspot normal bgcolor mediaunspun hover footer table
6       0.02557 NUM net sourceforge spamassassin example list talk razor use
rs lists received localhost usw com aug listNUM esmtp org subject admin
7       0.01535 NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject org admin com oct version request
8       0.00672 NUM comics dot com/comics/dilbert/images/clear ummailNUM uni
tedmedia blank dilbert NUM/click target html com/comics/dilbert/daily conten
t alt asp daily com/m NUMverdana fool NUMcenter
9       0.02227 com org received yahoo NUM localhost perl http use yahoogrou
ps sep aug list zzzzteana iiu zzzz subject message unsubscribe grp
10      0.01084 NUM com exmh redhat org users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.01279 NUMd NUM NUMdNUM font width http size height table www align
com src border img face color bgcolor center type
12      0.04991 world said new NUM security company also states government u
nited technology one people would companies xml years first president may
13      0.0121  http www href nbsp lockergnome com NUM font html table borde
r class size target blank title type color width web
14      0.1125  NUM one get people mail email make com time money would free
like send want order business net address report

[beta: 0.01926]
<450> LL/token: -6.90161
[beta: 0.01927]
<460> LL/token: -6.90154
[beta: 0.01928]
<470> LL/token: -6.90033
[beta: 0.01928]
<480> LL/token: -6.90052
[beta: 0.01928]
<490> LL/token: -6.89995

0       0.04824 com net received localhost org jul esmtp netnoteinc content
http zzzz aug mail mon sep subject smtp html date message
1       0.0198  font NUM size color face nbsp arial align verdana style cent
er div NUMe span width serif sans option helvetica aNUM
2       0.0522  org localhost linux ilug received com esmtp NUM lugh slashnu
ll dogma jmason oct aug tuatha http ist rssfeeds date sep
3       0.04719 com xent fork NUM localhost list org received sep mailto adm
in esmtp http subject postfix request taint spamassassin net version
4       0.01092 NUM width http height src img gif font www border cnet href
com/b bgcolor table online clickthru com/click size zdnet
5       0.00542 http www com html NUM radio net weblogs org php news theregi
```

```
ster com/NUM normal blogspot bgcolor mediaunspun table href hover
6       0.02524 NUM net sourceforge spamassassin example list talk razor use
rs lists received localhost usw com aug listNUM esmtp org subject admin
7       0.01532 NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject org admin com oct version request
8       0.0067  NUM comics dot com/comics/dilbert/images/clear ummailNUM uni
tedmedia blank dilbert NUM/click target html com/comics/dilbert/daily conten
t alt com/m NUMverdana daily NUMcenter end comic
9       0.02172 com org received NUM yahoo localhost perl http use yahoogrou
ps sep list aug zzzzteana iiu message unsubscribe subject zzzz grp
10      0.01065 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.01276 NUMd NUM NUMdNUM font width http size height table align www
com border src img face color bgcolor center type
12      0.04899 world said new security NUM states government company also u
nited technology would one people companies first xml years president time
13      0.0118  http www href nbsp lockergnome com NUM font html table borde
r size class target blank title type color width web
14      0.11086 NUM one get people mail email time make would com money like
free send received order net business want report

[beta: 0.01927]
<500> LL/token: -6.89945
[beta: 0.01928]
<510> LL/token: -6.89878
[beta: 0.0193]
<520> LL/token: -6.89913
[beta: 0.01929]
<530> LL/token: -6.89878
[beta: 0.01929]
<540> LL/token: -6.89942

0       0.04836 com net received localhost org jul esmtp netnoteinc http con
tent zzzz aug mail mon sep subject smtp message date html
1       0.01958 font NUM size face color nbsp arial align verdana style cent
er div NUMe span width serif sans option helvetica aNUM
2       0.05099 org localhost linux ilug received com esmtp NUM lugh dogma s
lashnull jmason oct aug tuatha http ist rssfeeds date sep
3       0.04532 com xent fork NUM localhost list org received sep mailto adm
in esmtp http subject postfix request taint spamassassin net version
4       0.01032 NUM width http height src img gif font www border cnet href
com/b bgcolor table online clickthru com/click size zdnet
5       0.005   http www com html radio NUM net org weblogs php news theregi
ster com/NUM normal blogspot table bgcolor mediaunspun hover footer
6       0.02527 NUM net sourceforge spamassassin example list talk razor use
rs lists received localhost usw com aug listNUM esmtp org subject admin
7       0.01544 NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject org admin com oct version request
8       0.00646 NUM comics dot com/comics/dilbert/images/clear ummailNUM uni
tedmedia blank dilbert NUM/click target html com/comics/dilbert/daily conten
t alt daily com/m NUMverdana end NUMcenter strip
9       0.02141 com org received yahoo NUM localhost perl http use yahoogrou
ps sep aug zzzzteana list iiu message subject unsubscribe zzzz grp
10      0.01093 NUM com exmh redhat org users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.01277 NUMd NUM NUMdNUM width font http size height table align www
com border src img face color bgcolor center gif
```

```
12      0.04791 world said new security states also company government unite
d NUM technology one people companies would first xml years president may
13      0.01068 http www href nbsp lockergnome com NUM font html table borde
r class size target blank title color width type web
14      0.10943 NUM one get people mail email time make would com money like
free send received order want net internet information

[beta: 0.01931]
<550> LL/token: -6.89875
[beta: 0.01931]
<560> LL/token: -6.89896
[beta: 0.01931]
<570> LL/token: -6.89778
[beta: 0.01931]
<580> LL/token: -6.89773
[beta: 0.01928]
<590> LL/token: -6.89822

0       0.04718 com net received localhost org jul esmtp netnoteinc content
http zzzz aug mail mon subject sep smtp date message html
1       0.01933 font NUM size face color nbsp arial align verdana style cent
er div NUMe span width serif sans option helvetica aNUM
2       0.05022 org localhost linux ilug received com esmtp NUM lugh slashnu
ll dogma jmason oct tuatha aug http ist rssfeeds date sep
3       0.04604 com xent fork NUM localhost list org received sep mailto adm
in esmtp http subject postfix request taint spamassassin net version
4       0.01038 NUM width http height src img gif font www border cnet href
com/b bgcolor table online clickthru com/click size zdnet
5       0.00505 http www com html NUM radio net weblogs org php news theregi
ster com/NUM normal blogspot table bgcolor mediaunspun footer hover
6       0.02464 NUM net sourceforge spamassassin example list talk razor use
rs lists received localhost usw com aug listNUM esmtp org subject admin
7       0.01526 NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject org admin com oct version request
8       0.00575 NUM comics dot com/comics/dilbert/images/clear ummailNUM uni
tedmedia blank dilbert NUM/click target html com/comics/dilbert/daily conten
t daily NUMverdana end NUMcenter comic strip begin
9       0.02151 com org received yahoo NUM localhost perl http use yahoogrou
ps sep aug list zzzzteana iiu message subject zzzz unsubscribe grp
10      0.01053 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.01257 NUMd NUM NUMdNUM font width http size height table www align
com border src img color face bgcolor center gif
12      0.04656 said world states security new government company also unite
d technology NUM one would people companies years xml first president americ
an
13      0.01036 http www href nbsp lockergnome com font NUM html table borde
r class size target blank title width color type web
14      0.10732 NUM one get people email make mail time would money like com
free send list order new also want use

[beta: 0.0193]
<600> LL/token: -6.89822
[beta: 0.01931]
<610> LL/token: -6.89754
[beta: 0.01928]
<620> LL/token: -6.89779
```

```
[beta: 0.01929]
<630> LL/token: -6.89789
[beta: 0.01929]
<640> LL/token: -6.8976

0       0.04769 com net received localhost org jul esmtp netnoteinc content
http zzzz aug mail mon sep subject smtp message date html
1       0.01956 font NUM size color face nbsp arial align verdana style cent
er div NUMe span width serif sans option helvetica aNUM
2       0.05059 org localhost linux ilug received com esmtp NUM lugh slashnu
ll dogma jmason oct aug tuatha http ist rssfeeds date sep
3       0.04478 com xent fork NUM localhost list org received sep mailto adm
in esmtp http subject postfix request taint spamassassin net version
4       0.01043 NUM width http height src img gif font www border cnet href
com/b bgcolor table online clickthru com/click size zdnet
5       0.00493 http www com html NUM radio net weblogs org php news theregi
ster com/NUM normal blogspot bgcolor mediaunspun href table hover
6       0.02467 NUM net sourceforge spamassassin example list talk razor use
rs lists received localhost usw com aug listNUM esmtp org subject admin
7       0.01515 NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject org admin com oct version request
8       0.00542 NUM comics dot com/comics/dilbert/images/clear ummailNUM uni
tedmedia blank dilbert NUM/click target html com/comics/dilbert/daily conten
t daily NUMverdana NUMcenter end comic strip begin
9       0.02114 com org received yahoo NUM localhost perl http use yahoogrou
ps sep aug zzzzteana list iiu message zzzz subject unsubscribe grp
10      0.01032 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.01245 NUMd NUM NUMdNUM font width http size height table align www
com border src img face color bgcolor center gif
12      0.04549 said world NUM new states security government company united
also technology one people would companies first xml years president informa
tion
13      0.00948 http www href nbsp lockergnome com font NUM html table borde
r class size target blank title width color type web
14      0.10736 NUM one get people email mail would time make like money com
free send also order new want list work

[beta: 0.0193]
<650> LL/token: -6.89744
[beta: 0.01929]
<660> LL/token: -6.89703
[beta: 0.01929]
<670> LL/token: -6.89725
[beta: 0.01928]
<680> LL/token: -6.89729
[beta: 0.01928]
<690> LL/token: -6.89719

0       0.04853 com net received localhost org jul esmtp netnoteinc content
zzzz http aug mail mon subject sep smtp date message html
1       0.0191  font NUM size color face nbsp arial align verdana style cent
er div NUMe span width sans option serif helvetica aNUM
2       0.05013 org localhost linux ilug received com esmtp NUM lugh dogma s
lashnull jmason oct aug tuatha http ist rssfeeds date sep
3       0.04577 com xent fork NUM localhost list org received sep mailto adm
in http esmtp subject postfix request taint spamassassin net version
```

```
4       0.01053 NUM width http height src img gif font www border cnet href 
com/b bgcolor table online clickthru com/click size zdnet
5       0.00489 http www com html NUM radio net org php weblogs news theregi
ster com/NUM table blogspot normal bgcolor mediaunspun footer hover
6       0.02413 NUM net sourceforge spamassassin example list talk razor use
rs lists received localhost usw com aug listNUM esmtp org subject admin
7       0.01459 NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject org admin com oct version request
8       0.00543 NUM comics dot com/comics/dilbert/images/clear ummailNUM uni
tedmedia blank dilbert NUM/click target html com/comics/dilbert/daily conten
t daily NUMverdana NUMcenter end alt comic strip
9       0.02097 com org received yahoo NUM localhost perl http use yahoogrou
ps sep list zzzzteana aug iiu message zzzz subject unsubscribe grp
10      0.01015 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.01238 NUMd NUM NUMdNUM font width http size height table align www 
com border face src img color bgcolor gif center
12      0.04468 world said states new government security NUM also company u
nited technology one people companies would first years xml president inform
ation
13      0.00887 http www href nbsp lockergnome com font NUM html table borde
r class size target blank width title color type name
14      0.10851 NUM one get people email time make mail would like com money 
free send list new also order want work

[beta: 0.01928]
<700> LL/token: -6.89716
[beta: 0.01928]
<710> LL/token: -6.89748
[beta: 0.01929]
<720> LL/token: -6.89698
[beta: 0.01929]
<730> LL/token: -6.89631
[beta: 0.01928]
<740> LL/token: -6.89709

0       0.04829 com net received localhost org jul esmtp netnoteinc content 
http zzzz aug mail mon sep subject smtp html message date
1       0.01953 font NUM size color face nbsp arial align verdana style cent
er div NUMe span width serif sans option aNUM helvetica
2       0.05005 org localhost linux ilug received com esmtp NUM lugh dogma s
lashnull jmason oct aug tuatha http ist rssfeeds sep date
3       0.04519 com xent fork NUM localhost list org received sep mailto adm
in esmtp http subject postfix request taint spamassassin net version
4       0.01045 NUM width http height src img gif font www border cnet href 
com/b bgcolor table online clickthru com/click size zdnet
5       0.00503 http www com html radio NUM net org weblogs php news theregi
ster com/NUM blogspot normal table mediaunspun bgcolor hover footer
6       0.02374 NUM net sourceforge spamassassin example list talk razor use
rs lists received localhost usw com aug listNUM esmtp org subject admin
7       0.01481 NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject org admin com oct version request
8       0.00543 NUM comics dot com/comics/dilbert/images/clear ummailNUM uni
tedmedia blank dilbert NUM/click target html com/comics/dilbert/daily conten
t daily alt NUMverdana NUMcenter end comic strip
9       0.02077 com org received yahoo NUM localhost perl http use yahoogrou
ps sep aug list zzzzteana iiu zzzz message subject unsubscribe grp
```

```
10      0.01014 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.01219 NUMd NUM NUMdNUM font width http size height table www align
com border src img color face bgcolor center gif
12      0.04475 said world states new security government NUM also company u
nited technology one people companies would first xml president years time
13      0.00886 http www href nbsp lockergnome com font NUM html table borde
r class size target blank color width title type name
14      0.10861 NUM one get people email make time would mail like com money
free send want work list order new use

[beta: 0.01929]
<750> LL/token: -6.89696
[beta: 0.01928]
<760> LL/token: -6.89614
[beta: 0.01928]
<770> LL/token: -6.89605
[beta: 0.0193]
<780> LL/token: -6.89622
[beta: 0.01929]
<790> LL/token: -6.89576

0       0.048   com net received localhost org jul esmtp netnoteinc content
http zzzz aug mail mon subject sep smtp message date html
1       0.01892 font NUM size color face nbsp arial align verdana style cent
er div NUMe span width sans serif option helvetica aNUM
2       0.04988 org localhost linux ilug received com esmtp NUM lugh slashnu
ll dogma jmason oct aug tuatha http ist rssfeeds date sep
3       0.04485 com xent fork NUM localhost list org received sep mailto adm
in esmtp http subject postfix request taint spamassassin net version
4       0.0107  NUM width http height src img gif font www border cnet href
com/b bgcolor table online clickthru com/click size zdnet
5       0.00496 http www com html radio NUM net org weblogs php news com/NUM
theregister blogspot normal mediaunspun table href hover footer
6       0.02372 NUM net sourceforge spamassassin example list talk razor use
rs lists received localhost usw com aug listNUM esmtp org subject admin
7       0.01468 NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject org admin com oct version request
8       0.00523 NUM comics dot com/comics/dilbert/images/clear ummailNUM uni
tedmedia blank dilbert NUM/click target html com/comics/dilbert/daily conten
t alt NUMverdana daily NUMcenter end comic begin
9       0.02079 com org received yahoo NUM localhost perl http use yahoogrou
ps sep zzzzteana aug list iiu zzzz message unsubscribe subject grp
10      0.01006 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.01256 NUMd NUM NUMdNUM font width http size height table align www
com border img face src color bgcolor center gif
12      0.04374 said world states new government NUM security company united
also technology one people would companies first xml years president time
13      0.00853 http www href nbsp lockergnome com font NUM table html borde
r class size target blank title width color type text
14      0.10702 NUM one get people make email mail would time like money com
send free new also use want order work

[beta: 0.01928]
<800> LL/token: -6.89596
[beta: 0.01929]
```

```
<810> LL/token: -6.89608
[beta: 0.01927]
<820> LL/token: -6.89599
[beta: 0.0193]
<830> LL/token: -6.89554
[beta: 0.01932]
<840> LL/token: -6.89622

0       0.04815 com net received localhost org jul esmtp netnoteinc content
http zzzz aug mail mon sep subject smtp message date html
1       0.01921 font NUM size color face nbsp arial align verdana style cent
er div NUMe span width sans serif option helvetica aNUM
2       0.04961 org localhost linux ilug received com esmtp NUM lugh slashnu
ll dogma jmason oct aug tuatha http ist rssfeeds date sep
3       0.04476 com xent fork NUM localhost list org received sep mailto adm
in esmtp http subject postfix request taint spamassassin net version
4       0.01066 NUM width http height src img gif font www border cnet href
com/b bgcolor table online clickthru com/click size zdnet
5       0.00486 http www com html NUM radio net org php weblogs news com/NUM
theregister normal blogspot table mediaunspun bgcolor footer hover
6       0.0238  NUM net sourceforge spamassassin example list talk razor use
rs lists received localhost usw com aug listNUM esmtp org subject admin
7       0.01457 NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject org admin com oct version request
8       0.00527 NUM comics dot com/comics/dilbert/images/clear ummailNUM uni
tedmedia blank dilbert NUM/click target html com/comics/dilbert/daily conten
t NUMverdana alt daily NUMcenter script end comic
9       0.02064 com org received yahoo NUM localhost perl http use yahoogrou
ps sep aug zzzzteana list iiu message unsubscribe zzzz subject grp
10      0.0097  NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.01226 NUMd NUM NUMdNUM font width http size height table align www
com border src img face color bgcolor center gif
12      0.04357 said world new states government security also NUM united co
mpany technology one people companies would first xml years president americ
an
13      0.00848 http www href nbsp lockergnome com font NUM html table borde
r class size target blank width title color type name
14      0.10707 NUM one people get email would time make mail like money fre
e send com use list net also order new

[beta: 0.0193]
<850> LL/token: -6.89539
[beta: 0.0193]
<860> LL/token: -6.89611
[beta: 0.01928]
<870> LL/token: -6.89567
[beta: 0.0193]
<880> LL/token: -6.89696
[beta: 0.0193]
<890> LL/token: -6.89607

0       0.04787 com net received localhost org jul esmtp netnoteinc content
http zzzz aug mail mon subject sep smtp date html message
1       0.01931 font NUM size face color nbsp arial align verdana style cent
er div NUMe span width serif sans option helvetica aNUM
2       0.04976 org localhost linux ilug received com esmtp NUM lugh slashnu
```

```
ll dogma jmason oct aug tuatha http ist rssfeeds date sep
3       0.04444 com xent fork NUM localhost list org received sep mailto adm
in esmtp http subject postfix request taint spamassassin net version
4       0.01084 NUM width http height src img gif font www border cnet href
com/b bgcolor table online clickthru com/click size zdnet
5       0.00513 http www com html NUM radio net org weblogs php news theregi
ster com/NUM blogspot normal bgcolor table mediaunspun href footer
6       0.02377 NUM net sourceforge spamassassin example list talk razor use
rs lists received localhost usw com aug listNUM esmtp org subject admin
7       0.0147  NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject org admin com oct version request
8       0.00512 NUM comics dot com/comics/dilbert/images/clear ummailNUM uni
tedmedia blank dilbert NUM/click target html com/comics/dilbert/daily conten
t daily alt NUMverdana NUMcenter end strip begin
9       0.02073 com org received yahoo NUM localhost perl http use yahoogrou
ps sep aug zzzzteana list iiu subject message zzzz unsubscribe grp
10      0.00977 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.0124  NUMd NUM NUMdNUM font width http size height table align www
com border src img face color bgcolor center type
12      0.04382 world said states new government NUM also company united sec
urity one technology people would companies xml first president years americ
an
13      0.00822 http www href nbsp lockergnome com font NUM html table borde
r class size target blank title width color type name
14      0.10648 NUM one get people email would time make mail like money com
free send want software information new work list

[beta: 0.01927]
<900> LL/token: -6.89616
[beta: 0.01929]
<910> LL/token: -6.89619
[beta: 0.01928]
<920> LL/token: -6.8962
[beta: 0.01929]
<930> LL/token: -6.89622
[beta: 0.01929]
<940> LL/token: -6.89541

0       0.04822 com net received localhost org jul esmtp netnoteinc content
http zzzz aug mail mon sep subject smtp html date message
1       0.01917 font NUM size face color nbsp arial align verdana style cent
er div NUMe span width sans serif option helvetica aNUM
2       0.0497  org localhost linux ilug received com esmtp NUM lugh slashnu
ll dogma jmason oct tuatha aug http ist rssfeeds date sep
3       0.04521 com xent fork NUM localhost list org received sep mailto adm
in http esmtp subject postfix request taint spamassassin net version
4       0.01067 NUM width http height src img gif font www border cnet href
com/b bgcolor table online clickthru com/click size zdnet
5       0.00486 http www com html NUM radio net org weblogs php news theregi
ster com/NUM normal blogspot table bgcolor mediaunspun footer hover
6       0.02319 NUM net sourceforge spamassassin example list talk razor use
rs lists received localhost usw com aug listNUM esmtp org subject admin
7       0.0144  NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject org admin com oct version request
8       0.00518 NUM comics dot com/comics/dilbert/images/clear ummailNUM uni
tedmedia blank dilbert NUM/click target html com/comics/dilbert/daily conten
```

```
t daily alt NUMverdana NUMcenter end begin comic
9       0.02035 com org received yahoo NUM localhost perl http use yahoogrou
ps sep zzzzteana aug list iiu zzzz subject unsubscribe message grp
10      0.01007 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.0122  NUMd NUM NUMdNUM font width http size height table align www
com border src img face color bgcolor gif center
12      0.04272 world said states new government also united NUM company sec
urity technology people one would companies first xml president years americ
an
13      0.00816 http www href nbsp lockergnome com font NUM html table borde
r class size target blank color title width type name
14      0.1069  NUM one get people time email would make mail like money com
free send use new list work want order

[beta: 0.01928]
<950> LL/token: -6.89512
[beta: 0.01929]
<960> LL/token: -6.89538
[beta: 0.01928]
<970> LL/token: -6.89522
[beta: 0.0193]
<980> LL/token: -6.89605
[beta: 0.01929]
<990> LL/token: -6.89555

0       0.04845 com net received localhost org jul esmtp netnoteinc content
http zzzz aug mail mon sep subject smtp date NUM message
1       0.01929 font NUM size color face nbsp arial align verdana style cent
er div NUMe span width sans serif option helvetica aNUM
2       0.04964 org localhost linux ilug received com esmtp NUM lugh slashnu
ll dogma jmason oct aug tuatha http ist rssfeeds sep date
3       0.04546 com xent fork NUM localhost list org received sep mailto adm
in http esmtp subject postfix request taint spamassassin net version
4       0.01073 NUM width http height src img gif font www border cnet href
com/b bgcolor table online clickthru com/click size zdnet
5       0.00485 http www com html NUM radio net org weblogs php news theregi
ster com/NUM normal blogspot table bgcolor mediaunspun href hover
6       0.02372 NUM net sourceforge spamassassin example list talk razor use
rs lists received localhost usw com aug listNUM esmtp org subject admin
7       0.01464 NUM net freshrpms rpm list zzzlist egwn localhost received h
ttp esmtp lists mailto subject org admin com oct version request
8       0.0052  NUM comics dot com/comics/dilbert/images/clear ummailNUM uni
tedmedia blank dilbert NUM/click target html com/comics/dilbert/daily conten
t daily alt NUMverdana NUMcenter end comic begin
9       0.02079 com org received yahoo NUM localhost perl http use yahoogrou
ps sep aug zzzzteana list iiu subject unsubscribe zzzz message grp
10      0.01004 NUM com exmh org redhat users taint spamassassin workers lis
tman received example list localhost mxNUM sep aug esmtp int corp
11      0.01239 NUMd NUM NUMdNUM font width http size height table www align
com border src img face color bgcolor center gif
12      0.04208 said world states government new NUM united also security co
mpany technology people one would companies first xml years president americ
an
13      0.00804 http www href nbsp lockergnome com font NUM html table borde
r class size target blank title width color type name
14      0.1069  NUM one get people email would time make mail like money com
```

```
free send use list also new work information

[beta: 0.01928]
<1000> LL/token: -6.89501

Total time: 5 minutes 57 seconds
```

```
Complete
```

In [12]:

```
# Organizing the results into a dictionary
topic_word_probability_dict = lmw.load_topic_word_distributions(output_direc
len(topic_word_probability_dict)

for _topic, _word_probability_dict in topic_word_probability_dict.items():
    print('Topic', _topic)
    for _word, _probability in sorted(_word_probability_dict.items(), key=la
        print(round(_probability, 4), '\t', _word)
    print()
```

```
Topic 0
0.0607	com
0.0288	net
0.0239	received
0.0195	localhost
0.0125	org
0.0115	jul
0.0114	esmtp
0.0109	netnoteinc
0.0101	content
0.0101	http
0.0097	zzzz
0.009	aug
0.0085	mail
0.0076	mon
0.0075	sep

Topic 1
0.1272	font
0.0445	NUM
0.0428	size
0.0337	color
0.0329	face
0.0293	nbsp
0.0208	arial
0.0183	align
0.0148	verdana
0.0147	style
0.0134	center
0.0125	div
0.0116	NUMe
0.0105	span
0.0098	width

Topic 2
0.0491	org
0.039	localhost
0.0382	linux
```

```
0.0265   ilug
0.0253   received
0.0226   com
0.0168   esmtp
0.0164   NUM
0.014    lugh
0.0138   slashnull
0.0137   dogma
0.0127   jmason
0.012    oct
0.0118   aug
0.0117   tuatha

Topic 3
0.0687   com
0.0561   xent
0.0475   fork
0.0305   NUM
0.0233   localhost
0.0225   list
0.0213   org
0.0211   received
0.0152   sep
0.0148   mailto
0.0144   admin
0.0141   http
0.014    esmtp
0.0137   subject
0.0113   postfix

Topic 4
0.0854   NUM
0.0669   width
0.0579   http
0.0399   height
0.0387   src
0.0385   img
0.0366   gif
0.0312   font
0.0268   www
0.0253   border
0.0241   cnet
0.0195   href
0.0187   com/b
0.0184   bgcolor
0.0179   table

Topic 5
0.0833   http
0.0475   www
0.0334   com
0.0148   html
0.0089   NUM
0.0082   radio
0.0066   net
0.0063   org
0.0061   weblogs
```

```
0.0058   php
0.0056   news
0.0051   theregister
0.005    com/NUM
0.0045   normal
0.0044   blogspot

Topic 6
0.0643   NUM
0.0586   net
0.0533   sourceforge
0.0254   spamassassin
0.0233   example
0.0196   list
0.0187   talk
0.0185   razor
0.0182   users
0.0174   lists
0.016    received
0.0144   localhost
0.014    usw
0.0125   com
0.0117   aug

Topic 7
0.0584   NUM
0.0541   net
0.0453   freshrpms
0.043    rpm
0.0391   list
0.0259   zzzlist
0.0172   egwn
0.0159   localhost
0.0142   received
0.0118   http
0.0103   esmtp
0.0097   lists
0.0096   mailto
0.0089   subject
0.0089   org

Topic 8
0.3745   NUM
0.0112   comics
0.0074   dot
0.0062   com/comics/dilbert/images/clear
0.0053   ummailNUM
0.0052   unitedmedia
0.0049   blank
0.0045   dilbert
0.0045   NUM/click
0.0036   target
0.0032   html
0.0031   com/comics/dilbert/daily
0.0026   content
0.0018   daily
0.0018   NUMverdana
```

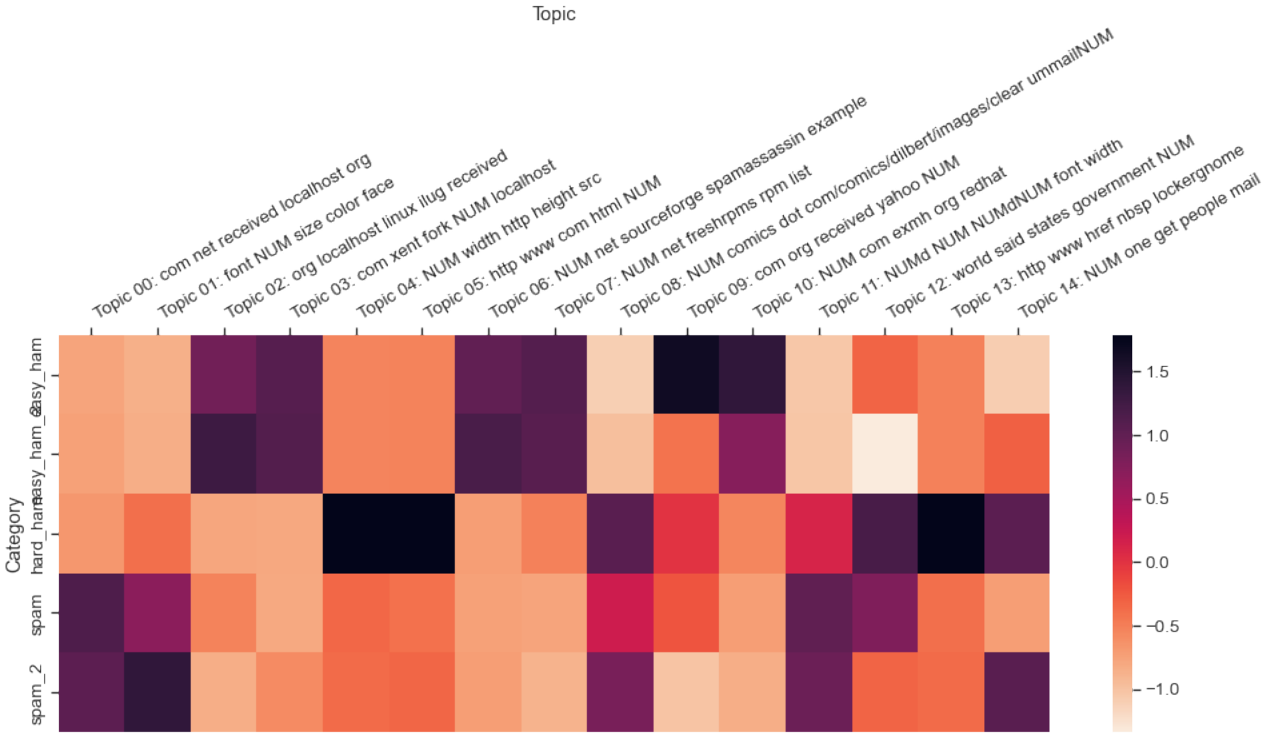
```
Topic 9
0.0449   com
0.028    org
0.0224   received
0.0192   yahoo
0.0184   NUM
0.0156   localhost
0.0142   perl
0.0116   http
0.011    use
0.0109   yahoogroups
0.0108   sep
0.01     aug
0.01     zzzzteana
0.0097   list
0.0081   iiu

Topic 10
0.0984   NUM
0.0553   com
0.0486   exmh
0.0253   org
0.0253   redhat
0.0234   users
0.0221   taint
0.0221   spamassassin
0.0212   workers
0.0211   listman
0.0188   received
0.0178   example
0.0162   list
0.0159   localhost
0.0127   mxNUM

Topic 11
0.2503   NUMd
0.0474   NUM
0.0384   NUMdNUM
0.0306   font
0.03     width
0.0185   http
0.016    size
0.0143   height
0.0136   table
0.012    align
0.0117   www
0.0093   com
0.0092   src
0.0092   border
0.0087   img

Topic 12
0.0046   world
0.0045   said
0.0035   states
0.0035   government
```

```
0.0034   NUM
0.0034   new
0.0032   united
0.0032   also
0.0031   security
0.003    company
0.0029   technology
0.0028   people
0.0027   one
0.0027   would
0.0024   companies

Topic 13
0.0431   http
0.0319   www
0.0313   href
0.0276   nbsp
0.0209   lockergnome
0.0182   com
0.015    font
0.0131   NUM
0.0118   html
0.0116   table
0.011    border
0.0095   class
0.0094   size
0.0089   target
0.0084   blank

Topic 14
0.04     NUM
0.0084   one
0.006    get
0.0059   people
0.005    mail
0.005    would
0.0049   make
0.0049   email
0.0048   time
0.0041   like
0.0038   money
0.0036   com
0.0034   free
0.0033   send
0.0033   use
```

Examine Topics by Category

In [13]:

```
directory = combined_df['directory'].tolist()

target_labels = ["easy_ham",
                 "easy_ham_2",
                 "hard_ham",
                 "spam",
                 "spam_2"]

lmw.plot_categories_by_topics_heatmap(directory,
                                      topic_distributions,
                                      topic_keys,
                                      output_directory_path + '/categories_b
                                      target_labels=target_labels,
                                      dim=(14,8))
```



In [14]:

```
print('Spam directory is related to the following topics : \n', pd.DataFrame
print('\n','Spam_2 directory is related to the following topics : \n', pd.Da
```

```
Spam directory is related to the following topics :
 com,net,received,localhost,org,jul,esmtp,netnoteinc,http,content,zzzz,aug,m
ail,mon,sep,subject,smtp,NUM,message,date        com,xent,fork,NUM,localhost
,list,org,received,sep,mailto,admin,http,esmtp,subject,postfix,request,taint
,spamassassin,net,version        NUM,width,http,height,src,img,gif,font,www,
border,cnet,href,com/b,bgcolor,table,online,clickthru,com/click,size,zdnet

 Spam_2 directory is related to the following topics :
 com,net,received,localhost,org,jul,esmtp,netnoteinc,http,content,zzzz,aug,m
ail,mon,sep,subject,smtp,NUM,message,date
 http,www,com,html,NUM,radio,net,org,weblogs,php,news,theregister,com/NUM,no
rmal,blogspot,table,bgcolor,mediaunspun,hover,footer
```

Clustering

In [15]:

```
NUM_CLUSTERS = 15
km = KMeans(n_clusters=NUM_CLUSTERS, max_iter=10000, n_init=10, random_state
km

df['kmeans_cluster'] = km.labels_

email_clusters = (df[['directory', 'kmeans_cluster']]
                  .sort_values(by=['kmeans_cluster'],
                               ascending=False)
                  .groupby('kmeans_cluster').head(20))  # top 20 movies for
email_clusters = email_clusters.copy(deep=True)

feature_names = tf.get_feature_names()
topn_features = 15
ordered_centroids = km.cluster_centers_.argsort()[:, ::-1]

sample_silhouette_values = silhouette_samples(tf_matrix, km.labels_)

# get key features for each cluster
for cluster_num in range(NUM_CLUSTERS):

    cluster_silhouette_values = sample_silhouette_values[km.labels_ == clust

    key_features = [feature_names[index]
                    for index in ordered_centroids[cluster_num, :topn_fe
    print('CLUSTER #'+str(cluster_num+1), ":", cluster_silhouette_values.mea
    print('Cluster Size', cluster_silhouette_values.shape[0])
    print('Key Features:', key_features)
    print('-'*80)
```

```
CLUSTER #1 : 0.24944832267230824
Cluster Size 218
Key Features: ['yahoo', 'yahoogroups com', 'yahoogroups', 'zzzzteana', 'yaho
o com', 'grp scd', 'grp scd yahoo', 'scd', 'scd yahoo com', 'scd yahoo', 'gr
p', 'zzzzteana yahoogroups', 'zzzzteana yahoogroups com', 'groups', 'fortean
a']
----------------------------------------------------------------------------
----
CLUSTER #2 : 0.09463603457153336
Cluster Size 605
Key Features: ['rssfeeds', 'oct', 'spamassassin taint org', 'spamassassin ta
int', 'taint org', 'taint', 'rssfeeds spamassassin taint', 'rssfeeds spamass
assin', 'spamassassin', 'sep', 'tue oct', 'jmason org', 'jmason', 'path rssf
eeds spamassassin', 'utf url']
----------------------------------------------------------------------------
----
CLUSTER #3 : -0.015095537273865055
Cluster Size 2121
Key Features: ['net', 'perl', 'aug', 'zzzz', 'sep', 'jul', 'use perl', 'netn
oteinc com', 'netnoteinc', 'font', 'mon', 'mail', 'zzzz localhost', 'perl or
g', 'webnote']
----------------------------------------------------------------------------
----
CLUSTER #4 : 0.21142142557381002
Cluster Size 474
Key Features: ['3d', 'font', 'width 3d', 'size 3d', 'width', 'size', 'color
3d', 'td', 'color', 'face 3d', 'face', '20', '3d http', 'font face 3d', 'ali
```

```
gn']
----------------------------------------------------------------------------
----
CLUSTER #5 : 0.35141369159442093
Cluster Size 658
Key Features: ['rpm', 'freshrpms', 'freshrpms net', 'zzzlist', 'rpm zzzlist'
, 'net', 'rpm list', 'egwn', 'list', 'egwn net', 'zzzlist freshrpms', 'zzzli
st freshrpms net', 'rpm zzzlist freshrpms', 'http lists freshrpms', 'lists f
reshrpms']
----------------------------------------------------------------------------
----
CLUSTER #6 : 0.17302184480598956
Cluster Size 367
Key Features: ['sourceforge net', 'sourceforge', 'spamassassin talk', 'talk'
, 'net', 'spamassassin', 'example sourceforge', 'example sourceforge net', '
lists', 'usw', 'lists sourceforge net', 'lists sourceforge', 'example', 'tal
k admin', 'mailto spamassassin']
----------------------------------------------------------------------------
----
CLUSTER #7 : 0.17673802557736054
Cluster Size 1741
Key Features: ['fork', 'xent com', 'xent', 'fork admin xent', 'admin xent',
'admin xent com', 'fork admin', 'mailto fork', 'sep', 'http xent com', 'http
xent', 'list', 'xent com subject', 'fork request xent', 'request xent']
----------------------------------------------------------------------------
----
CLUSTER #8 : 0.2464098316002266
Cluster Size 206
Key Features: ['exmh users', 'exmh', 'users', 'redhat com', 'listman', 'redh
at', 'mx1', 'example com', 'spamassassin taint org', 'spamassassin taint', '
taint org', 'taint', 'spamassassin', 'listman redhat com', 'listman redhat']
----------------------------------------------------------------------------
----
CLUSTER #9 : 0.05947292054464942
Cluster Size 676
Key Features: ['font', 'width', 'td', 'size', 'height', 'nbsp', 'href http',
'src http', 'face', 'href', 'color', 'border', 'img', 'src', 'table']
----------------------------------------------------------------------------
----
CLUSTER #10 : 0.26208587402740147
Cluster Size 203
Key Features: ['exmh', 'exmh workers', 'workers', 'redhat com', 'listman', '
redhat', 'deepeddy', 'mx1', 'spamassassin taint org', 'spamassassin taint',
'taint org', 'taint', 'vircio', 'spamassassin', 'vircio com']
----------------------------------------------------------------------------
----
CLUSTER #11 : 0.17842591134414898
Cluster Size 799
Key Features: ['ilug', 'linux', 'ie', 'linux ie', 'lugh', 'ilug linux', 'tua
tha org', 'tuatha', 'lugh tuatha', 'lugh tuatha org', 'admin linux', 'ilug a
dmin', 'ilug admin linux', 'aug', 'ilug linux ie']
----------------------------------------------------------------------------
----
CLUSTER #12 : 0.3254622837978718
Cluster Size 295
Key Features: ['razor', 'razor users', 'sourceforge net', 'sourceforge', 'us
ers', 'net', 'example sourceforge', 'example sourceforge net', 'lists', 'lis
```

```
ts sourceforge', 'lists sourceforge net', 'usw', 'example', 'users example s
ourceforge', 'razor users example']
----------------------------------------------------------------------------
----
CLUSTER #13 : 0.082248157326386
Cluster Size 623
Key Features: ['example com', 'rssfeeds', 'oct', 'example', 'rssfeeds exampl
e com', 'rssfeeds example', 'tue oct', 'sep', 'spam', 'jmason org', 'jmason'
, 'path rssfeeds example', 'level url', 'spam level url', 'level url http']
----------------------------------------------------------------------------
----
CLUSTER #14 : 0.37067939907264863
Cluster Size 103
Key Features: ['spamassassin devel', 'devel', 'sourceforge net', 'sourceforg
e', 'net', 'spamassassin', 'example sourceforge net', 'example sourceforge',
'lists', 'lists sourceforge', 'lists sourceforge net', 'mailto spamassassin
devel', 'spamassassin devel admin', 'devel admin', 'spamassassin devel examp
le']
----------------------------------------------------------------------------
----
CLUSTER #15 : 0.06698882245980195
Cluster Size 264
Key Features: ['tim', 'comcast net', 'comcast', 'tim one', 'one comcast net'
, 'one comcast', 'tim one comcast', 'spambayes', 'skip', 'guido', 'python',
'ham', 'python org', 'subject spambayes', 'sep subject spambayes']
----------------------------------------------------------------------------
----
```

In [16]:

```python
df.pivot_table(index='kmeans_cluster', columns='is_spam', values='directory'
```

Out[16]:

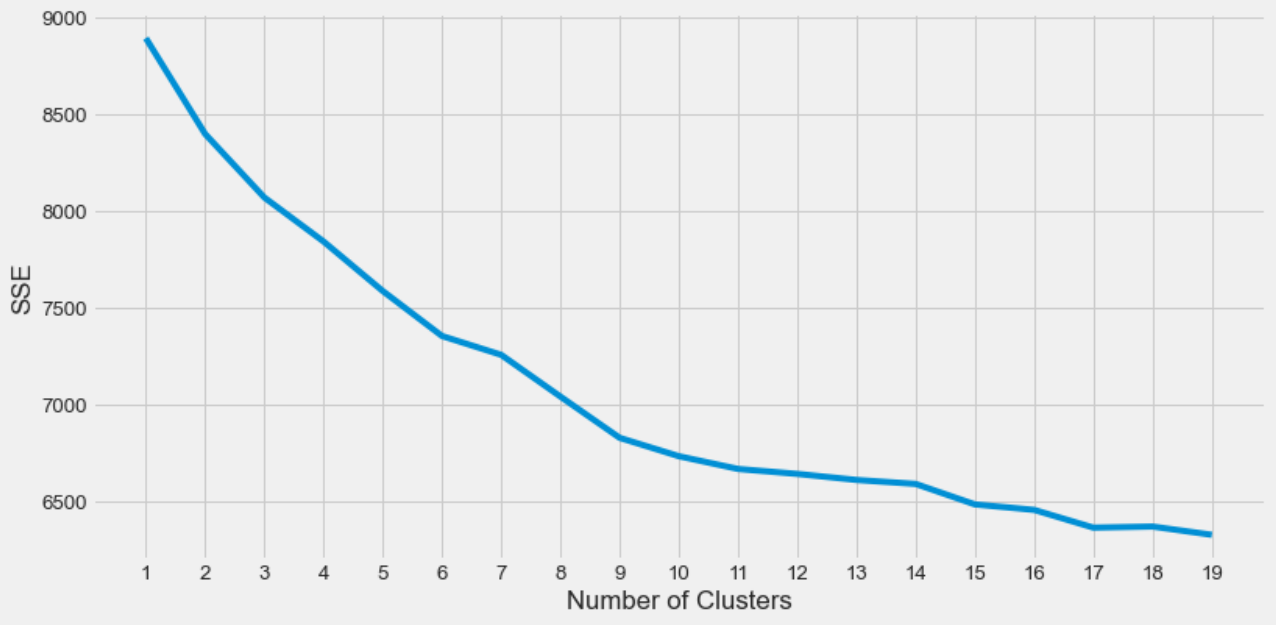
is_spam	0	1
kmeans_cluster		
0	218.0	NaN
1	605.0	NaN
2	736.0	1385.0
3	59.0	415.0
4	658.0	NaN
5	348.0	19.0
6	1691.0	50.0
7	206.0	NaN
8	254.0	422.0
9	203.0	NaN
10	691.0	108.0
11	295.0	NaN
12	623.0	NaN
13	103.0	NaN
14	264.0	NaN

In [17]:

```
scores = []

for k in range(1, 20):
    kmeans = KMeans(init="random", n_clusters=k, n_init=10, max_iter=300, ra
    kmeans.fit(tf_matrix)
    scores.append(kmeans.inertia_)
```

In [18]:

```
plt.style.use("fivethirtyeight")
plt.figure(figsize=(12,6))
plt.plot(range(1, 20), scores)
plt.xticks(range(1, 20))
plt.xlabel("Number of Clusters")
plt.ylabel("SSE")
plt.show()
```



Model Building

Using Original Data

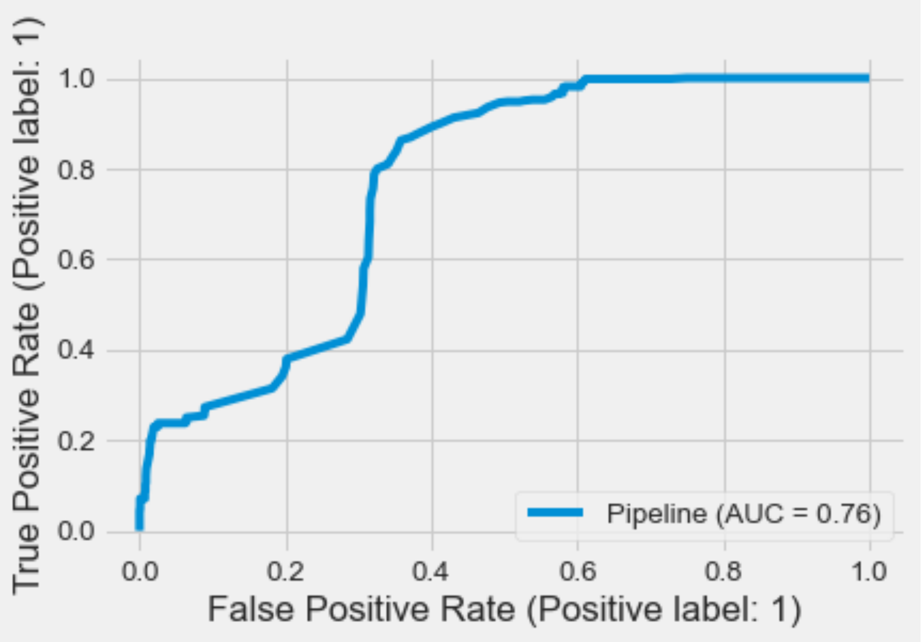
In [19]:

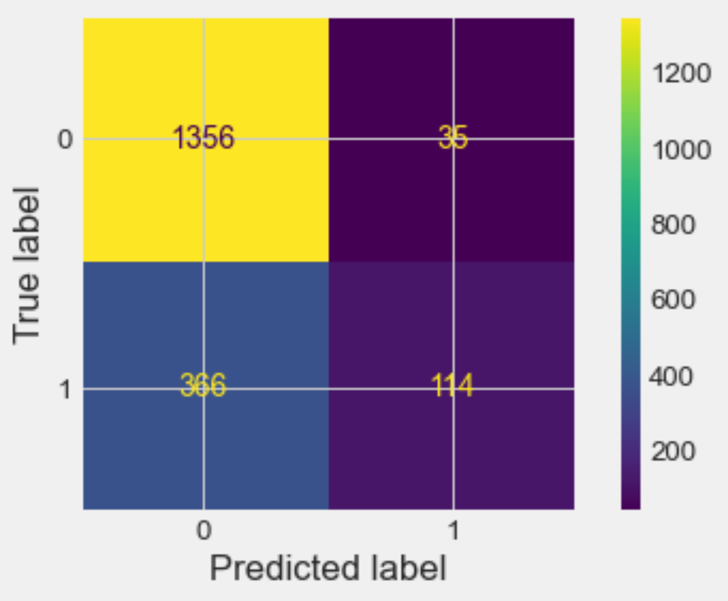
```
categorical_features = ['in_reply','subj_caps','attachments']
numeric_features = ['body_lines']

X = df[categorical_features + numeric_features]
y = df['is_spam']

X_train, X_test, y_train, y_test = train_test_split(X,
                                                    y,
                                                    test_size=0.2,
                                                    stratify=y,
                                                    random_state=11)


numeric_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(missing_values=np.nan, strategy="median")),
    ('scaler', RobustScaler(with_centering=False))])

categorical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(missing_values=np.nan, strategy="most_frequent
    ('onehot', OneHotEncoder(handle_unknown='ignore'))])

preprocessor = ColumnTransformer(
    transformers=[
        ('num', numeric_transformer, numeric_features),
        ('cat', categorical_transformer, categorical_features)])

clfPipeline = Pipeline(steps = [['preprocessor', preprocessor],['classifier'


clfPipeline.fit(X_train, y_train)

y_pred = clfPipeline.predict(X_test)

print("Accurracy:", accuracy_score(y_test, y_pred))
print("Recall:", recall_score(y_test, y_pred))
print("Precision:", precision_score(y_test, y_pred))
print("F1", f1_score(y_test, y_pred))
```

```
Accurracy: 0.7856761090326029
Recall: 0.2375
Precision: 0.7651006711409396
F1 0.36248012718600947
```

In [20]:

```
disp = RocCurveDisplay.from_estimator(clfPipeline, X_test, y_test)
disp = ConfusionMatrixDisplay.from_estimator(clfPipeline, X_test, y_test)
```





Using Previously Identified Clusters

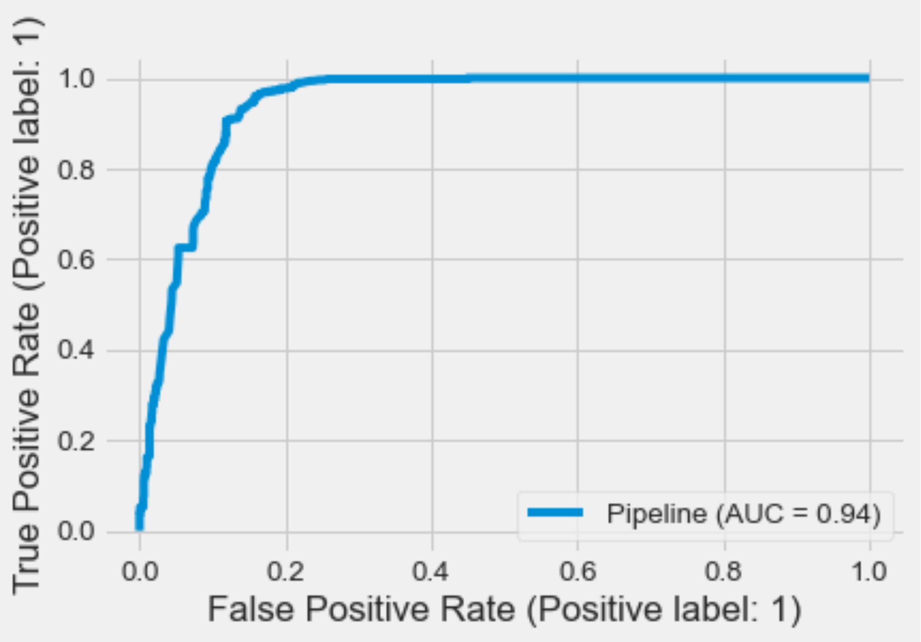
In [21]:

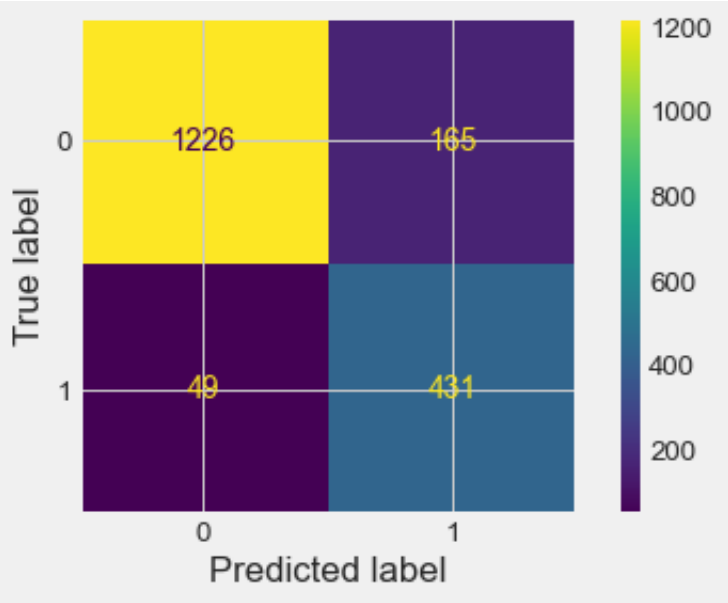
```
categorical_features = ['in_reply','subj_caps','attachments','kmeans_cluster
numeric_features = ['body_lines']

X = df[categorical_features + numeric_features]
y = df['is_spam']

X_train, X_test, y_train, y_test = train_test_split(X,
                                                    y,
                                                    test_size=0.2,
                                                    stratify=y,
                                                    random_state=11)

numeric_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(missing_values=np.nan, strategy="median")),
    ('scaler', RobustScaler(with_centering=False))])

categorical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(missing_values=np.nan, strategy="most_frequent
    ('onehot', OneHotEncoder(handle_unknown='ignore'))])

preprocessor = ColumnTransformer(
    transformers=[
        ('num', numeric_transformer, numeric_features),
        ('cat', categorical_transformer, categorical_features)])

clfPipeline = Pipeline(steps = [['preprocessor', preprocessor],['classifier'


clfPipeline.fit(X_train, y_train)

y_pred = clfPipeline.predict(X_test)

print("Accurracy:", accuracy_score(y_test, y_pred))
print("Recall:", recall_score(y_test, y_pred))
print("Precision:", precision_score(y_test, y_pred))
print("F1", f1_score(y_test, y_pred))
```

```
Accurracy: 0.8856226616782469
Recall: 0.8979166666666667
Precision: 0.7231543624161074
F1 0.8011152416356878
```

In [22]:

```
disp = RocCurveDisplay.from_estimator(clfPipeline, X_test, y_test)
disp = ConfusionMatrixDisplay.from_estimator(clfPipeline, X_test, y_test)
```


True Positive Rate (Positive label: 1)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (Positive label: 1)
Pipeline (AUC = 0.94)


True label
0
1
1226
165
49
431
0
1
Predicted label
1200
1000
800
600
400
200