```
!git clone https://github.com/bellevue-university/dsc650.git
```

```
Cloning into 'dsc650'...
remote: Enumerating objects: 120326, done.
remote: Counting objects: 100% (128/128), done.
remote: Compressing objects: 100% (47/47), done.
remote: Total 120326 (delta 54), reused 117 (delta 49), pack-reused 120198
Receiving objects: 100% (120326/120326), 360.60 MiB | 16.11 MiB/s, done.
Resolving deltas: 100% (7337/7337), done.
Updating files: 100% (114699/114699), done.
```

```
import os
import json
from pathlib import Path
path="/content/dsc650/dsc650/assignments/assignment04"
os.chdir(path)
current_dir = Path(os.getcwd()).absolute()
enron_data_dir = "/content/dsc650/data/external/enron"
```

```
from pyspark.sql import SparkSession
spark = SparkSession\
    .builder\
    .appName("Assignment04")\
    .getOrCreate()
```

Assignment 4.1

```
import pandas as pd
from chardet.universaldetector import UniversalDetector
from pyspark.sql.types import StructType, StringType, StructField
```

```
def read_raw_email(email_path):
    detector = UniversalDetector()

    try:
        with open(email_path) as f:
            original_msg = f.read()
    except UnicodeDecodeError:
        detector.reset()
        with open(email_path, 'rb') as f:
            for line in f.readlines():
                detector.feed(line)
                if detector.done:
                    break
        detector.close()
        encoding = detector.result['encoding']
        with open(email_path, encoding=encoding) as f:
            original_msg = f.read()

    return original_msg

def make_spark_df():
    records = []
    processed_files = set()

    for root, dirs, files in os.walk(enron_data_dir):
        for file_path in files:
            current_path = Path(root).joinpath(file_path)
            #print(os.path.join(root, file_path))
            if current_path in processed_files:
                continue

            processed_files.add(current_path)

            record = {}
            record['id']  = str(current_path.relative_to(enron_data_dir))
            record['id'] = record['id'].replace('\\', '/')
            record['username'] = os.path.basename(os.path.dirname(current_path))
            record['og_msg'] = read_raw_email(current_path)

            records.append(record)
            ## TODO: Complete the code to code to create the Spark dataframe
    schemaString = "id username og_msg"
```

```
        fields = [StructField(field_name, StringType(), True) for field_name in schemaString.split()]
        schema = StructType(fields)
        return spark.createDataFrame(data=records,schema=schema)

df = make_spark_df()


df.show(10)
```

```
+--------------------+----------+--------------------+
|                  id|  username|              og_msg|
+--------------------+----------+--------------------+
|meyers-a/sent_ite...|sent_items|Message-ID: <2231...|
|meyers-a/sent_ite...|sent_items|Message-ID: <1495...|
|meyers-a/sent_ite...|sent_items|Message-ID: <1423...|
|meyers-a/sent_ite...|sent_items|Message-ID: <4447...|
|meyers-a/sent_ite...|sent_items|Message-ID: <1679...|
|meyers-a/sent_ite...|sent_items|Message-ID: <4681...|
|meyers-a/sent_ite...|sent_items|Message-ID: <1392...|
|meyers-a/sent_ite...|sent_items|Message-ID: <2649...|
|meyers-a/sent_ite...|sent_items|Message-ID: <2536...|
|meyers-a/sent_ite...|sent_items|Message-ID: <3912...|
+--------------------+----------+--------------------+
only showing top 10 rows
```

Assignment **4.2**

```
plain_msg_example = """
Message-ID: <6742786.1075845426893.JavaMail.evans@thyme>
Date: Thu, 7 Jun 2001 11:05:33 -0700 (PDT)
From: jeffrey.hammad@enron.com
To: andy.zipper@enron.com
Subject: Thanks for the interview
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Hammad, Jeffrey </O=ENRON/OU=NA/CN=RECIPIENTS/CN=NOTESADDR/CN=CBBE377A-24F58854-862567DD-591AE7>
X-To: Zipper, Andy </O=ENRON/OU=NA/CN=RECIPIENTS/CN=AZIPPER>
X-cc:
X-bcc:
X-Folder: \Zipper, Andy\Zipper, Andy\Inbox
X-Origin: ZIPPER-A
X-FileName: Zipper, Andy.pst

Andy,

Thanks for giving me the opportunity to meet with you about the Analyst/ Associate program.  I enjoyed talking to you, and look forward to cc

Thanks and Best Regards,

Jeff Hammad
"""

html_msg_example = """
Message-ID: <21013632.1075862392611.JavaMail.evans@thyme>
Date: Mon, 19 Nov 2001 12:15:44 -0800 (PST)
From: insynconline.6jy5ympb.d@insync-palm.com
To: tstaab@enron.com
Subject: Last chance for special offer on Palm OS Upgrade!
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: InSync Online <InSyncOnline.6jy5ympb.d@insync-palm.com>
X-To: THERESA STAAB <tstaab@enron.com>
X-cc:
X-bcc:
X-Folder: \TSTAAB (Non-Privileged)\Staab, Theresa\Deleted Items
X-Origin: Staab-T
X-FileName: TSTAAB (Non-Privileged).pst

<html>

<html>
<head>
<title>Paprika</title>
<meta http-equiv="Content-Type" content="text/html;">
```

```
</head>
<body bgcolor="#FFFFFF" TEXT="#333333" LINK="#336699" VLINK="#6699cc" ALINK="#ff9900">
<table border="0" cellpadding="0" cellspacing="0" width="582">
<tr valign="top">
  <td width="582" colspan="9"><nobr><a href="http://insync-online.p04.com/u.d?BEReaQA5eczXB=1"><img src="http://images4.postdirect.com/master
</tr>
<tr valign="top">
  <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="4" height="1" hspace="0" vsp
  <td width="20"><img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="20" height="1" hspace="0" vspace="0" border="
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?LkReaQA5eczXL=21"><img src="http://images4.postdirect.com/master-images/40470
  <td width="20"><img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="20" height="1" hspace="0" vspace="0" border="
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?BkReaQA5eczXO=31"><img src="http://images4.postdirect.com/master-images/40470
  <td width="20"><img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="20" height="1" hspace="0" vspace="0" border="
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?JkReaQA5eczXRs=41"><img src="http://images4.postdirect.com/master-images/4047
  <td width="19"><img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="19" height="1" hspace="0" vspace="0" border="
  <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="4" height="1" hspace="0" vsp
</tr>
</table>
<table border="0" cellpadding="0" cellspacing="0" width="582">
<tr valign="top">
  <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="4" height="1" hspace="0" vsp
  <td width="574"><br>
    <table border="0" cellpadding="0" cellspacing="0" width="574" bgcolor="#99ccff">
    <tr>
      <td width="50"><img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="50" height="1" hspace="0" vspace="0" borc
      <td width="474"><font face="verdana, arial" size="-2"color="#000000">
        <br>
        Dear THERESA,
        <br><br>
        Due to overwhelming demand for the Palm OS&#174; v4.1 Upgrade with Mobile Connectivity, we are
        extending the special offer of 25% off through November 30, 2001. So there's still time to significantly
        increase the functionality of your Palm&#153; III, IIIx, IIIxe, IIIc, V or Vx handheld. Step up to the
        new Palm OS v4.1 through this extended special offer. You'll receive the brand new Palm OS v4.1
        <b>for just $29.95 when you use Promo Code <font color="#FF0000">OS41WAVE</font></b>. That's a
        <b>$10 savings</b> off the list price.
        <br><br>
        <a href="http://insync-online.p04.com/u.d?NkReaQA5eczXRh=51">Click here to view a full product demo now</a>.
        <br><br>
        <a href="http://insync-online.p04.com/u.d?MkReaQA5eczXRm=61"><img src="http://images4.postdirect.com/master-images/404707/title1.gif"
        <br><br>
        You can do a lot more with your Palm&#153; handheld when you upgrade to the Palm OS v4.1. All your
        favorite features just got even better and there are some terrific new additions:
        <br><br>
        <LI> Handwrite notes and even draw pictures right on your Palm&#153 handheld</LI>
        <LI> Tap letters with your stylus and use Graffiti&#174; at the same time with the enhanced onscreen keyboard</LI>
        <LI> Improved Date Book functionality lets you view, snooze or clear multiple alarms all with a single tap </LI>
        <LI> You can easily change time-zone settings</LI>

        <br><br>
        <a href="http://insync-online.p04.com/u.d?WkReaQA5eczXRb=71"><img src="http://images4.postdirect.com/master-images/404707/title2.gif"
        <br><br>
        <LI> <nobr>Mask/unmask</nobr> private records or hide/unhide directly within the application</LI>
        <LI> Lock your device automatically at a designated time using the new Autolocking feature</LI>
        <LI> Always remember your password with our new Hint feature*</LI>

        <br><br>
        <a href="http://insync-online.p04.com/u.d?VEReaQA5eczXRQ=81"><img src="http://images4.postdirect.com/master-images/404707/title3.gif"
        <br><br>
        <LI> Use your GSM compatible mobile phone or modem to get online and access the web</LI>
        <LI> Stay connected with email, instant messaging and text messaging to GSM mobile phones</LI>
        <LI> Send applications or records through your cell phone to schedule meetings and even "beam"
             important information to others</LI>

        <br><br>
        All this comes in a new operating system that can be yours for just $29.95! <a href="http://insync-online.p04.com/u.d?MkReaQA5eczXRV=
        upgrade to the new Palm&#153; OS v4.1</a> and you'll also get the latest Palm desktop software. Or call
        <nobr>1-800-881-7256</nobr> to order via phone.
        <br><br>
        Sincerely,<br>
        The Palm Team
        <br><br>
        P.S. Remember, this extended offer opportunity of 25% savings absolutely ends on November 30, 2001
        and is only available through the Palm Store when you use Promo Code <b><font color="#FF0000">OS41WAVE</font></b>.
        <br><br>
        <img src="http://images4.postdirect.com/master-images/404707/bottom_button.gif" align="right" alt="" width="295" height="60" hspace="
        <br><img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="474" height="1" hspace="0" vspace="0" border="0" a
        </font></td>
```

```
        <td width="50"><img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="50" height="1" hspace="0" vspace="0" borc
      </tr>
      </table></td>
      <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="4" height="1" hspace="0" v
    </tr>
    <tr>
    <td colspan="3"><img src="http://images4.postdirect.com/master-images/404707/bottom.gif" width="582" height="67" hspace="0" vspace="0" borc
    </tr>
  </table>
  <table border="0" cellpadding="0" cellspacing="0" width="582">
    <tr>
      <td width="54"><img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="54" height="1" hspace="0" vspace="0" border
      <td width="474"><font face="arial, verdana" size="-2" color="#000000"><br>
      * This feature is available on the Palm&#153; IIIx, Palm&#153; IIIxe, and Palm&#153; Vx. <br><br>
      ** Note: To use the MIK functionality, you need either a Palm OS&#174; compatible modem or a phone
      with  <nobr>built-in</nobr> modem or data capability that has either an infrared port or cable exits.  If you
      are using a phone, you must have data services from your mobile service provider.  <a href="http://insync-online.p04.com/u.d?RkReaQA5eczX
      a list of tested and supported phones that you can use with the MIK. Cable not provided.
      <br><br>
      ------------------<br>
      To modify your profile or unsubscribe from Palm newsletters, <a href="http://insync-online.p04.com/u.d?KkReaQA5eczXRE=121">click here</a>
      Or, unsubscribe by replying to this message, with "unsubscribe" as the subject line of the message.
      <br><br>
      ------------------<br>
      Copyright&#169; 2001 Palm, Inc. Palm OS, Palm Computing, HandFAX, HandSTAMP, HandWEB, Graffiti,
      HotSync, iMessenger, MultiMail, Palm.Net, PalmConnect, PalmGlove, PalmModem, PalmPoint, PalmPrint,
      and the Palm Platform Compatible Logo are registered trademarks of Palm, Inc. Palm, the Palm logo,
      AnyDay, EventClub, HandMAIL, the HotSync Logo, PalmGear, PalmGlove, PalmPix, Palm Powered, the Palm
      trade dress, PalmSource, Smartcode, and Simply Palm are trademarks of Palm, Inc. All other brands and
      product names may be trademarks or registered trademarks of their respective owners.</font>
      <img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="474" height="1" hspace="0" vspace="0" border="0" alt=""></
      <td width="54"><img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="54" height="1" hspace="0" vspace="0" border
    </tr>
  </table><br><br><br><br>
  <!-- The following image is included for message detection -->
  <img src="http://p04.com/1x1.dyn" border="0" alt="" width="1" height="1">
  <img src="http://p04.com/1x1.dyn?0vEGou8Ig30ba2L2bLn" width=1 height=1></body>
  </html>

  </html>
  """
  plain_msg_example = plain_msg_example.strip()
  html_msg_example = html_msg_example.strip()


  from bs4 import BeautifulSoup
  from email.parser import Parser
  from dateutil.parser import parse
  from collections import namedtuple
  from email.policy import default


  output_columns = [
          'payload',
          'text',
          'Message_D',
          'Date',
          'From',
          'To',
          'Subject',
          'Mime-Version',
          'Content-Type',
          'Content-Transfer-Encoding',
          'X-From',
          'X-To',
          'X-cc',
          'X-bcc',
          'X-Folder',
          'X-Origin',
          'X-FileName',
          'Cc',
          'Bcc'
  ]

  columns = [column.replace('-', '_') for column in output_columns]

  ParsedEmail = namedtuple('ParsedEmail', columns)
```

```
def parse_html_payload(payload):
    """
    This function uses Beautiful Soup to read HTML data
    and return the text.  If the payload is plain text, then
    Beautiful Soup will return the original content
    """
    soup = BeautifulSoup(payload, 'html.parser')
    return str(soup.get_text()).encode('utf-8').decode('utf-8')

def parse_email(original_msg):
    result = {}
    # parse 'original_msg'
    msg = Parser(policy=default).parsestr(original_msg)

    ## TODO: Use Python's email library to read the payload and the headers
    ## https://docs.python.org/3/library/email.examples.html

    # get email body content from parsed 'original_msg'
    payload = msg.get_payload()

    # extract text from email body 'payload'
    result['text'] = parse_html_payload(payload)

    # extract individual parts of email message
    for item in msg:
        result[item] = msg[item]

    # create and return tuple of parsed email parts
    tuple_result = tuple([str(result.get(column, None)) for column in columns])
    return ParsedEmail(*tuple_result)


print(columns)
```

```
['payload', 'text', 'Message_D', 'Date', 'From', 'To', 'Subject', 'Mime_Version', 'Content_Type', 'Content_Transfer_Encoding', 'X_From'
```

```
parsed_msg = parse_email(plain_msg_example)
```

```
<ipython-input-30-641fbc64e851>:7: MarkupResemblesLocatorWarning: The input looks more like a filename than markup. You may want to ope
  soup = BeautifulSoup(payload, 'html.parser')
```

```
print(parsed_msg.text)
```

```
Andy,

Thanks for giving me the opportunity to meet with you about the Analyst/ Associate program.  I enjoyed talking to you, and look forward

Thanks and Best Regards,

Jeff Hammad
```

```
parsed_html_msg = parse_email(html_msg_example)


print(parsed_html_msg.text)
```

```


Paprika
```

```
Dear THERESA,

Due to overwhelming demand for the Palm OS® v4.1 Upgrade with Mobile Connectivity, we are
extending the special offer of 25% off through November 30, 2001. So there's still time to significantly
increase the functionality of your Palm™ III, IIIx, IIIxe, IIIc, V or Vx handheld. Step up to the
new Palm OS v4.1 through this extended special offer. You'll receive the brand new Palm OS v4.1
for just $29.95 when you use Promo Code OS41WAVE. That's a
$10 savings off the list price.

 Click here to view a full product demo now.



You can do a lot more with your Palm™ handheld when you upgrade to the Palm OS v4.1. All your
favorite features just got even better and there are some terrific new additions:

  Handwrite notes and even draw pictures right on your Palm™ handheld
  Tap letters with your stylus and use Graffiti® at the same time with the enhanced onscreen keyboard
  Improved Date Book functionality lets you view, snooze or clear multiple alarms all with a single tap
  You can easily change time-zone settings



  Mask/unmask private records or hide/unhide directly within the application
  Lock your device automatically at a designated time using the new Autolocking feature
  Always remember your password with our new Hint feature*
```

```
print(parsed_html_msg.Subject)
```

```
 Last chance for special offer on Palm OS Upgrade!
```

Assignment 4.3

```
from chardet.universaldetector import UniversalDetector
from pyspark.ml.pipeline import Transformer
from pyspark.sql.functions import udf
from pyspark.ml.feature import HashingTF, Tokenizer
from pyspark.ml import Pipeline
from pyspark.ml.feature import CountVectorizer


## This creates a schema for the email data
email_struct = StructType()

for column in columns:
    email_struct.add(column, StringType(), True)


## This creates a user-defined function which can be used in Spark
parse_email_func = udf(lambda z: parse_email(z), email_struct)

def parse_emails(input_df):
    new_df = input_df.select('username', 'id', 'og_msg', parse_email_func('og_msg').alias('parsed_email'))
    for column in columns:
        new_df = new_df.withColumn(column, new_df.parsed_email[column])

    new_df = new_df.drop('parsed_email')
    return new_df

class ParseEmailsTransformer(Transformer):
    def _transform(self, dataset):
        """
        Transforms the input dataset.

        :param dataset: input dataset, which is an instance of :py:class:`pyspark.sql.DataFrame`
        :returns: transformed dataset
```

```
        """
        return dataset.transform(parse_emails)

## Use the custom ParseEmailsTransformer, Tokenizer, and CountVectorizer
## to create a spark pipeline

# set up tokenizer for email text
tokenizer = Tokenizer(inputCol="text", outputCol="words")

# set up count vectorizer for tokenized words
cv = CountVectorizer(inputCol="words", outputCol="features",
                     vocabSize=3, minDF=2.0)
# cv = CountVectorizer(inputCol="words", outputCol="features")

# set up pipline for processing / transforming dataframe
email_pipeline = Pipeline(
    ## TODO: Complete code
    stages=[ParseEmailsTransformer(), tokenizer, cv]
)

model = email_pipeline.fit(df)
result = model.transform(df)


result.select('id', 'words', 'features').show()
```

```
+--------------------+--------------------+--------------------+
|                  id|               words|            features|
+--------------------+--------------------+--------------------+
|meyers-a/sent_ite...|[, , -----origina...|      (3,[0],[28.0])|
|meyers-a/sent_ite...|[tom:, , i, would...|(3,[0,1,2],[33.0,...|
|meyers-a/sent_ite...|[bill:, , please,...|(3,[0,1,2],[7.0,6...|
|meyers-a/sent_ite...|[kathy:, , please...| (3,[0,1],[7.0,1.0])|
|meyers-a/sent_ite...|[tom:, , if, it, ...|(3,[0,1,2],[5.0,1...|
|meyers-a/sent_ite...|[tom:, , buffalo,...|       (3,[0],[2.0])|
|meyers-a/sent_ite...|[bill:, , for, ye...|(3,[0,1,2],[8.0,6...|
|meyers-a/sent_ite...|[tom:, , earlier,...|(3,[0,1,2],[4.0,1...|
|meyers-a/sent_ite...|[tom:, , just, wa...|(3,[0,1,2],[19.0,...|
|meyers-a/sent_ite...|[bill:, , last, n...|(3,[0,1,2],[3.0,2...|
|meyers-a/sent_ite...|[amy,, , please, ...|(3,[0,1,2],[2.0,3...|
|meyers-a/deleted_...|[, , start, date:...|      (3,[0],[29.0])|
|meyers-a/deleted_...|[, , start, date:...|      (3,[0],[11.0])|
|meyers-a/deleted_...|[, , start, date:...|       (3,[0],[9.0])|
|meyers-a/deleted_...|[, , start, date:...|      (3,[0],[11.0])|
|meyers-a/deleted_...|[fyi, -, , the, s...| (3,[0,1],[9.0,1.0])|
|meyers-a/deleted_...|[, , start, date:...|(3,[0,1,2],[72.0,...|
|meyers-a/deleted_...|[, , start, date:...|      (3,[0],[11.0])|
|meyers-a/deleted_...|[to, whom, it, ma...|(3,[0,1,2],[9.0,7...|
|meyers-a/deleted_...|[, , start, date:...|       (3,[0],[3.0])|
+--------------------+--------------------+--------------------+
only showing top 20 rows
```

✓ 2s completed at 12:33 AM