```
## Mounting Drive
import os
from google.colab import drive
drive.mount('/content/drive', force_remount = True)
os.chdir(r"/content/drive/My Drive/dsc650/dsc650/assignments/assignment04")
!pwd
```

```
     Mounted at /content/drive
     /content/drive/My Drive/dsc650/dsc650/assignments/assignment04
```

```
## Needed to install openjdk to eliminate pysaprk error message.
import os
import json
from pathlib import Path
import zipfile
import email
from email.policy import default
from email.parser import Parser
from datetime import timezone
from collections import namedtuple

import pandas as pd
## import s3fs
from bs4 import BeautifulSoup
from dateutil.parser import parse
## Install chardet
!pip install chardet
from chardet.universaldetector import UniversalDetector
## install pyspark
!pip install pyspark
from pyspark.ml import Pipeline
from pyspark.ml.feature import CountVectorizer
from pyspark.ml.feature import HashingTF, Tokenizer
from pyspark.sql import SparkSession
from pyspark.sql.functions import col
from pyspark.ml.pipeline import Transformer
from pyspark.sql.functions import udf
from pyspark.sql.types import StructType, StringType

import pandas as pd

current_dir = Path(os.getcwd()).absolute()
results_dir = current_dir.joinpath('results')
results_dir.mkdir(parents=True, exist_ok=True)
data_dir = current_dir.joinpath('data')
data_dir.mkdir(parents=True, exist_ok=True)
enron_data_dir =  '/content/drive/My Drive/dsc650/data/external/enron'

## os.environ["PYSPARK_PYTHON"] = "C:/Users/jkmey/anaconda3/envs/dsc650/python.exe"

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]

## Show the output for column names to prove code executed.
print(columns)
```

```
print(columns)

ParsedEmail = namedtuple('ParsedEmail', columns)

## Print the ParsedEmail to confirm the code executed.
print(ParsedEmail)

## Import findspark
## import findspark
## findspark.init()

spark = SparkSession\
    .builder\
    .appName("Assignment04")\
    .getOrCreate()
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: chardet in /usr/local/lib/python3.9/dist-packages (4.0.0)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting pyspark
  Downloading pyspark-3.3.2.tar.gz (281.4 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 281.4/281.4 MB 5.4 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Collecting py4j==0.10.9.5
  Downloading py4j-0.10.9.5-py2.py3-none-any.whl (199 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 199.7/199.7 KB 22.9 MB/s eta 0:00:00
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.3.2-py2.py3-none-any.whl size=281824028 sha256=ac14de62de4e15ebfcc3dc6c26b1bb7d172dfb932
  Stored in directory: /root/.cache/pip/wheels/6c/e3/9b/0525ce8a69478916513509d43693511463c6468db0de237c86
Successfully built pyspark
Installing collected packages: py4j, pyspark
  Attempting uninstall: py4j
    Found existing installation: py4j 0.10.9.7
    Uninstalling py4j-0.10.9.7:
      Successfully uninstalled py4j-0.10.9.7
Successfully installed py4j-0.10.9.5 pyspark-3.3.2
['payload', 'text', 'Message_D', 'Date', 'From', 'To', 'Subject', 'Mime_Version', 'Content_Type', 'Content_Transfer_Encoding', 'X_From',
<class '__main__.ParsedEmail'>
```

The following code loads data to your local JupyterHub instance. You only need to run this once.

```
'''
Commenting out this entire cell and manually placed the files into the Assignment 4 data folder.
The files were pulled from C:/Users/jkmey/Documents/Github/DSC650_Course_Assignments/dsc650/data/external/enron/
and moved to C:/Users/jkmey/Documents/Github/DSC650_Course_Assignments/dsc650/dsc650/assignments/assignment04/data/enron.
I did not zip the files because I was running into issues when executing the code with enron.zip. Left files uncompressed.
'''
'''
def copy_data_to_local():
    dst_data_path = data_dir.joinpath('enron.zip')
    endpoint_url='https://storage.budsc.midwest-datascience.com'
    enron_data_path = 'data/external/enron.zip'

    s3 = s3fs.S3FileSystem(
        anon=True,
        client_kwargs={
            'endpoint_url': endpoint_url
        }
    )


    s3.get(enron_data_path, str(dst_data_path))

    with zipfile.ZipFile(dst_data_path) as f_zip:
        f_zip.extractall(path=data_dir)

copy_data_to_local()
'''
```

```
'\ndef copy_data_to_local():\n    dst_data_path = data_dir.joinpath('enron.zip')
\n    endpoint_url='https://storage.budsc.midwest-datascience.com'\n    enron_dat
a_path = 'data/external/enron.zip'\n\n    s3 = s3fs.S3FileSystem(\n        anon=T
rue,\n        client_kwargs={\n            'endpoint_url': endpoint_url\n
```

This code reads emails and creates a Spark dataframe with three columns.

Assignment 4.1

```
## Import additional packages from pyspark
from pyspark.sql.types import StructField
from pyspark.sql.types import StructType
from pyspark.sql.types import StringType

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 = []
    print("enron_data_dir", enron_data_dir)
    for root, dirs, files in os.walk(enron_data_dir):
        for file_path in files:
            ## Current path is now the file path to the current email.
            ## Use this path to read the following information
            ## original_msg
            ## username (Hint: It is the root folder)
            ## id (The relative path of the email message)
            current_path = Path(root).joinpath(file_path)
            record = {}
            username = os.path.basename(os.path.dirname(root))
            id = username+"/"+os.path.basename(root)+"/"+file_path
            record["id"] = id
            record["username"] = username
            record["original_msg"] = read_raw_email(current_path)
            records.append(record)

    ## Complete the code to create the Spark dataframe
    schemaString = "id username original_msg"
    fields = [StructField(field_name, StringType(), True) for field_name in schemaString.split()]
    schema = StructType(fields)
    return spark.createDataFrame(records, schema)

df = make_spark_df()
```

```
 enron_data_dir /content/drive/My Drive/dsc650/data/external/enron
```

```
df.show()
```

```
 +------------------+--------+--------------------+
 |                id|username|        original_msg|
 +------------------+--------+--------------------+
 |davis-d/2_trash/1_| davis-d|Message-ID: <1774...|
 |davis-d/2_trash/3_| davis-d|Message-ID: <2833...|
 |davis-d/2_trash/4_| davis-d|Message-ID: <1972...|
 |davis-d/2_trash/2_| davis-d|Message-ID: <2467...|
 |2_trash/candis/12_| 2_trash|Message-ID: <5686...|
 |2_trash/candis/10_| 2_trash|Message-ID: <1964...|
 |2_trash/candis/11_| 2_trash|Message-ID: <7345...|
 |2_trash/candis/13_| 2_trash|Message-ID: <7218...|
 |2_trash/candis/14_| 2_trash|Message-ID: <3016...|
 |2_trash/candis/15_| 2_trash|Message-ID: <1233...|
 |2_trash/candis/17_| 2_trash|Message-ID: <1365...|
 |2_trash/candis/16_| 2_trash|Message-ID: <2215...|
 |2_trash/candis/19_| 2_trash|Message-ID: <8556...|
 | 2_trash/candis/1_| 2_trash|Message-ID: <1807...|
```

```
|2_trash/candis/18_| 2_trash|Message-ID: <2251...|
| 2_trash/candis/3_| 2_trash|Message-ID: <2977...|
| 2_trash/candis/2_| 2_trash|Message-ID: <2705...|
| 2_trash/candis/4_| 2_trash|Message-ID: <3065...|
| 2_trash/candis/5_| 2_trash|Message-ID: <2798...|
| 2_trash/candis/6_| 2_trash|Message-ID: <3108...|
+------------------+--------+--------------------+
only showing top 20 rows
```

```
df.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- username: string (nullable = true)
 |-- original_msg: string (nullable = true)
```

Assignment 4.2

Use `plain_msg_example` and `html_msg_example` to create a function that parses an email message.

```
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 co

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" bord
      <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" bord
    </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" bord
  </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()


## Test out the email reader for the html path.
email_path_html = '/content/drive/My Drive/dsc650/dsc650/assignments/assignment04/examples/html//message-0.txt'
print(read_raw_email(email_path_html))
```

```
Message-ID: <9603832.1075862117623.JavaMail.evans@thyme>
Date: Mon, 26 Nov 2001 09:43:37 -0800 (PST)
From: irene@m-ul.com
To: pmims@enron.com
Subject: 4 FREE Airline Tickets!!
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Travel Agent <irene@m-ul.com>@ENRON
X-To: pmims@enron.com
X-cc:
X-bcc:
X-Folder: \PMIMS (Non-Privileged)\Mims, Patrice L.\Deleted Items
X-Origin: Mims-Thurston-P
X-FileName: PMIMS (Non-Privileged).pst


<html>
<head>
<title>Untitled Document</title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
</head>

<body bgcolor="#FFFFFF"><center>
<table width="540" border="1" cellspacing="0" cellpadding="3" bordercolor="#999999">
 <tr bgcolor="#0099FF">
 <td bgcolor="#FFFFFF" height="48">
 <div align="center"><font face="Verdana, Arial, Helvetica, sans-serif" size="2">If
 you do not wish to receive email from Bargain Bazaar, please Click Here:<br>
 <a href="http://www.m-ul.com/e/unsub.cgi?l=9994&m=1736394363">UNSUBSCRIBE</a>
 </font></div>
 </td>
 </tr>
</table>
<br>
```

```
<table width="540" border="1" cellspacing="0" cellpadding="0" bordercolor="#666600">
<tr>
<td>
<table width="100%" border="1" cellspacing="0" cellpadding="0" bordercolor="#CCFFFF">
<tr>
<td>
<table width="100%" border="0" cellspacing="0" cellpadding="5">
<tr align="center">
<td><img src="http://www.m-ul.com/email/images/accessories/travelagent-header.jpg" width="525" height="80"></td>
</tr>
<tr align="center">
<td>
<table cellspacing="0" cellpadding="0" border="0" align="center" width="601">
<tr>
<td width="18" height="159" valign="top"></td>
<td width="607" height="159" valign="top"><a href="http://www.m-ul.com/e/c.cgi?j=20011125_62&e=1736394363&r=u424&d=6&p=1"><img src="
<td width="4" height="159" valign="top"></td>
</tr>
<tr>
<td height="2" colspan="3" valign="top">
```

```
## Test out the email reader for the html path.
email_path_plain = '/content/drive/My Drive/dsc650/dsc650/assignments/assignment04/examples/plain/message-0.txt'
print(read_raw_email(email_path_plain))
```

```
Message-ID: <14562086.1075853934677.JavaMail.evans@thyme>
Date: Mon, 24 Jan 2000 01:37:00 -0800 (PST)
From: office.chairman@enron.com
To: all.downtown@enron.com
Subject: Over $50 -- You made it happen!
Mime-Version: 1.0
Content-Type: text/plain; charset=ANSI_X3.4-1968
Content-Transfer-Encoding: quoted-printable
X-From: Office of the Chairman
X-To: All Enron Downtown
X-cc:
X-bcc:
X-Folder: \Dana_Davis_Dec2000\Notes Folders\Discussion threads
X-Origin: Davis-D
X-FileName: ddavis2.nsf

On Wall Street, people are talking about Enron.  At Enron, we're talking=20
about people...our people.  You are the driving force behind every success=
=20
that our company has experienced, including our high-performing stock price=
,=20
which surpassed the $50 mark only a few days ago. You made it happen!

To show our appreciation for your hard work and commitment to Enron=01,s=20
continued success, eligible regular full-time and regular part-time employe=
es=20
who were on the payroll of a wholly-owned Enron company at Dec. 31, 1999 wi=
ll=20
receive 50 Enron stock options.  A special stock option award certificate a=
nd=20
a copy of the stock plan will be sent to you in the next few weeks.

About the stock options grant:
The grant was effective Jan. 18, 2000, therefore, the option grant price is=
=20
$55.50.
Options will have a seven-year term, which means you must exercise these=20
options before seven years have passed.
Options will vest 25 percent on Feb. 28, 2000 and 25 percent each subsequen=
t=20
January 18th thereafter until fully vested, as long as you are an employee =
of=20
Enron.  This means you may exercise 25 percent of these options as early as=
=20
Feb. 28.
This grant is made in accordance with the terms and provisions of the Enron=
=20
Corp. Stock Plans and the award documents, which you will receive in the=20
coming weeks.=20

Due to varying international restrictions and legalities, Enron employees i=
n=20
certain international locations will receive some other form of recognition=
. =20
Your local management and human resources representative will communicate=
=20
further details.
```

```
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 = {}
    msg = Parser(policy=default).parsestr(original_msg)
    ## Use Python's email library to read the payload and the headers
    ## https://docs.python.org/3/library/email.examples.html
    result['payload'] = msg.get_payload()
    result['text'] = parse_html_payload(result['payload'])

    ## Use error handling to setup key,value within msg.items()
    try:
        for key, value in msg.items():
            result[key.replace('-', '_')] = value
    except Exception as e:
        print(f"Problem parsing email: email_path e")

    ## use error handling to parse date and convert date with isoformat().
    try:
        result["Date"] = parse(result["Date"], ignoretz = False).isoformat()
    except Exception as e:
        print(f"Problem converting date: {result.get('date')} {e}")

    tuple_result = tuple([str(result.get(column, None)) for column in columns])
    return ParsedEmail(*tuple_result)
```

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

```
<ipython-input-10-6a57cbc2b755>:7: MarkupResemblesLocatorWarning: The input looks more like a filename than markup. You may want to oper
  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
```

Assignment 4.3

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

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


## Import bs4
import bs4


## 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', 'original_msg', parse_email_func('original_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
## Setup tokenizer and vectorizer to put into pipeline.
tokenizer = Tokenizer(inputCol ="text", outputCol="words")
```

```
vectorizer = CountVectorizer(inputCol = tokenizer.getOutputCol(), outputCol = "features", vocabSize = 3)

## Setup Stages to include tokenizer and vectorizer created above.
stages = [tokenizer, vectorizer]
email_pipeline = Pipeline(stages=stages)
    ## Complete code
new_df = parse_emails(df)

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


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

```
+------------------+--------------------+--------------------+
|                id|               words|            features|
+------------------+--------------------+--------------------+
|davis-d/2_trash/1_|[, >, , , , , >, ...|(3,[0,1,2],[697.0...|
|davis-d/2_trash/3_|[----------------...|(3,[0,1,2],[118.0...|
|davis-d/2_trash/4_|[-----original, m...|(3,[0,2],[17.0,1.0])|
|davis-d/2_trash/2_|[fyi..., thanks.,...|(3,[0,1,2],[57.0,...|
|2_trash/candis/12_|[----------------...|      (3,[0],[41.0])|
|2_trash/candis/10_|[hi, mommy!, , ye...|(3,[0,1,2],[4.0,1...|
|2_trash/candis/11_|[hey, sweetie,, ,...| (3,[0,1],[8.0,1.0])|
|2_trash/candis/13_|[----------------...|(3,[0,1,2],[65.0,...|
|2_trash/candis/14_|[----------------...|(3,[0,2],[58.0,1.0])|
|2_trash/candis/15_|[----------------...|(3,[0,1,2],[29.0,...|
|2_trash/candis/17_|[----------------...|      (3,[0],[43.0])|
|2_trash/candis/16_|[----------------...|(3,[0,1,2],[18.0,...|
|2_trash/candis/19_|[----------------...|      (3,[0],[14.0])|
| 2_trash/candis/1_|[are, you, on, th...| (3,[0,1],[1.0,1.0])|
|2_trash/candis/18_|[, , , , --------...|      (3,[0],[11.0])|
| 2_trash/candis/3_|[candis, all, you...|(3,[0,1,2],[16.0,...|
| 2_trash/candis/2_|[listen, girly!, ...| (3,[0,2],[9.0,1.0])|
| 2_trash/candis/4_|[what, is, your, ...| (3,[0,2],[1.0,1.0])|
| 2_trash/candis/5_|[candis, -, , why...|       (3,[0],[3.0])|
| 2_trash/candis/6_|[----------------...|      (3,[0],[16.0])|
+------------------+--------------------+--------------------+
only showing top 20 rows
```