

Assignment 4

September 27, 2021

```
[1]: 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
from chardet.universaldetector import UniversalDetector

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, StructField

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 = data_dir.joinpath('enron')

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)

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

```

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

Followed below steps: Copied directory from external/enron to /assignment04/data
jovyan@jupyter-adityasumbaraju:~/dsc650/dsc650/assignments/assignment04/data\$
cp -r /home/jovyan/dsc650/data/external/enron . jovyan@jupyter-adityasumbaraju:~/dsc650/dsc650/assignments/assignment04/data\$

```

[23]: #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()

```

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

0.1 Assignment 4.1

```

[7]: def read_raw_email(email_path):
    detector = UniversalDetector()

    try:
        with open(email_path, encoding="utf8", errors='ignore') 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()
            #print(original_msg)

    return original_msg

def make_spark_df():
    records = []

    for root, dirs, files in os.walk(enron_data_dir):
        for file_path in files:
            current_path = Path(root).joinpath(file_path)
            #print (current_path)
            username = ""
            if (os.path.dirname(os.path.dirname(current_path)).split('/')[2]
↳ == "enron"):
                username = os.path.dirname(os.path.dirname(current_path)).
↳split('/')[1]
            else:

```

```

        os.path.dirname(os.path.dirname(current_path)).split('/')[ -2]
    record = {}
    record['id'] = os.path.relpath(current_path, os.path.dirname(os.
→path.dirname(root)))
    record['username'] = username
    record['original_msg'] = read_raw_email(current_path)
    records.append(record)

    ## TODO: Complete the code to 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)
    # return spark.createDataFrame(records)

df = make_spark_df()

```

[8]: df.show()

```

+-----+-----+-----+
|          id|username|    original_msg|
+-----+-----+-----+
|may-1/notes_inbox...|  may-1|Message-ID: <3024...|
|may-1/notes_inbox...|  may-1|Message-ID: <7226...|
|may-1/notes_inbox...|  may-1|Message-ID: <2591...|
|may-1/notes_inbox...|  may-1|Message-ID: <3243...|
|may-1/notes_inbox...|  may-1|Message-ID: <1407...|
|may-1/notes_inbox...|  may-1|Message-ID: <2447...|
|may-1/notes_inbox...|  may-1|Message-ID: <9631...|
|may-1/notes_inbox...|  may-1|Message-ID: <1745...|
|may-1/notes_inbox/8_|  may-1|Message-ID: <2812...|
|may-1/notes_inbox...|  may-1|Message-ID: <3340...|
|may-1/notes_inbox...|  may-1|Message-ID: <6783...|
|may-1/notes_inbox...|  may-1|Message-ID: <2907...|
|may-1/notes_inbox...|  may-1|Message-ID: <2344...|
|may-1/notes_inbox...|  may-1|Message-ID: <5532...|
|may-1/notes_inbox...|  may-1|Message-ID: <2778...|
|may-1/notes_inbox...|  may-1|Message-ID: <2589...|
|may-1/notes_inbox...|  may-1|Message-ID: <1429...|
|may-1/notes_inbox...|  may-1|Message-ID: <9382...|
|may-1/notes_inbox...|  may-1|Message-ID: <1723...|
|may-1/notes_inbox...|  may-1|Message-ID: <1140...|
+-----+-----+-----+
only showing top 20 rows

```

```
[9]: df.printSchema()
```

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

0.2 Assignment 4.2

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

```
[10]: 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
↳contributing to the success that the program has enjoyed.

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"><a href="http://insync-online.p04.com/u.d?

↪AkReaQA5eczXE=11"></nobr></td>

</tr>

<tr valign="top">

<td width="4" bgcolor="#CCCCCC"></td>

<td width="20"></

↪td>

<td width="165">
<a href="http://insync-online.p04.com/u.d?

↪LkReaQA5eczXL=21">
</td>

<td width="20"></

↪td>

```

    <td width="165"><br><a href="http://insync-online.p04.com/u.d?
    ↪BkReaQA5eczX0=31"></a><br></td>
    <td width="20"></
    ↪td>
    <td width="165"><br><a href="http://insync-online.p04.com/u.d?
    ↪JkReaQA5eczXR=41"></a><br></td>
    <td width="19"></
    ↪td>
    <td width="4" bgcolor="#CCCCCC"></td>
</tr>
</table>
<table border="0" cellpadding="0" cellspacing="0" width="582">
<tr valign="top">
    <td width="4" bgcolor="#CCCCCC"></td>
    <td width="574"><br>
        <table border="0" cellpadding="0" cellspacing="0" width="574"
        ↪bgcolor="#99ccff">
            <tr>
                <td width="50"></td>
                <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

for just \$29.95 when you use Promo Code **OS41WAVE**. That's a
\$10 savings off the list price.

[Click here](http://insync-online.p04.com/u.d?NkReaQA5eczXRh=51)
to view a full product demo now

[!\[\]\(082f818d99f166a3ba574d9284d73064_img.jpg\)](http://insync-online.p04.com/u.d?MkReaQA5eczXRm=61)

You can do a lot more with your Palm's 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's 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

[!\[\]\(64f7c7e956682d89489e8b2ffcb346b7_img.jpg\)](http://insync-online.p04.com/u.d?WkReaQA5eczXRb=71)

- 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*

[!\[\]\(bd1e8030ddcf14aba902529f6d3d03af_img.jpg\)](http://insync-online.p04.com/u.d?VEReaQA5eczXRQ=81)

- Use your GSM compatible mobile phone or modem to get online and access the web
- Stay connected with email, instant messaging and text messaging to GSM mobile phones


```

        <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=91">Click here
        to
            upgrade to the new Palm 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>
        
        <br>
        </font></td>
        <td width="50"></td>
        </tr>
        </table></td>
        <td width="4" bgcolor="#CCCCCC"></td>
        </tr>
        <tr>
        <td colspan="3"></td>
        </tr>
    </table>
    <table border="0" cellpadding="0" cellspacing="0" width="582">
        <tr>
            <td width="54"></
            td>
            <td width="474"><font face="arial, verdana" size="-2" color="#000000"><br>

```

```

    * This feature is available on the Palm®; IIIx, Palm®; IIIxe, and
    ↪Palm®; Vx. <br><br>
    ** Note: To use the MIK functionality, you need either a Palm OS®;
    ↪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?
    ↪RkReaQA5eczXRK=101">Click here</a> for
    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®; 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>
    </td>
    <td width="54"></
    ↪td>
    </tr>
</table><br><br><br><br>
<!-- The following image is included for message detection -->

</body>
</html>

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

```

```
html_msg_example = html_msg_example.strip()
```

```
[23]: def parse_html_payload(payload):  
    """  
    This function uses BeautifulSoup to read HTML data  
    and return the text. If the payload is plain text, then  
    BeautifulSoup 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)  
    ## TODO: Use Python's email library to read the payload and the headers  
    if msg.is_multipart():  
        for payload in msg.get_payload():  
            body=payload.get_payload()  
        else:body=msg.get_payload()  
  
    result['From']=msg['From']  
    result['Subject']=msg['Subject']  
    result['id']=msg['Message-ID']  
    result['X-From']=msg['X-From']  
    result['payload']=msg.get_payload()  
    result['text'] = parse_html_payload(result['payload'])  
    ## https://docs.python.org/3/library/email.examples.html  
    tuple_result = tuple([str(result.get(column, None)) for column in columns])  
    return ParsedEmail(*tuple_result)
```

```
[24]: parsed_msg = parse_email(plain_msg_example)
```

```
[25]: 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 to contributing
to the success that the program has enjoyed.

Thanks and Best Regards,

Jeff Hammad

```
[26]: parsed_html_msg = parse_email(html_msg_example)
```

```
[27]: 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*

Use your GSM compatible mobile phone or modem to get online and access the web
Stay connected with email, instant messaging and text messaging to GSM mobile phones

Send applications or records through your cell phone to schedule meetings and even "beam"

important information to others

All this comes in a new operating system that can be yours for just \$29.95! Click here to

upgrade to the new Palm OS v4.1 and you'll also get the latest Palm desktop software. Or call

1-800-881-7256 to order via phone.

Sincerely,

The Palm Team

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 OS41WAVE.

* This feature is available on the Palm IIIx, Palm IIIxe, and Palm Vx.

** Note: To use the MIK functionality, you need either a Palm OS® compatible modem or a phone

with built-in 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. Click here for

a list of tested and supported phones that you can use with the MIK. Cable not provided.

To modify your profile or unsubscribe from Palm newsletters, click here.

Or, unsubscribe by replying to this message, with "unsubscribe" as the subject line of the message.

Copyright© 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.

0.3 Assignment 4.3

```
[34]: ## This creates a schema for the email data
email_struct = StructType()

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

[37]: ## 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

email_pipeline = ↵
    ↪ Pipeline(stages=[ParseEmailsTransformer(), Tokenizer(inputCol='original_msg', ↵
    ↪ outputCol='words'),
                                CountVectorizer(inputCol='words', ↵
    ↪ outputCol='features')])
    ## TODO: Complete code
)
model = email_pipeline.fit(df)
result = model.transform(df)

[38]: result.select('id', 'words', 'features').show()
```

id	words	features
may-1/notes_inbox...	[message-id:, <30... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <72... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <25... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <32... (167993,[0,1,2,6,...	
may-1/notes_inbox...	[message-id:, <14... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <24... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <96... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <17... (167993,[0,1,2,3,...	
may-1/notes_inbox/8_	[message-id:, <28... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <33... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <67... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <29... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <23... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <55... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <27... (167993,[0,7,8,13...	
may-1/notes_inbox...	[message-id:, <25... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <14... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <93... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <17... (167993,[0,1,2,3,...	
may-1/notes_inbox...	[message-id:, <11... (167993,[0,1,2,3,...	

only showing top 20 rows

[]: