MSc Web and Social Media Analytics

Tutorial Pack 10
(90 minutes)

(To be completed during LW10 tutorial)

LEARNING OBJECTIVES

- **To practice implementing a pre-processing chain in Python**
- **To prepare a raw social media dataset for analysis**

LEARNING OUTCOMES

- **By the end of this tutorial students will have;**
  - ○ **Implemented a pre-processing chain in Python.**
  - ○ **Built-up their understanding of how to read files in Python.**
  - ○ **Explored using the Python "with" statement.**

RESOUCRES AND TOOLS REQUIRED

- ○ **Python 3 via google collab or repl.it**
- ○ **Gravity film tweets**
- ○ **LW9 lecture notes covering the pre-processing pipeline**

IMPORTANT:

This pack is designed for you to go at your own speed. At the end of each section there is a series of practice questions and exercises. You should attempt to answer **all questions**.

Any questions you do not complete today should be completed before your next tutorial.

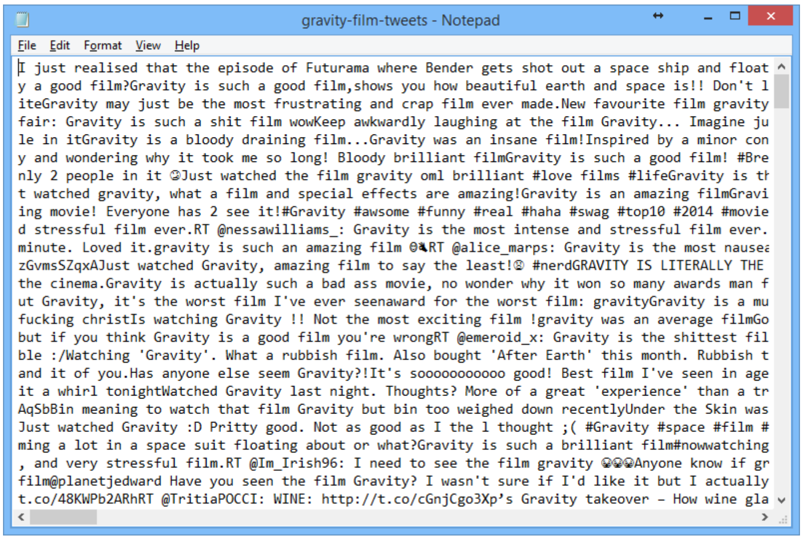
You may find this tutorial pack and the exercises useful preparation for the coursework.

1. Introduction

In last week's lecture we discussed the significance of the data preparation and pre-processing stages of model development. In today's session we are going to practice applying some of these techniques to a sample dataset collected following the release of the film Gravity.

2. Accessing the sample dataset

To complete this tutorial, you will need a copy of "gravity-film-tweets.txt". This can be found on the blackboard website under week 10.



```
I just realised that the episode of Futurama where Bender gets shot out a space ship and float
y a good film?Gravity is such a good film,shows you how beautiful earth and space is!! Don't l
iteGravity may just be the most frustrating and crap film ever made.New favourite film gravity
fair: Gravity is such a shit film wowKeep awkwardly laughing at the film Gravity... Imagine ju
le in itGravity is a bloody draining film...Gravity was an insane film!Inspired by a minor con
y and wondering why it took me so long! Bloody brilliant filmGravity is such a good film! #Bre
nly 2 people in it ☺Just watched the film gravity oml brilliant #love films #lifeGravity is th
t watched gravity, what a film and special effects are amazing!Gravity is an amazing filmGravi
ing movie! Everyone has 2 see it!#Gravity #awsome #funny #real #haha #swag #top10 #2014 #movie
d stressful film ever.RT @nessawilliams_: Gravity is the most intense and stressful film ever.
minute. Loved it.gravity is such an amazing film 😍RT @alice_marps: Gravity is the most nausea
zGvmsSZqxAJust watched Gravity, amazing film to say the least!☹ #nerdGRAVITY IS LITERALLY THE
the cinema.Gravity is actually such a bad ass movie, no wonder why it won so many awards man f
ut Gravity, it's the worst film I've ever seenaward for the worst film: gravityGravity is a mu
fucking christIs watching Gravity !! Not the most exciting film !gravity was an average filmGo
but if you think Gravity is a good film you're wrongRT @emeroid_x: Gravity is the shittest fil
ble :/Watching 'Gravity'. What a rubbish film. Also bought 'After Earth' this month. Rubbish t
and it of you.Has anyone else seem Gravity?!It's sooooooooooo good! Best film I've seen in age
it a whirl tonightWatched Gravity last night. Thoughts? More of a great 'experience' than a tr
AqSbBin meaning to watch that film Gravity but bin too weighed down recentlyUnder the Skin was
Just watched Gravity :D Pritty good. Not as good as I the l thought ;( #Gravity #space #film #
ming a lot in a space suit floating about or what?Gravity is such a brilliant film#nowwatching
, and very stressful film.RT @Im_Irish96: I need to see the film gravity 😂😂😂Anyone know if gr
film@planetjedward Have you seen the film Gravity? I wasn't sure if I'd like it but I actually
t.co/48KWPb2ARhRT @TritiaPOCCI: WINE: http://t.co/cGnjCgo3Xp's Gravity takeover – How wine gla
```

Figure 1 Sample dataset when viewed in notepad.

This file should then be uploaded to your Google Collab session. To do this you should create a new notebook (for example Week10.ipynb), switch to the **Files** view and then click on the option to "*Upload to session storage*"

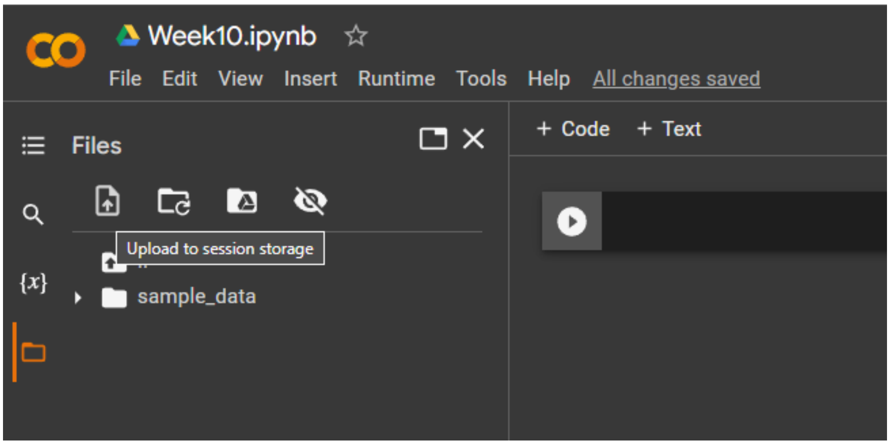


Figure 2 Uploading a file to the Google Collab session

Note that **any files** you upload will be **removed** when the Google Collab session is terminated. For this reason you should always keep a local copy of any data you collect. Once the file has been selected and uploaded you should see it appear in the list of available files.

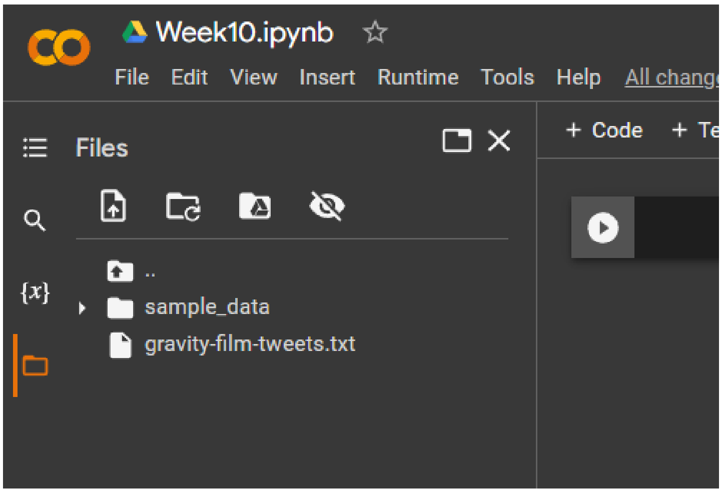
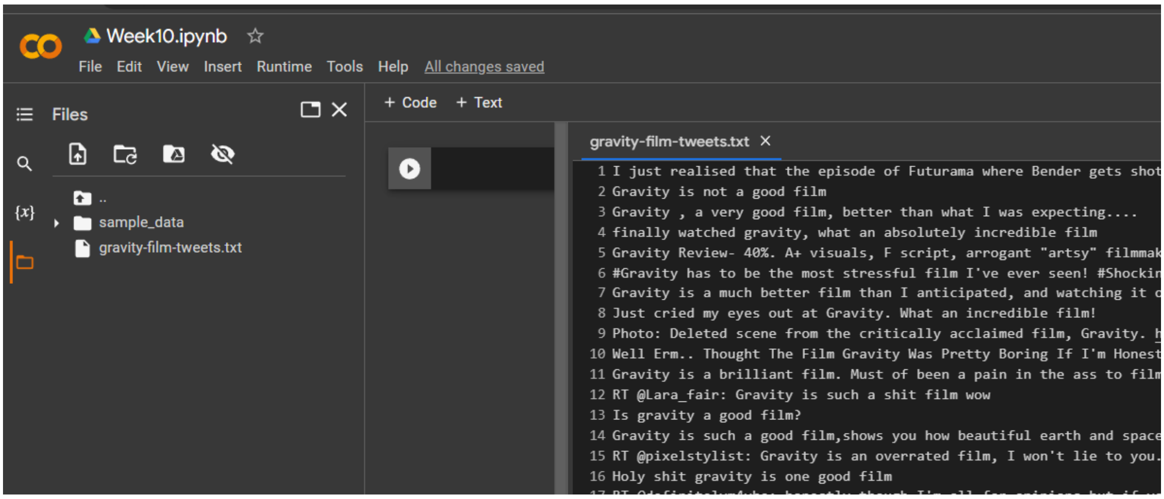


Figure 3 Screenshot of sample data once it has been successfully uploaded.

You can double click on the file to open the text viewer in Google Collab. The numbers on the left-hand side represent the line number of each collected tweet. By scrolling to the bottom of that file you should observe a total of 3607 lines – although this includes some blank lines and some unusually short tweets.



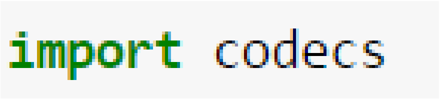
```
 1 I just realised that the episode of Futurama where Bender gets shot
 2 Gravity is not a good film
 3 Gravity , a very good film, better than what I was expecting....
 4 finally watched gravity, what an absolutely incredible film
 5 Gravity Review- 40%. A+ visuals, F script, arrogant "artsy" filmmak
 6 #Gravity has to be the most stressful film I've ever seen! #Shockin
 7 Gravity is a much better film than I anticipated, and watching it c
 8 Just cried my eyes out at Gravity. What an incredible film!
 9 Photo: Deleted scene from the critically acclaimed film, Gravity. h
10 Well Erm.. Thought The Film Gravity Was Pretty Boring If I'm Honest
11 Gravity is a brilliant film. Must of been a pain in the ass to film
12 RT @Lara_fair: Gravity is such a shit film wow
13 Is gravity a good film?
14 Gravity is such a good film,shows you how beautiful earth and space
15 RT @pixelstylist: Gravity is an overrated film, I won't lie to you.
16 Holy shit gravity is one good film
```

Figure 4 Previewing the sample data file in Google Collab.

3. Opening the dataset

To be able to access the contents of this file from Python we need to understand the basics of reading and writing to files. In the context of programming this is known as FILE IO (or File Input/Output).

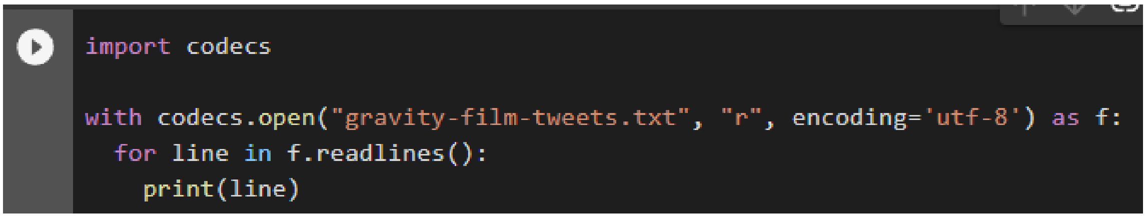
Python functions for reading and writing to text files can be found in the **codecs** package. We therefore need to import the codecs package into our program.



```
import codecs
```

To open the data we use the **open** function from the codecs package.

1. The first parameter is the name of the file we would like to open.
2. The second parameter is the file mode, in this case we want to open it for reading so we specify “r” (“w” is used for writing and “a” is used for appending).
3. Finally, the third parameter allows us to specify the encoding of the file. In this case it contains UTF-8[1] or unicode data.



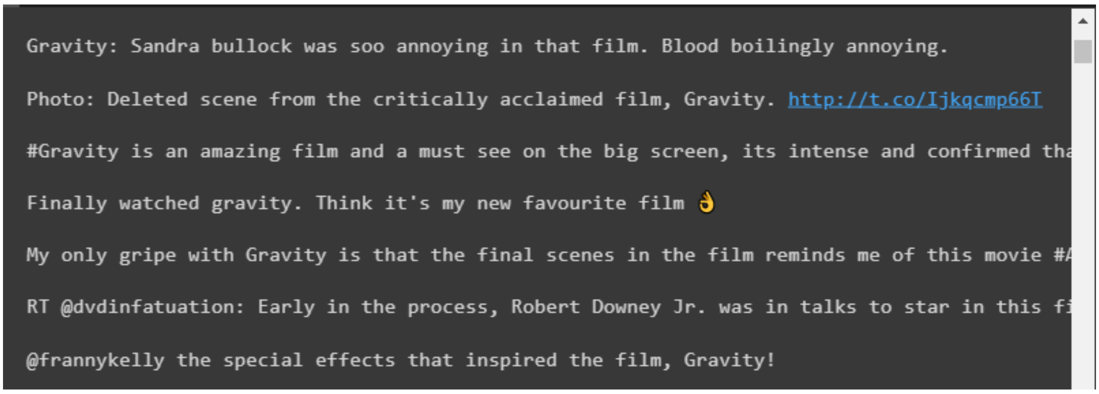
```
import codecs

with codecs.open("gravity-film-tweets.txt", "r", encoding='utf-8') as f:
  for line in f.readlines():
    print(line)
```

Figure 5 Code used to read in each line in the sample dataset.

You will notice that I have combined the open function and the **with** statement. Its purpose here is to tell Python that anything inside the current code block relies on the file being opened sucessfully and assigned into the variable f. It will also auto close the file at the end of the current block.

The next line says that I want to go through each line in the file (obtained by calling the **readlines**() method of the file represented by the variable f) and assign its contents into the variable line and then print it to the console.

When the code is run you should see output simialr to the following.



```
Gravity: Sandra bullock was soo annoying in that film. Blood boilingly annoying.

Photo: Deleted scene from the critically acclaimed film, Gravity. http://t.co/Ijkqcmp66T

#Gravity is an amazing film and a must see on the big screen, its intense and confirmed tha

Finally watched gravity. Think it's my new favourite film 👌

My only gripe with Gravity is that the final scenes in the film reminds me of this movie #A

RT @dvdinfatuation: Early in the process, Robert Downey Jr. was in talks to star in this fi

@frannykelly the special effects that inspired the film, Gravity!
```

Figure 6 Example output

[1] For more information about character encodings you can see https://en.wikipedia.org/wiki/UTF-8

Questions

Create a new notebook on Google Collab called **Week 10**. Follow the steps shown previously to upload a copy of the Gravity dataset to your session. Using the code shown in **Figure 5**, test that you are able to read the dataset from within Python by printing out each tweet on its own line.

You should then attempt the following questions. You may find it useful to refer to your lecture notes for Learning Week 9.

Q1 *Modify the code used to print out each tweet in the sample data file such that only tweets with a length of 10 or more characters are displayed.*

Q2 *Add a new cell to your notebook*. In this cell create a function that is called "*common_case*". This function should accept a single string parameter as an input and return the lowercase equivilent of the supplied parameter.

Q3 *Add a new cell to your notebook.* In this cell create a function called "without_leading_trailing_whitespace". This function should accept a single string parameter as an input and return a string without any leading (or trailing) whitespace characters. (Hint: consider using the *strip*() method that is part of the string class).

Q4 *Add a new cell to your notebook.* In this cell create a function called "no_multi_punctuation". This function should accept a single string parameter as an input and return a string where cases of multiple exclaimation marks (!!!!) or question marks (???) are replaced with a single exclaimation or question mark. (Hint: a regular expression can be used here).

Q5 *Add a new cell to your notebook.* In this cell create a function called "no_retweets". This function should accept a single string parameter as an input and return a string without any words that start with a retweet symbol "@" (Hint: consider using the **startswith**() method that is part of the string class).

Q6 *Add a new cell to your notebook.* In this cell create a function called "no_http_links". This function should accept a single string parameter as an input and return a string without any words that start with either https or http (Hint: consider using the **startswith**() method that is part of the string class).

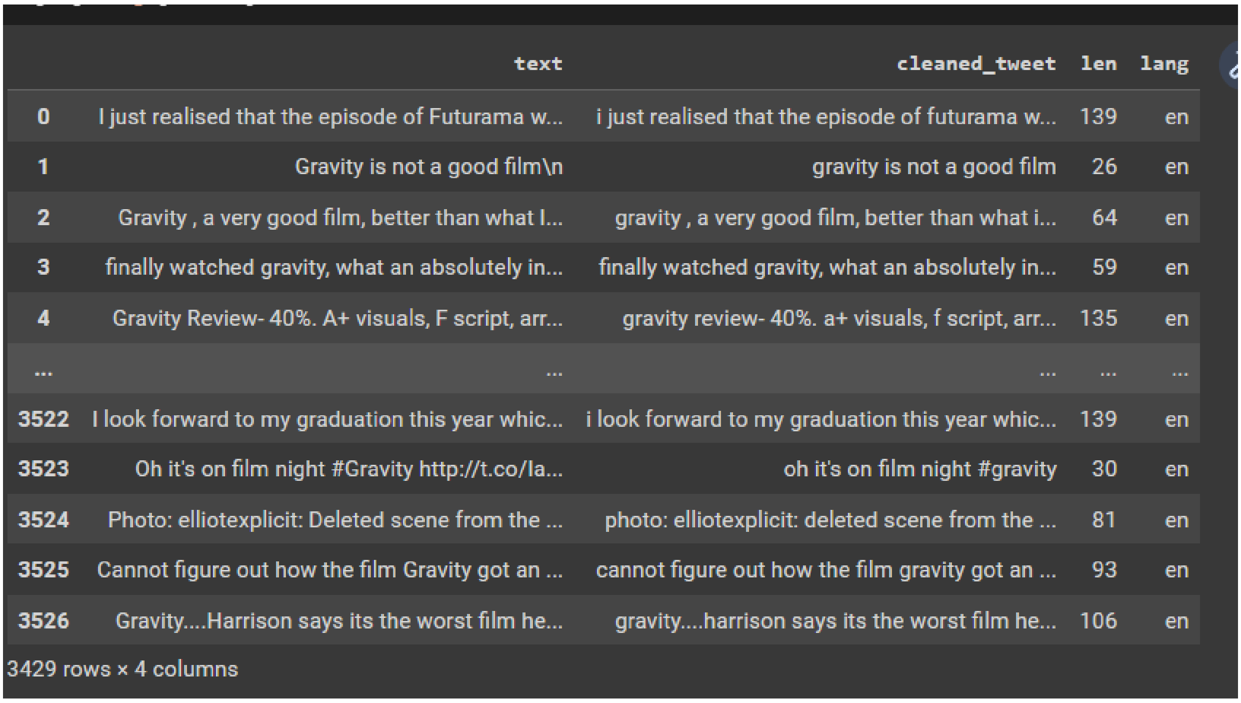
Q7 *Combine the steps from Q2-Q6 into a single function called "preprocessing_pipeline". This function should accept a string parameter and return a pre-processed string as the result.*

Q8 Modify the code used to read in the sample data so that the tweets are read into a pandas dataframe. Apply the pre-processing pipeline developed in Q7 so that a new column is added to your dataframe called "cleaned_tweets" containing the processed results.

Q9 Add a column to your dataframe called “len”. Populate this field by using a suitable Python function to count the length of each “cleaned_tweet”. Filter you dataframe so that only tweets with a length of 10 or greater are selected.

Q10 Add a column to your dataframe called “lang”. Populate this field by using a suitable Python function to detect the language of each tweet. Filter you dataframe so that only tweets in English are selected.

Your resulting dataframe should resemble the following:



	text	cleaned_tweet	len	lang
0	I just realised that the episode of Futurama w...	i just realised that the episode of futurama w...	139	en
1	Gravity is not a good film\n	gravity is not a good film	26	en
2	Gravity , a very good film, better than what I...	gravity , a very good film, better than what i...	64	en
3	finally watched gravity, what an absolutely in...	finally watched gravity, what an absolutely in...	59	en
4	Gravity Review- 40%. A+ visuals, F script, arr...	gravity review- 40%. a+ visuals, f script, arr...	135	en
...	...	...	...	...
3522	I look forward to my graduation this year whic...	i look forward to my graduation this year whic...	139	en
3523	Oh it's on film night #Gravity http://t.co/la...	oh it's on film night #gravity	30	en
3524	Photo: elliotexplicit: Deleted scene from the ...	photo: elliotexplicit: deleted scene from the ...	81	en
3525	Cannot figure out how the film Gravity got an ...	cannot figure out how the film gravity got an ...	93	en
3526	Gravity....Harrison says its the worst film he...	gravity....harrison says its the worst film he...	106	en

3429 rows × 4 columns