Tutorial Pack 11
(90 minutes)

(To be completed during LW11 tutorial)

LEARNING OBJECTIVES

- **To perform exploratory analysis on a sample dataset from Reddit Counter class**
- **To practice applying the LDA algorithm to group comments into themes**

LEARNING OUTCOMES

- **By the end of this tutorial students will have;**
  - **Read in a sample dataset in CSV format into a DataFrame.**
  - **Carried out some initial exploratory analysis.**
  - **Utilised the plotting functionality present in Pandas to create line graphs and column charts.**
  - **Conducted a word frequency analysis using the Counter class.**
  - **Visualised a dataset using a WordCloud representation.**
  - **Carried out filtering and grouping of data within a DataFrame.**
  - **Applied the LDA algorithm to identify abstract topics of discussion.**

RESOUCRES AND TOOLS REQUIRED

- **Lecture Slides for LW 10/ LW9**
- **Google Collab Notebook**

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.

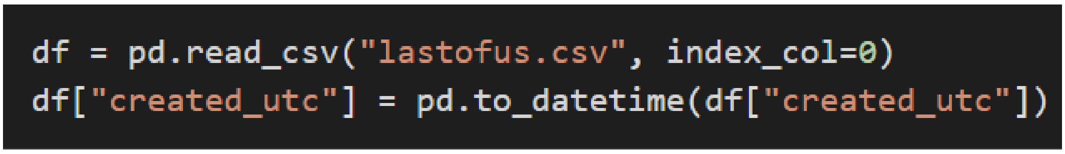
You may find this tutorial pack and the exercises useful preparation for the coursework.

1. Last of Us HBO Series

Today’s tutorial will provide an opportunity to practice the techniques related to data exploration and topic modelling covered in last week’s lecture. To begin you will need to download a copy of the sample dataset “lastofus.csv”. This file can be found in the Week 11 folder on Blackboard.

This dataset includes 480 comments posted to the Last of Us subreddit during early 2023 following the release of the HBO TV adaptation. The dataset is formatted as a CSV (comma separated values) file, it can be read in using the read_csv() function found in the Pandas package.

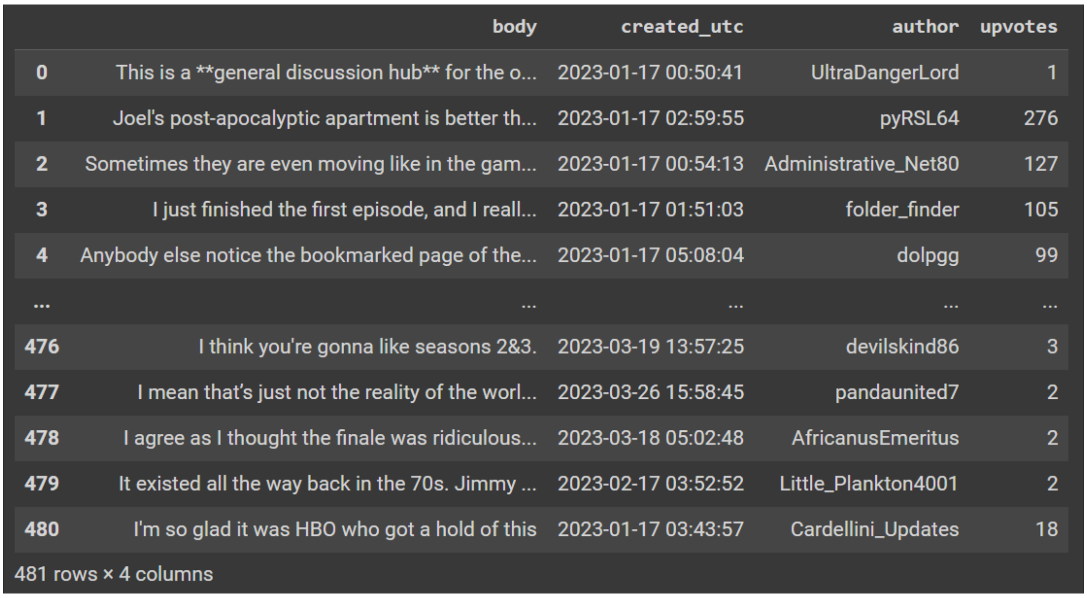
As the dataset contains a field containing the date each comment was posted, it needs to be manually converted to a datetime column when it is read in from a CSV file. This is because CSV files only contain data, and not information about the data type associated with each column.



```
df = pd.read_csv("lastofus.csv", index_col=0)
df["created_utc"] = pd.to_datetime(df["created_utc"])
```

Figure 1 Reading in the sample Last of Us dataset using pandas

One the data has been read in; it should resemble the following.



	body	created_utc	author	upvotes
0	This is a **general discussion hub** for the o...	2023-01-17 00:50:41	UltraDangerLord	1
1	Joel's post-apocalyptic apartment is better th...	2023-01-17 02:59:55	pyRSL64	276
2	Sometimes they are even moving like in the gam...	2023-01-17 00:54:13	Administrative_Net80	127
3	I just finished the first episode, and I reall...	2023-01-17 01:51:03	folder_finder	105
4	Anybody else notice the bookmarked page of the...	2023-01-17 05:08:04	dolpgg	99
...	...	...	...	...
476	I think you're gonna like seasons 2&3.	2023-03-19 13:57:25	devilskind86	3
477	I mean that's just not the reality of the worl...	2023-03-26 15:58:45	pandaunited7	2
478	I agree as I thought the finale was ridiculous...	2023-03-18 05:02:48	AfricanusEmeritus	2
479	It existed all the way back in the 70s. Jimmy ...	2023-02-17 03:52:52	Little_Plankton4001	2
480	I'm so glad it was HBO who got a hold of this	2023-01-17 03:43:57	Cardellini_Updates	18

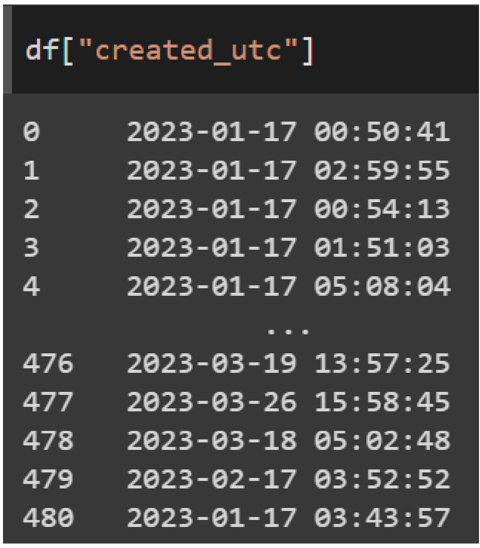
481 rows × 4 columns

Figure 2 The sample data once read into a pandas dataframe

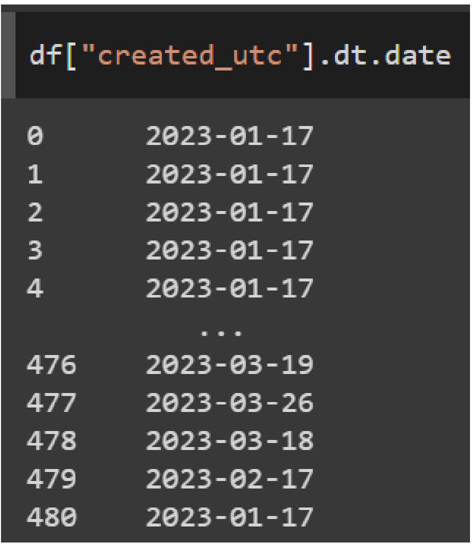
Additional Study Notes

Whilst the datetime object time provides a means to accurately store both dates and times, we are often only interested in the date part.

In this case we can access a the dt.date property of a datetime object as shown below.

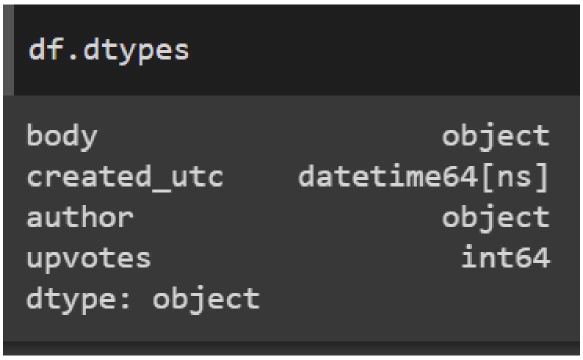


```
df["created_utc"]

0      2023-01-17 00:50:41
1      2023-01-17 02:59:55
2      2023-01-17 00:54:13
3      2023-01-17 01:51:03
4      2023-01-17 05:08:04
               ...
476    2023-03-19 13:57:25
477    2023-03-26 15:58:45
478    2023-03-18 05:02:48
479    2023-02-17 03:52:52
480    2023-01-17 03:43:57
```



```
df["created_utc"].dt.date

0       2023-01-17
1       2023-01-17
2       2023-01-17
3       2023-01-17
4       2023-01-17
           ...
476     2023-03-19
477     2023-03-26
478     2023-03-18
479     2023-02-17
480     2023-01-17
```

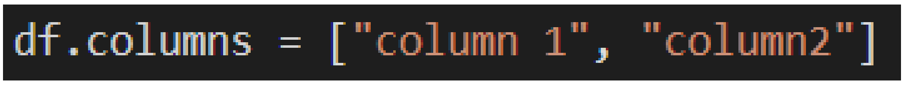
More generally we can check the internal datatype assigned to a dataframe column through its dtypes property. Note that the object data type in pandas is used to store strings.



```
df.dtypes

body                   object
created_utc    datetime64[ns]
author                 object
upvotes                 int64
dtype: object
```

Figure 3 Data types available in the sample dataset

The names of columns present in a dataframe can be accessed through the columns property. This property can also be overridden to rename a set of columns. In this case you can set the property equal to a list of new column names.



```
df.columns = ["column 1", "column2"]
```

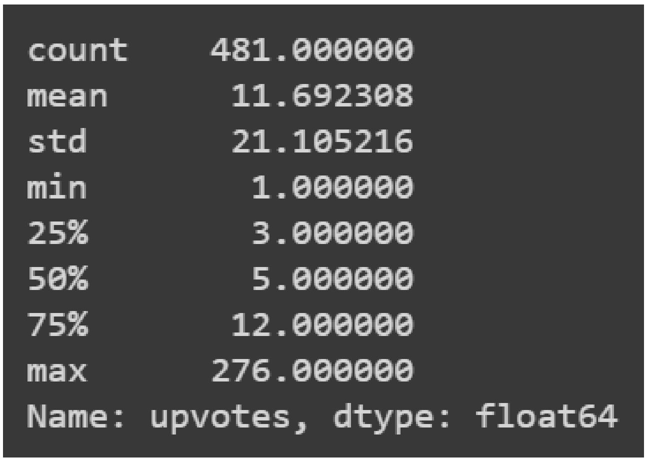
Figure 4 Renaming the names of columns in a dataframe

Questions

Create a new notebook on Google Collab called **Week 11** and upload a copy of the "lastofus.csv" data to your session.

You should then attempt the following questions. You may find it useful to refer to your lecture notes for Learning Weeks 10/9.

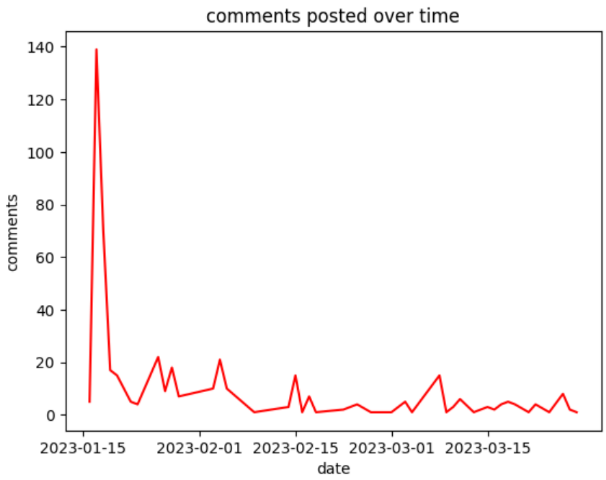
Q1 Using a suitable method, produce descriptive statistics for column containing the number of *upvotes* associated with each comment. Your output should resemble the screenshot below.



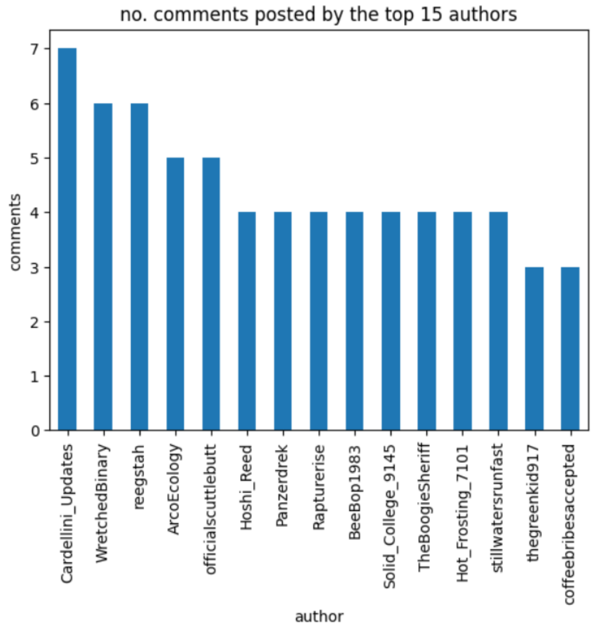
```
count    481.000000
mean      11.692308
std       21.105216
min        1.000000
25%        3.000000
50%        5.000000
75%       12.000000
max      276.000000
Name: upvotes, dtype: float64
```

Q2 Using a suitable Python statement, identify the date and time when the earliest comment in the dataset was posted.

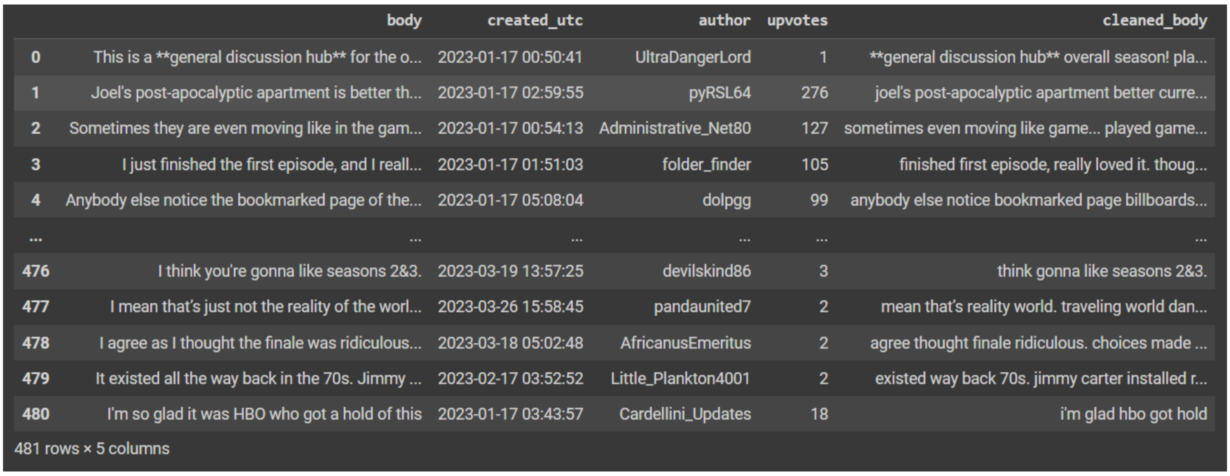
Q3 Using a suitable method, produce a chart of the number of comments posted each day. With appropraiate formatting, your output should resemble the figure below. Hint: the number of rows in a group can be counted using the Size() method.



Q4 Using a suitable method, create a vertical bar chart to show the number of comments posted by the top 15 most active authors. With appropraiate formatting, your output should resemble the figure below.



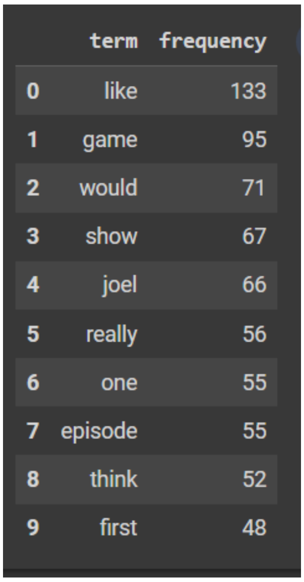
Q5 Apply **TWO** pre-processing steps to the "body" column and derive a new column called "cleaned_body". This column should contain text in a common-case and *exclude* English stopwords. Your dataframe should now resemble the screenshot below.



	body	created_utc	author	upvotes	cleaned_body
0	This is a **general discussion hub** for the o...	2023-01-17 00:50:41	UltraDangerLord	1	**general discussion hub** overall season! pla...
1	Joel's post-apocalyptic apartment is better th...	2023-01-17 02:59:55	pyRSL64	276	joel's post-apocalyptic apartment better curre...
2	Sometimes they are even moving like in the gam...	2023-01-17 00:54:13	Administrative_Net80	127	sometimes even moving like game... played game...
3	I just finished the first episode, and I reall...	2023-01-17 01:51:03	folder_finder	105	finished first episode, really loved it. thoug...
4	Anybody else notice the bookmarked page of the...	2023-01-17 05:08:04	dolpgg	99	anybody else notice bookmarked page billboards...
...	...	...	...	...	...
476	I think you're gonna like seasons 2&3.	2023-03-19 13:57:25	devilskind86	3	think gonna like seasons 2&3.
477	I mean that's just not the reality of the worl...	2023-03-26 15:58:45	pandaunited7	2	mean that's reality world. traveling world dan...
478	I agree as I thought the finale was ridiculous...	2023-03-18 05:02:48	AfricanusEmeritus	2	agree thought finale ridiculous. choices made ...
479	It existed all the way back in the 70s. Jimmy ...	2023-02-17 03:52:52	Little_Plankton4001	2	existed way back 70s. jimmy carter installed r...
480	I'm so glad it was HBO who got a hold of this	2023-01-17 03:43:57	Cardellini_Updates	18	i'm glad hbo got hold

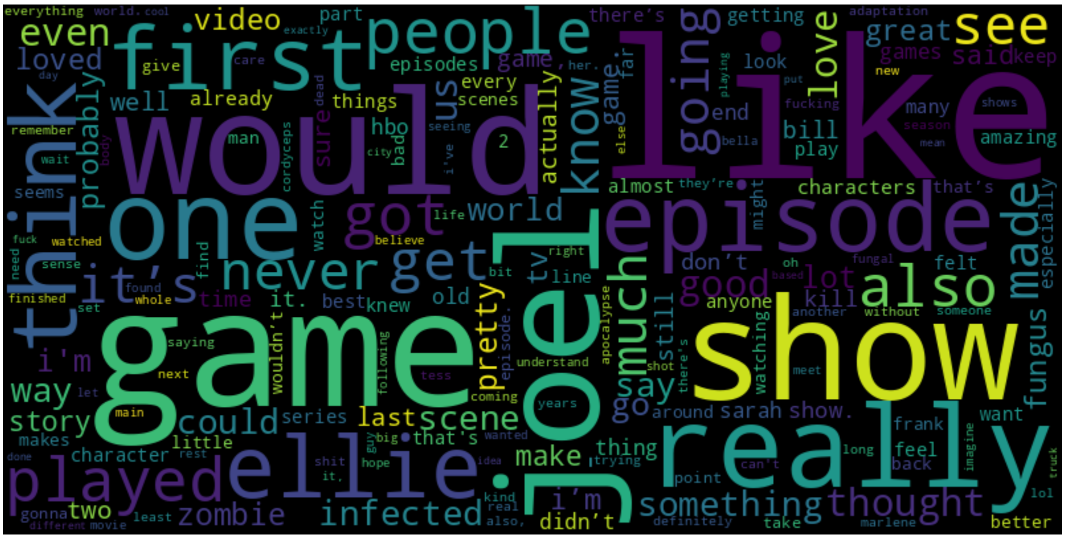
481 rows × 5 columns

Q6 Carry out a term-frequency analysis to identify the top 10 most commonly used words. Hint: you should use the text stored in the newly created “cleaned_body” column. Your output should resemble the table below.



	term	frequency
0	like	133
1	game	95
2	would	71
3	show	67
4	joel	66
5	really	56
6	one	55
7	episode	55
8	think	52
9	first	48

Q7 Using a suitable method, create a wordcloud to represent the top 200 most frequently used words. Hint: you should use the word counter created in the previous step. Your output should resemble the table below.



Q8 “joel” and “ellie” are the names of the two key chracters in the TV series. Using a suitable Python statement, determine the number of individual comments where the term “ellie” is mentioned. Hint: use a suitable filter on the dataframe and access the size property of the filtered result.

Q9 Apply an LDA topic model to your dataset. Experiment with 3,5 and 10 topics. Comment on and provide a logical interpreation of the results. Which model do you think best represents the discussions taking place?