LAB 01 – Web Science
Week starting on 13/01/25

Please work on these problems and, if needed, ask for help; if you cannot complete it today, please continue working on it.

No solutions are given for the lab problems.
Students should solve them with the help of lab tutors.

Lab materials are given in the Teams, File area. Look for the folders/subfolders for the week!
Class Materials – Labs – lab130125 …

I have given a set of tweets, and it is in the Teams File area (***lab130125/data1***).

1. Examine user objects & their keys; print each key-value pair
   ii) Explore the User Object for each of the Tweets
   iii) Identify useful keys

ToDo
- Read the file in json format
- Print key value pairs
- Look at the User Object
- Print the elements and observe (key-value pairs)
- Identify a list of hashtags, user mentions used
- Count how many tweets with extended text (that is more than 140 characters in tweet text) – use pandas data frame
- Identify and count retweets
- Observe the difference in User Objects of retweets

2. A file with links to a set of json files is given, and it is in the Teams File area (***lab130125/data2***).
   a. Download the data in **json** format, save the data on a file
   b. Summarise the data – statistics of the data, how many posts, how many comments etc.?
   c. Study the data structure (key-value pairs)

ToDo
- Read the file
- Extract link by link
- Write code to download each file
- Store them in a panadas data frame
- Summarise them
- Print out and see – key-value pairs

3. You will find some access restrictions if you just try to download the ***json*** file
   a. We will discuss how to fix this later
   b. I have provided another file called ***reddit_data.json***
   c. Write code to process

d. Store them in a panadas data frame
e. Summarise them
f. Print out and see – key-value pairs

4. Content Processing (This is based on Lecture 2 – only a portion is covered so far)
   Look at the Tweets text field and create vector representation of each of the tweet text.
   [Hint – read the text field; do the tokenisation; remove stop words; assign weights and normalise them]