Scientific Programming with Python - Exercise 2

Twitter tweets can hold a mountain of information, however sifting through the texts can be an ordeal. For this exercise, you will work with a CSV file containing meta-data and texts of tweets regarding the topic of Bitcoin over the last few years. Your task is to write a program that creates a data set from the contents of the tweets CSV that holds for the following data for each month:

- The most used 'hashtag' (#<tag>) of the month (not including #bitcoin, #bitcoins or #btc with any capitalization – e.g. #BTC, #BITcoin, …)
- The most mentioned username (@<username>) of the month
- The most referenced website (http or https) of the month

The data set will be saved in a CSV file with the following headers: Month,Hashtag,Mention,Website

Download the CSV file (zipped) here:

https://www.dropbox.com/s/motecidh0d80xri/tweets.zip?dl=0

Notes and Clarifications:

- The source CSV file will be named "tweets.csv".
- The output CSV file will be named "tweet-data.csv".
- The source file will be in **utf-8** encoding and must be saved in the same encoding.
- Each row in the output CSV will be one month (e.g. 2013-02, 2014-01, …).
- Rows will be ordered in ascending order by date.
- The hashtag and mentions will contain the leading symbol in the output file (i.e. # or @).
- A valid hashtag or mention will contain 1 or more alphanumeric (in any language), underscore (_) or dash (-) characters after the leading symbol (i.e. they must not contain any punctuation or whitespace other than dashes or underscores).
- The website will be the name of the website without the leading protocol (http or https) and without the trailing path (e.g. https://www.youtube.com/watch?v=oHg5SJYRHA0 will be counted as www.youtube.com when considering most referenced websites).
- In case of a tie, the first item in a sorted list of strings will be selected.
- In case there were no hastags/mentions/website references for a given month, the data set should hold the value 'None' for that cell.
- The tweets.csv file is approximately 1GB in size, so performance is an issue. Your code should run in under 2 minutes on a standard i5 PC, and in no case should it exceed a run time of 5 minutes.
- You are allowed use of any Python library learned in class, excluding the Pandas library.
    - For the functionality of 'Mode' selection **ONLY**, you may use the `pandas.Series.mode()` function.

Do not forget to write clean, readable code, and document your functions!