Take-home test for Unit 5

Intro

You are given a collection of text files, each containing 1000 recent tweets posted by several popular Twitter accounts. Each line in a file is one tweet, so if we read the files the usual way, each line will corresponds to a separate tweet:

```
with open('nytimes.tweet') as lines:
  for line in lines:
    print(line)
```

If you are getting a decoding error, to correctly open UTF-8 encoded text files, use a variant of the function `open`:

```
with open('nytimes.tweet', encoding='utf-8') as lines:
  for line in lines:
    print(line)
```

A username prefixed with `@` is called a **mention**. For example, the tweet below contains four mentions (`@voguemagazine`, `@FentyOfficial`, `@ethjgreen`, and `@yusefhairnyc`):

The above tweet is represented by the following string in the text file `rihanna.tweet`, any mention can be identified a word that starts with `@`:

```
So proud to be on another cover of @voguemagazine wearing my own designs from @FentyOfficial !!! On
stands October 16th! Photographer: @ethjgreen Fashion Editor: Tonne Goodman Hair: @yusefhairnyc Makeup:
Kanako Takase  https//vogue.cm/Tp75OAW  pic.twitter.com/LU9TEe1NEh
```

Task

We will write a program `test5.py` for **finding the most frequently mentioned usernames** in each of the provided files. To correctly identify mentions, we have to cleanup each tweet, keeping only letters, digits, and symbols `@` and `_`. After each tweet is cleaned up, we have to go through its words, and if the word starts with `@`, it is a mention.

Step-by-step:

1. Modify function `cleanedup` so that it keeps not only lowercase letters, but also digits `0123456789`, and symbols `@` and `_`
2. Write a new function `findMentions` that takes a filename as a parameter and reports the 5 usernames most frequently mentioned in that file. In order to do that, the function should create a dictionary of counts for all username mentions (words starting with `@`). After reading through the file and accumulating the counts for all mentioned usernames, use the dictionary to create a list like this:

   ```
   [[15, '@alice'], [20, '@bob'], [7, '@carol'], ... ]
   ```

   Use `sort` to sort the above list and print out 5 most frequently mentioned usernames.
3. Check each file in the current folder (using `os.listdir('.')`), if the file name ends with `.tweet`, print the file name and call `findMentions` on the file to find its most frequent mentions. The resulting output should have the same format as shown in the example below.

Example output

If you copy all provided `.tweet` files in the folder with your script, its output will look as follows (note that the exact order of the files might be different, depending on your operating system):

```
BillGates.tweet
     @trevornoah 7
     @rogerfederer 8
     @theeconomist 11
     @warrenbuffett 15
     @melindagates 18

ladygaga.tweet
     @markronson 9
     @realdonaldtrump 9
     @ahsfx 10
     @btwfoundation 11
     @applemusic 13

BarackObama.tweet
     @natlparkservice 3
     @flotus 4
     @ofa 5
     @vp 5
     @michelleobama 9

amyschumer.tweet
     @nbcsnl 11
     @everytown 12
     @bridgeteverett 14
     @rachelfeinstein 15
     @comedycentral 49

nytimes.tweet
     @jodikantor 2
     @mega2e 2
     @caityweaver 3
     @nytmag 5
     @nytparenting 5

doctorow.tweet
     @rgibli 2
     @xokasia 2
     @cbc 3
     @doctorow 3
     @sensanders 7

rihanna.tweet
     @eminem 12
     @fentyofficial 19
     @rihanna 21
     @savagexfenty 29
     @fentybeauty 48

Kaepernick7.tweet
     @yourrightscamp 16
     @kikifbaby 20
     @mikailsprice 26
     @darthkaepernick 28
     @kaepernick7 138
```

```
justinbieber.tweet
     @spotify 10
     @fallontonight 12
     @applemusic 15
     @theellenshow 15
     @skrillex 20

aoc.tweet
     @rokhanna 3
     @berniesanders 4
     @rashidatlaib 5
     @ayannapressley 6
     @ilhanmn 9

espn.tweet
     @dwyanewade 11
     @nfl 17
     @nba 21
     @thecheckdown 29
     @kingjames 32

ID_AA_Carmack.tweet
     @racketlang 2
     @beatsaber 3
     @boztank 3
     @joerogan 3
     @elonmusk 5
```

Last updated 2020-09-24 16:28:45 -0400