Laboratorio di Informatica 2018-19

Vi chiediamo di creare uno o più script Python secondo quanto specificato nei punti precedenti. Non vi chiediamo di consegnare lo script/gli script che creerete, tuttavia vi conviene salvare il vostro lavoro e conservarlo anche per le prossime lezioni esercitazioni, dato che riutilizzerete anche nelle esercitazioni successive gli script che oggi creerete.

Esercizio zero: File txt con tweets

Negli esercizi seguenti dovrete utilizzare un file txt contenente dei tweet scaricati dal social network Twitter (Tweets.zip). Ogni file che memorizza i tweet ha la seguente struttura: ogni riga del file contiene i dati di un tweet inviato da un utente, all'interno della riga le informazioni sono separate da 3 caratteri consecutivi /// Le informazioni si ripetono sempre nello stesso ordine. Qua sotto trovate un esempio (la parola TWEET è sempre presente all'inizio di una riga).

TWEET /// data del tweet /// username del twittatore /// numero di follower del twittatore /// testo

Ad esempio, in un ipotetico file "Italia.txt" potrà essere presente un tweet come segue:

```
TWEET|||Thu Dec 12 23:58:38 +0000 2013|||Luca30elode|||213|||"E tu che fai in Italia?"
```

Dove `TWEET` è una parola chiave che indica l'inizio di un nuovo tweet, scritto il 12 dicembre alle 23.58 da `Luca30elode` che ha 213 follower, seguito dal testo del tweet. Per maggiori dettagli sul funzionamento del social network Twitter e dei concetti collegati, si rimanda alla documentazione su twitter caricata sulla piattaforma di elearning.

Esercizio 1

Implementate una funzione in linguaggio python

tweetOrari(nomefile, listaParole)

che prenda in input 2 parametri formali:

- una stringa con il nome del file da analizzare (il file contenente i tweet);
- una lista contenente delle parole che devono essere presenti nei tweet analizzati (gli elementi della lista sono stringhe).

Esempio di invocazione: tweetOrari('tweet.txt', ['buon', 'anno']) #NB: […] è un parametro unico. La funzione deve restituire **un dizionario** che abbia come chiave la fascia oraria in cui i tweet sono stati twittati e come valore il numero di tweet che contengono tutte le parole presenti nella lista. Ad esempio dizionario[0]=123 indica che tra le 0.00 e le 0.59 sono stati twittati 123 tweet (contenenti le parole della lista passata). Mentre dizionario[23]=0 indica che tra le 23.00 e le 23.59 sono stati effettuati 0 tweet. Come listaParole può essere passata anche la lista vuota [], in questo caso dovranno essere contati tutti i tweet contenuti nella fascia oraria. Se in una fascia oraria non sono presenti tweet che soddisfano le condizioni, deve essere comunque inserita la chiave corrispondente alla fascia oraria associata alla numerosità 0, es. dizionario[23]=0

Nota Bene: le possibili chiavi del dizionario sono valori interi compresi tra 0 e 23, estremi inclusi.

Esercizio 2

Implementate una funzione in linguaggio python

calcolaFollower(nomefile, daescludere)

che prenda in input 2 parametri formali:

- una stringa con il nome del file da analizzare (il file contenente i tweet);
- il numero dei primi tweet del file da escludere.

La funzione deve restituire **un dizionario** che abbia come chiave il nome utente del twittatore, e come valore un numero che ne descrive l'impatto, calcolato come la somma del numero di

follower dei tweet che l'utente ha twittato. Ad esempio, se nel vostro file l'utente "Andrea" ha inviato 3 tweet ogni volta con 100 follower allora diz["Andrea"] = 300. Il numero di follower di un utente può cambiare (aumentare/diminuire) da un tweet all'altro, quindi per calcolare il risultato vanno sommati il numero di follower di ogni singolo tweet dell'utente. Dai tweet processati vanno esclusi i primi tweet del file, es. se la funzione viene invocata così: calcolaFollower('tweet.txt', 5) i primi 5 tweet del file non devono essere considerati per il calcolo del risultato. Al parametro formale da escludere può essere anche passato il valore 0, in questo caso significa che tutti i tweet del file devono essere processati.

Esercizio 3

Implementate una funzione in linguaggio python

contaMenzioni(nomefile, blackList)

che prenda in input 2 parametri formali:

- una stringa con il nome del file da analizzare (il file contenente i tweet);
- una lista contenente gli username degli utenti twittatori i cui tweet inviati non devono essere processati (gli elementi della lista sono stringhe).

Alcuni tweet contengono citazioni ad altri utenti (sono riconoscibili perché è presente l'username dell'utente citato preceduto dal carattere @). La funzione deve restituire **un dizionario** che abbia come chiave il nome utente della persona menzionata (l'@ non deve essere incluso) e come valore il numero di volte che è stata menzionata nei tweet analizzati. Esempio, considerando un file di soli 3 tweet:

```
TWEET|||Thu Dec 12 03:58:38 +0000 2013|||UtenteA|||213|||"@UtenteB: Andiamo al cinema?"
TWEET|||Thu Dec 12 07:08:12 +0000 2013|||UtenteB|||213|||"RT @UtenteA: Andiamo al cinema?"
TWEET|||Thu Dec 12 12:03:03 +0000 2013|||UtenteA|||213|||"@UtenteB: Quindi andiamo oppure no?"
```

Poiché ci sono 3 tweet che menzionano (cioè contengono *nel testo del tweet* il carattere "@" seguito dal nome utente) allora diz["UtenteA"] = 1 e diz["UtenteB"] = 2.

Dai tweet processati devono essere esclusi i tweet inviati dagli utenti i cui username sono presenti nel parametro blackList. Nell'esempio precedente, se viene invocata la funzione contaMenzioni('tweet.txt', ['UtenteB','UtenteC']), non deve essere processato il secondo tweet (quello cioè inviato da UtenteB). Se lo user da escludere appare nel testo di un tweet non è un problema, purchè il tweet sia stato inviato da un utente assente dalla blackList (nell'esempio precedente, il 1° ed il 3° tweet devono essere processati).

Si suggerisce:

- l'uso di find("@") per identificare l'inizio del nome utente nel testo e la slicing per estrarlo.
- di eliminare dal testo la punteggiatura che da fastidio o di sostituirla con uno spazio (utilizzando le funzioni di ricerca o sostituzione all'interno di una stringa)
- Si ricorda che uno username non può essere composto da segni di punteggiatura

Esercizio 4

Implementate una funzione in linguaggio python

contaReTweetInviati(nomefile, blackList)

che prenda in input 2 parametri formali:

- una stringa con il nome del file da analizzare (il file contenente i tweet);
- una lista contenente gli username di utenti i cui tweet inviati non devono essere processati (gli elementi della lista sono stringhe).

Può accadere che un tweet inviato da @UtenteA sia semplicemente la riproposizione di un tweet scritto da un altro utente. In tal caso il tweet di @UtenteA si chiama re-tweet ed è identificato dalla sigla "RT" all'inizio del tweet seguito da uno spazio, da un @ e dal nome dell'utente retwittato. Nell'esempio seguente sono riportati 3 retweet fatti dall'UtenteA:

```
TWEET|||Thu Dec 12 03:58:38 +0000 2013|||UtenteA|||213|||"RT @UtenteB: Ho preso 30 e lode!"
  TWEET|||Thu Dec 12 07:08:12 +0000 2013|||UtenteA|||213|||"RT @UtenteC: E' quasi natale"
    TWEET|||Thu Dec 12 12:03:03 +0000 2013|||UtenteA|||213|||"RT @UtenteD: Oggi nevica!"
```

La funzione deve contare il numero di ReTweet effettuati dagli utenti, restituendo **un dizionario** che abbia come chiave il nome utente e come valore il numero di tweet che l'utente ha re-twittato (l'@ non deve essere incluso nella chiave del dizionario).

Ad esempio, se @UtenteA ha re-twittato 3 tweet, allora nel dizionario restituito si dovrebbe avere diz["UtenteA"] = 3 (l'@ non deve essere incluso nella chiave del dizionario). Dai tweet processati devono essere esclusi i tweet inviati dagli utenti i cui username sono presenti nel parametro blackList. Nell'esempio precedente, se viene invocata la funzione contaReTweetInviati('tweet.txt',['UtenteA']) nessuno dei 3 tweet deve essere considerato. Il tweet deve essere processato se lo user da escludere appare nel testo, purchè il tweet sia stato inviato da un utente assente dalla blackList. Nell'esempio precedente, se viene invocata la funzione contaReTweetInviati('tweet.txt', ['UtenteB', 'UtenteC'] tutti e 3 i tweet devono essere processati.

Esercizio 5

Implementate una funzione in linguaggio python

classificaReTweet(nomefile, blackList)

che prenda in input 2 parametri formali:

- una stringa con il nome del file da analizzare (il file contenente i tweet);
- una lista contenente gli username degli utenti i cui tweet inviati non devono essere processati (gli elementi della lista sono stringhe).

Può accadere che un tweet inviato da @UtenteA sia semplicemente la riproposizione di un tweet scritto da un altro utente. In tal caso il tweet di @UtenteA si chiama re-tweet ed è identificato dalla sigla "RT" all'inizio del tweet seguito da uno spazio, da un @ e dal nome dell'utente retwittato. Nell'esempio seguente sono riportati 3 retweet fatti dall'UtenteA:

```
TWEET|||Thu Dec 12 03:58:38 +0000 2013|||UtenteA|||213|||"RT @UtenteB: Ho preso 30 e lode!"
  TWEET|||Thu Dec 12 07:08:12 +0000 2013|||UtenteA|||213|||"RT @UtenteC: E' quasi natale"
    TWEET|||Thu Dec 12 12:03:03 +0000 2013|||UtenteA|||213|||"RT @UtenteC: Oggi nevica!"
```

La funzione che dovete implementare deve contare per ogni utente, il numero di volte in cui i suoi messaggi sono stati retwittati. La funzione deve restituire **un dizionario** che abbia come chiave il nome utente e come valore il numero di volte che l'utente è stato tweet re-twittato. Per esempio, nel caso precedente si dovrebbe avere: diz["UtenteC"] = 2 (l'@ non deve essere incluso nella chiave del dizionario).

Dai tweet processati devono essere esclusi i tweet inviati dagli utenti i cui username sono presenti nel parametro blackList. Nell'esempio precedente, se viene invocata la funzione classificaReTweet('tweet.txt', ['UtenteA']) nessuno dei 3 tweet deve essere considerato. Se lo user da escludere appare nel testo di un messaggio o un suo messaggio è oggetto di retweet il tweet deve essere processato, purché il tweet sia stato inviato da un utente assente dalla

blackList. Nell'esempio precedente, se viene invocata la funzione classificaReTweet('tweet.txt', ['UtenteB', 'UtenteC']) tutti e 3 i tweet devono essere processati.

Esercizio 6
Implementate una funzione in linguaggio python

contaHashTag(nomefile)

che prenda in input, come parametro formale, il file contenente i tweet da analizzare.
Un tweet può contenere nel testo degli hashtag, un hashtag è una parola preceduta dal carattere # che viene utilizzata per etichettare il contenuto del tweet. Esempio, considerate il file dei tweet "Milan.txt" contenente i tweet che contengono la keyword "Milan" sia composto di soli 2 tweet:

```
TWEET|||Thu Dec 12 03:58:38 +0000 2013|||UtenteA|||213|||"#Napoli #Juventus fuori dalla Champions"
TWEET|||Thu Dec 12 04:18:31 +0000 2013|||UtenteA|||3|||"Però il #Napoli ha fatto una bella partita"
TWEET|||Thu Dec 12 12:03:12 +0000 2013|||Utentec|||433|||"#Milan ancora in corsa nel #campionato"
```

La funzione deve restituire un **dizionario** che abbia come chiave l'hashtag e come valore il numero di volte che l'hashtag appare nei tweet. Nell'esempio precedente, se si invoca la funzione diz=contaHashTag('tweet.txt'), si avrà: diz["Napoli"] = 2, diz["Juventus"] = 1, …
Dalla chiave del dizionario deve essere escluso il simbolo #
Si suggerisce:

- l'uso di find("#") per identificare l'inizio di un hashtag nel testo e la slicing per estrarlo
- di prestare attenzione agli hashtag che appaiono come ultima parola del tweet
- Si ricorda che la punteggiatura e gli spazi non fanno parte del nome di un hashtag.

Esercizio 7
Implementate una funzione in linguaggio python

esploraHashTag(nomefile, elencoHashTag)

che prenda in input 2 parametri formali:

- una stringa con il nome del file da analizzare (il file contenente i tweet);
- una lista contenente gli hashTag da analizzare (gli elementi della lista sono stringhe).

Un tweet può contenere nel testo degli hashtag, un hashtag è una parola preceduta dal carattere # che viene utilizzata per etichettare il contenuto del tweet. Esempio, considerate il file dei tweet "Milan.txt" contenente i tweet che contengono la keyword "Milan" sia composto di soli 2 tweet:

```
TWEET|||Thu Dec 12 03:58:38 +0000 2013|||UtenteA|||213|||"#Napoli #Juventus fuori dalla #Champions"
TWEET|||Thu Dec 12 04:18:31 +0000 2013|||UtenteA|||3|||"Però il #Napoli ha fatto una bella partita"
TWEET|||Thu Dec 12 12:03:12 +0000 2013|||Utentec|||433|||"#Milan ancora in corsa nel #campionato"
```

La funzione deve restituire un **dizionario** che abbia come chiave l'hashtag presente in elencoHashTag e come valore il numero di volte che l'hashtag appare nei tweet. Nell'esempio precedente, se si invoca la funzione diz=esploraHashTag('tweet.txt', ['campionato','Champions']),
si avrà diz["Champions"] = 1 e diz["Juventus"] = 1. Se un hashtag non è presente nei tweet, deve essere comunque inserito nel dizionario con numerosità 0. Dalle chiavi del dizionario devono essere esclusi i simboli #
Si suggerisce

- di prestare attenzione agli hashtag che appaiono come ultima parola del tweet
- di eliminare o sostituire con uno spazio la punteggiatura che da fastidio

- di prestare attenzione ad hashtag con parti in comune, es. se si cercano le occorrenze di #campionato non devono essere contate le occorrenze di #campionato2014 a meno che quest'ultimo non sia presente esplicitamente nella lista degli hashtag da cercare

Si ricorda che la punteggiatura e gli spazi non fanno parte del nome di un hashtag.