Python - Sezione 2

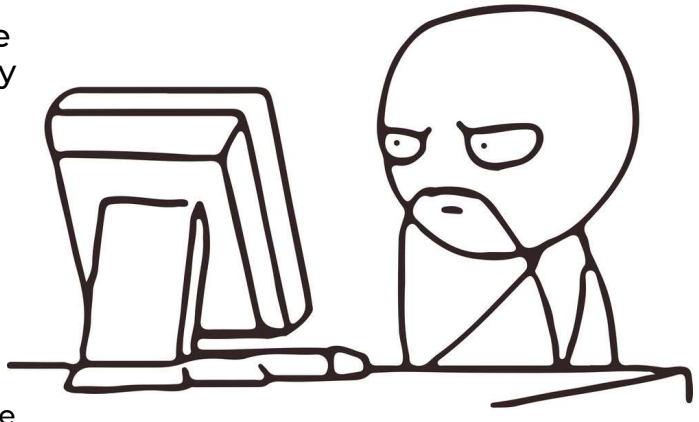
Lab riassuntivo 1

LINEE GUIDA:

- Scrivere il codice in un file Jupyter
- Consegnare Jupyter e txt esportati in un archivio .zip

1. Scaricare il *corpus* di testo fornito e convertirlo in oggetto lista, al cui interno ogni frase è una lista separata di parole (Occhio alle maiuscole!)
   - Contare numero di frasi, lunghezza totale del dataset in parole (=tutte) e lunghezza media di ogni frase (in parole)
2. Realizzare un «vocabulary» (o «lessico») contenente *una sola volta* le parole presenti nel corpus e un *conteggio di frequenza* delle parole nel vocabulary
   - Calcolare lunghezza del vocabulary e ritornare la/e parola/e con freq minima e freq massima, visualizzandone le rispettive frequenze
   - Ordinare le parole nel vocabulary in ordine alfabetico crescente e decrescente
3. Rimuovere dal vocabulary, in sequenza:
   - Le parole che appaiono meno di 2 volte (togliere freq < 2, mantenere freq=2)
   - Le parole che appaiono più di 100 volte (togliere freq > 100, mantenere freq=100)
   - Le parole presenti nel file '*stop words*' fornito (occhio alle maiuscole!)

   Dopo *ognuno* dei 3 step precedenti, ricalcolare lunghezza del vocabulary e ritornare la/e parola/e con freq minima e freq massima, visualizzandone le rispettive frequenze



4. Traduzione numerica (o "vettorizzazione")
   - I modelli di machine learning possono processare solo numeri, non stringhe. Definire quindi:
     - Una funzione *word2index()* che traduca ogni parola del corpus in un numero: il numero deve essere l'*indice* (posizione) con cui tale parola appare nel vocabulary. Definire un indice residuale da usare per le parole che non sono nel vocabulary.
     - Una funzione *index2word()* che ri-traduca il corpus *vettorizzato* nelle stringhe originali del vocabulary. Le parole che erano state tradotte con l'indice residuale vanno qui tradotte con la stringa 'OOV'.
   - Eseguire le 2 funzioni sul corpus. Esportare in formato txt il corpus tradotto in numeri e quello riconvertito in parole.