ALGORITMI E STRUTTURE DATI
STATISTICA PER I BIG DATA
APPELLO DEL 15 LUGLIO 2022

In questo esame faremo degli esperimenti con i Bloom filter. In particolare studiamo due data set, uno costituito da parole inglesi ed uno costituito da parole italiane. L'esperimento per ciascuno data set consiste nel determinare, per una serie di possibili taglie del Bloom filter, il numero di funzioni hash da utilizzare per minimizzare i falsi positivi.

I due data set sono contenuti nei file `inglesiPos.txt` contenente 45000 parole inglesi e `italianePos.txt` contenente 140000 parole italiane. In entrambi i file ogni linea di testo contiene una sola parola. Per valutare l'incidenza di falsi negativi usiamo il file `inglesiNeg.txt`, contenente circa 53000 parole inglesi non presenti in `inglesiPos.txt`, e il file `italianeNeg.txt`, contenente circa 140000 parole italiane non presenti in `italianePos.txt`. Per ciascun data set il vostro programma deve inserire le parole del file `Pos` in un Bloom filter di grandezza m volte il numero di parole per m in `range(2,8)`; calcolare la percentuale di falsi positivi del file `Neg` per valori di $2 \leq k \leq 8$; riportare per ogni m, il valore k che ottiene la percentuale minore di falsi positivi.

Istruzioni per la consegna. Il codice da sviluppare consiste di un programma python che utilizza la classe che si trova nel file `bloom.py`. La classe è simile a quella presentata a lezione.

Tutto il codice consegnato deve essere contenuto in un file con estensione `.py` che ha come nome il cognome dello studente scritto in minuscolo. Se il cognome contiene un apostrofo, uno spazio o un accento questi dovranno essere omessi.

La cartella che ha ricevuto contiene il pdf di questa traccia, i file `inglesiNeg.txt`, `inglesiPos.txt`, `italianePos.txt`, `italianeNeg.txt`, il file `bloom.py` e il file `result.txt` che contiene un possibile output del programma.