ALGORITMI E STRUTTURE DATI
STATISTICA PER I BIG DATA
APPELLO DEL 9 NOVEMBRE 2022

In questo esame faremo degli esperimenti con i Bloom filter. Come sappiamo i Bloom filter sono caratterizzati da due parametri: k, il numero di funzioni hash usate, ed m, l'espansione che determina la grandezza $N = n * m$ dell'array di bit utilizzato per un data set di grandezza n. Questi due parametri determinano la probabilità di un falso positivo.

Per effettuare i nostri esperimenti, consideriamo il data set di 90000 parole inglesi che troviamo nel file `englishWords.txt` e costruiamo un Bloom filter con k hash function e taglia $N = m*90000$. Ci interessa stampare la lista di tutti i falsi positivi che otteniamo effettuando una query per ciascuna delle parole che si trovano nel file `paroleItaliane.txt`.

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 ed ha i medesimi metodi. Non è efficiente dal punto di vista dell'allocazione di memoria ma è stata modificata per evitare problemi di compatibilità sulle diverse piattaforme. Il programma chiede interattivamente i valore di m e k e poi stampa la lista di falsi positivi.

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 `englishWords.txt`, `paroleItaliane.txt` e il file `bloom.py`.