ALGORITMI E STRUTTURE DATI
STATISTICA PER I BIG DATA
APPELLO DEL 5 SETTEMBRE 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.

Nel nostro studio, per ogni numero di hash function $3 \leq k \leq 13$ vogliamo stimare la minima espansione $m \in [2, 2.5, 3, 3.5, \ldots, 9.5]$ che garantisce una probabilità di falsi positivi $< .04$.

Per effettuare lo studio, consideriamo il data set di 90000 parole inglesi che si trovano nel file `englishWords.txt` e costruiamo un Bloom filter con k hash function e taglia $N = m * 90000$ per i valori di k e m di interesse per lo studio. Per stimare la probabilità di falsi positivi calcoliamo la percentuali di 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.

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`, il file `bloom.py` e il file `result.txt` che contiene un possibile output del programma.