NOME:
COGNOME:

Esercizio 1 (5 punti)
Dite se le seguenti asserzioni sono vere o false, e giustificate:
a)In un sistema di retrieval Booleano, lo stemming non abbassa mai la precisione
b) In un sistema di retrieval Booleano, lo stemming non abbassa mai la recall
c)Lo stemming aumenta la taglia del vocabolario
d)Lo stemming è utilizzato durante l'indicizzazione , ma non durante l'elaborazione di una query

Esercizio 2. (2 punti)
Quale è l'idf di un termine che occorre in tutti i documenti della collezione? Confrontate l'effetto di questo indice con l'uso di stop list.

Esercizio 3 (6 punti)
Il fattore di Dice è dato da:
$Dice(x,y) = \frac{2 | x | f | y |}{| x | + | y |}$ dove |s| rappresenta la "arity" (taglia) di un set s.
se |x| è il numero di documenti recuperati da un sistema di information retrieval e |y| è l'insieme dei documenti rilevanti nell'intera collezione, dimostrate che la *F-measure bilanciata* è equivalente a Dice(x,y). (Suggerimento: usate le espressioni di P e R in funzione dei "true positive" ecc.).

Esercizio 4 (9 punti)
Descrivete l'algoritmo HITS (con formule) e dimostratene la convergenza

Esercizio 5 (8 punti)
Cos'è il modulo e l'angolo del gradiente di un'immagine in un punto? Come si calcolano? A che servono?

R1:
a)Falso. Lo stemming aumenta il set dei documenti recuperati ma non il set dei documenti rilevanti
b)Vero. Dal momento che aumenta il set dei documenti rilevati, può aumentare la recall
c) Falso. Lo stemming diminuisce la taglia del vocabolario
d)Falso: si applica lo stesso processo a documenti e query.

R2:
E' zero. L'effetto è lo stesso di inserire la parola in una stop list: questa parola viene ignorata.

R3:

$$F = \frac{2P \cdot R}{P + R}$$

$$P = \frac{tp}{tp + fp} \quad R = \frac{tp}{tp + fn} \quad F = \frac{2tp}{2tp + fp + fn} \quad | \, x \models tp + fp \quad | \, y \models tp + fn$$

$$\Rightarrow Dice(x, y) = F$$

R4: vedi slides sull'argomento (Web IR, link analysis, algoritmo HITS, convergenza del power method)

R5: Data un'immagine *I* e un punto *P* di coordinate *P= (x,y)*, posso interpretare *I* come una funzione reale a due variabili e calcolarne le derivate parziali *Dx* e *Dy*. Per cui il gradiente di *I* in *P* è il vettore delle derivate parziali di *I* in *P*. Dato che *I* non è una funzione continua, le derivate parziali possono solo essere approssimate. Per calcolarle, si possono usare degli operatori locali (e.g., l'operatore di Sobel) che consistono nell'effettuare una serie di somme e differenze tra i valori dell'intensità dei pixel circostanti *P* con lo scopo, appunto, di approssimare il valore delle derivate in *P*.
Il vettore gradiente $G(x,y) = (Gx, Gy)^T$:

$$G(x,y) = \nabla I(x,y) = \begin{pmatrix} \frac{\partial I(x,y)}{\partial_x} \\ \frac{\partial I(x,y)}{\partial_y} \end{pmatrix}$$

può essere equivalentemente rappresentato utilizzando il suo modulo (*r*) e l'angolo (*phi*) rispetto all'asse *X* dell'immagine:

$$D(p) = \begin{pmatrix} r \\ \phi \end{pmatrix} = \begin{pmatrix} \sqrt{G_x^2 + G_y^2} \\ \arctan(\frac{G_y}{G_x}) \end{pmatrix}$$

D(P) può essere usato per individuare le discontinuità luminose in *I*. Infatti *D(P)* punta nella direzione di massima crescita di *I* in *P*. Per cui, se in *P I* cresce (o decresce) velocemente (ovvero, se in *P* è presente una discontinuità luminosa), allora *r(P)* avrà un valore elevato e *phi(P)* punterà nella direzione ortogonale al cambiamento di illuminazione.