Prova scritta di
METODI PER IL RITROVAMENTO DELL'INFORMAZIONE
C.d.L. in Informatica - A.A. 2020-21
Docenti: P. Basile, P. Lops - 1 Settembre 2021

Nome e Cognome : _____

Matricola : _____

1) Siano dati l'insieme delle categorie `C = {c1,c2}` e una collezione di documenti definiti sul vocabolario `V = {T1,T2,T3,T4,T5}`.

a) Costruire un classificatore *bayesiano* per C, addestrandolo sul seguente training set TR:

`TR = {<D1,c2>, <D2,c2>, <D3,c1>, <D4,c1>}`

dove per ogni documento si riporta di seguito l'elenco delle parole in esso presenti, con le relative occorrenze dei termini ne:

`D1={T1:1, T2:4}` `D2={T1:2, T3:2}`
`D3={T2:1, T4:4}` `D4={T2:2, T3:3}`

NB: illustrare chiaramente tutte le fasi di costruzione del classificatore

(PUNTI 7)

b) Determinare la classe di appartenenza del seguente documento `d={T1:1,T3:2,T5:1}`

(PUNTI 3)

2) Dati i documenti D1, D2 e D3 e la query `Q` rappresentati come vettori di pesi TF-IDF non normalizzati:

	T1	T2	T3	T4	T5	T6
D1	1	2	2	0	0	0
D2	0	0	1	2	2	0
D3	3	4	0	0	3	0
Q	0	0	5	2	0	0

a) Calcolare il ranking dei documenti rispetto alla query `Q` utilizzando la similarità del coseno.

(PUNTI 3)

b) Assumendo che `D1` e `D2` siano rilevanti, mentre `D3` non sia invece rilevante, riformulare la query utilizzando l'algoritmo di Rocchio (utilizzare α=0.75 e β=0.25, dove α e β sono i pesi da assegnare al centroide degli esempi positivi e negativi, rispettivamente).

(PUNTI 6)

3) Illustrare in maniera sintetica il problema della *overspecialization* (*sovraspecializzazione*) dei content-based recommender systems

(PUNTI 5)

4) Descrivere la metrica *nDCG* (normalized Discounted Cumulative Gain), illustrandone calcolo e principi di base.

(PUNTI 3)

5) Sia *q* una query che ha 5 documenti rilevanti nella collezione. Supponiamo che un algoritmo di ritrovamento applicato a *q* riporti il seguente ranking R_q: `D1 D5 D7 D9 D4`. Supponendo di avere dei giudizi di rilevanza non binari espressi in una scala a 5 valori (1-5), e assumendo che `D1` e `D7` abbiano rilevanza pari a 5, mentre `D4` abbia rilevanza pari a 4, calcolare il valore dell'*nDCG* per *q*.

(PUNTI 4)