

Prova scritta di
METODI PER IL RITROVAMENTO DELL'INFORMAZIONE
C.d.L. in Informatica - A.A. 2020-21
Docenti: P. Basile, P. Lops - 9 Luglio 2021

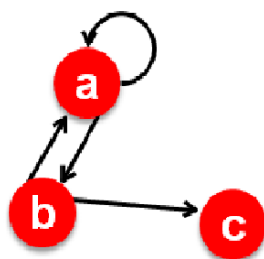
Nome e Cognome : _____

Matricola : _____

- 1) Descrivere in maniera sintetica i principi alla base del PageRank, focalizzando l'attenzione sulla formulazione basata sulle matrici di adiacenza stocastiche

(PUNTI 6)

- 2) Riportare la matrice di adiacenza stocastica corrispondente al seguente grafo:



(PUNTI 6)

- 3) 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:

$$TR = \{ \langle D1, c1 \rangle, \langle D2, c2 \rangle, \langle D3, c1 \rangle, \langle D4, c2 \rangle \}$$

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

$D1 = \{T1:2, T2:2\}$

$D2 = \{T1:1, T3:2\}$

$D3 = \{T2:1, T4:3\}$

$D4 = \{T2:2, T3:4\}$

NB: illustrare chiaramente tutte le fasi di costruzione del classificatore

(PUNTI 7)

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

(PUNTI 3)

- 4) Sia q una query che ha 5 documenti rilevanti nella collezione. Supponiamo che un algoritmo di ritrovamento riporti il seguente ranking R_q (R indica che il documento è rilevante; N indica che il documento è non rilevante; il risultato più a sinistra è il top della lista):

R_q : RNRNRNNRNN

- a) Fornire la descrizione sintetica delle metriche: *Recall*, *R-precision* ed *Average Precision*

(PUNTI 4)

- b) Calcolare *Recall*, *R-precision* ed *Average Precision* per la query q

(PUNTI 6)