

Appello del 13/04/21

1) 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 D3 D7 D9 D4

Supponiamo che D1, D7 e D4 siano documenti rilevanti per q

- Fornire la descrizione sintetica delle metriche: F1, R-Precision ed Average Precision

(PUNTI 3)

- Calcolare F1, R-Precision ed Average Precision per la query q

(PUNTI 3)

2) Siano dati l'insieme delle categorie $C = \{c_1, c_2\}$ e una collezione di 100 documenti definiti sul vocabolario $V = \{T_1, T_2, T_3, T_4, T_5\}$.

- Costruire un classificatore bayesiano per C , addestrandolo sul seguente training set TR :

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

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

$D1 = \{T1:2, T2:3, T4:4\}$ $D2 = \{T1:1, T3:2\}$

$D3 = \{T2:2, T3:4\}$ $D4 = \{T1:4, T4:2\}$

NB: illustrare chiaramente tutte le fasi di costruzione del classificatore

(PUNTI 7)

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

(PUNTI 2)

3) Descrivere in maniera sintetica i principi alla base del PageRank, focalizzando l'attenzione sulla formulazione basata sul Flow model

(PUNTI 5)

4) Descrivere l'ontologia linguistica WordNet ed indicare come è organizzata e rappresentata la conoscenza lessicale al suo interno.

(PUNTI 5)

5) Descrivere, commentando opportunamente, la funzione

per il calcolo delle predizioni dei rating in un algoritmo di filtraggio collaborativo di tipo user to user.

(PUNTI 5)

1)

R_q

D 1 X

D 5 0

D 3 0

D 7 X

D 9 0

3 documenti rilevanti:

$$R\text{-Precision} = \frac{1}{3}$$

$$P = \frac{2}{5}$$

$$R_2 = \frac{2}{3}$$

P R

1 1/3

2/4 2/3

$$AP = \frac{1}{m} \sum_{k=1}^m \text{Precision}(p_{2k}) \quad m = \# \text{ rel.}$$

P

R

1

0,1

1

0,2

1

0,3

1

 $\longrightarrow 1/3$

0,3

0,4

0,3

0,5

0,3

0,6

0,5

 $\longrightarrow 2/3$

0

0,7

0

0,8

0

0,9

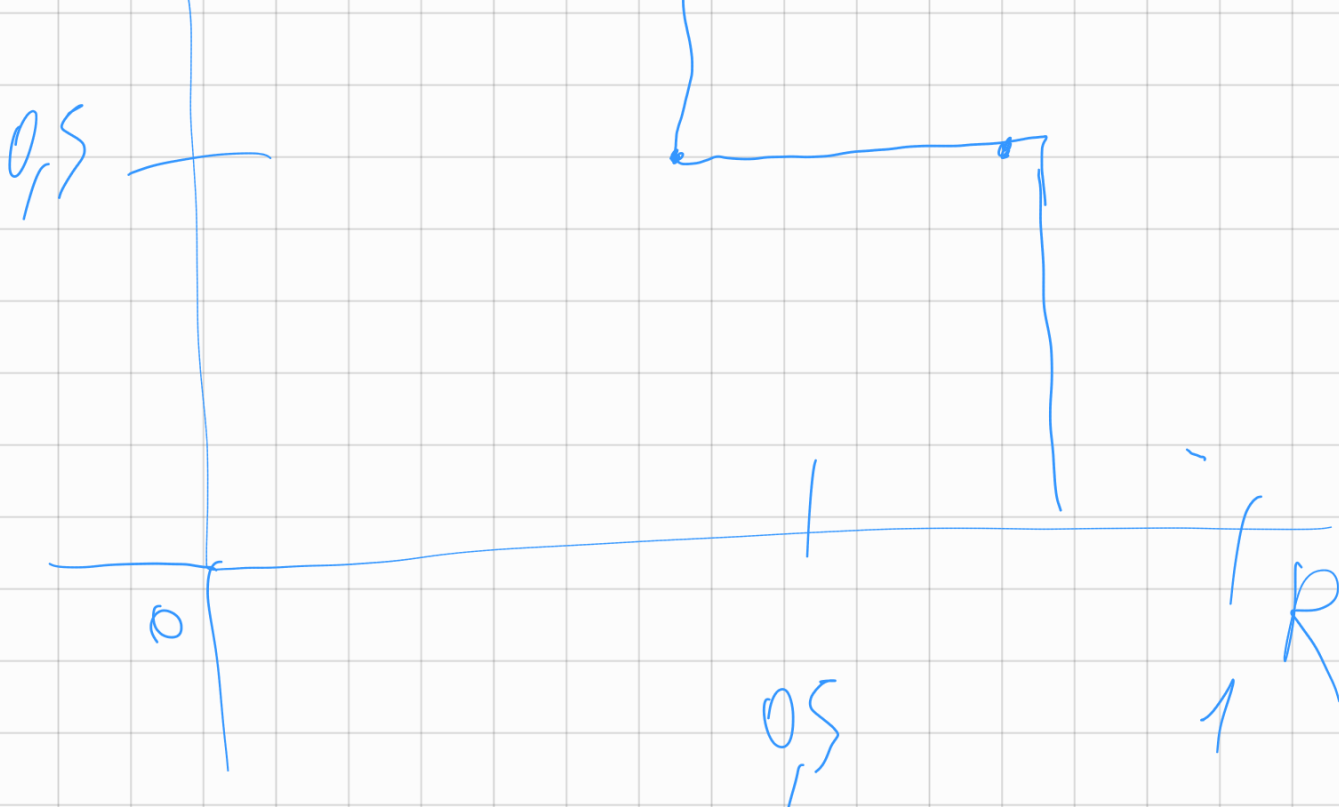
0

1

$$F1 = \frac{2}{3} \cdot \frac{2}{3} \cdot 2$$

$$\frac{2}{3} + \frac{2}{3}$$

P



DEFINIZIONI

R-PRECISION : è la precisione alla Posizione R nella classifica, dove R è il numero totale dei documenti rilevanti per la query. Il sistema perfetto dovrebbe avere score 1.0. È utile per osservare il comportamento dell'algoritmo per query.

F1: misura le performance dell'algoritmo che prende in considerazione sia il recall che la precision
si calcola come:

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

F_1 è un caso particolare della misura F_β

$$F_\beta = \frac{(1+\beta^2)PR}{\beta^2 P + R}$$

Se $\beta = 0$ abbiamo $F_\beta = P$

Recall è calcolo la completezza del sistema e quindi l'abilità di trovare tutti i componenti rilevanti. ~~*~~ ①

h) WordNet si basa sull'ontologia linguistica che rappresenta in maniera esplicita e formale le conoscenze linguistiche umane. Viene realizzato un mapping tra word form e word meaning. Una word form è polysemic se ad essa possono essere associate più word meaning.

Due parole sono Synonyms se ad esse è associata la stessa Word meaning. Lo scopo principale è quello di trasferire ad un computer tutta la conoscenza linguistica.

* ① Average Precision: media di tutte le precisioni nelle varie posizioni del ranking dove trovano i documenti rilevanti su tutti i documenti rilevanti.

