Esercizio Distance Function Data Technology

Matteo Colella, Matteo Angelo Costantini, Dario Gerosa

23/02/2018

Di seguito viene mostrata una soluzione per il problema della creazione della *distance function*, richiesta dall'esercizio facoltativo. È stato deciso di utilizzare una soluzione combinata di distanze avvaldoci di *edit distance* e di *tf-idf*.

IBM Corporation
AT&T Corporation
Microsoft Corporation
Google Inc
Repubblica Democratica del Congo
Repubblica Democratica di Corea
Repubblica Democratica Tedesca
Associazione Calcio Milan
Torino Football Club
Football Club Internazionale Milano

Tabella 1: Documenti reference

kongo
korea
milna
intrnazionale
torino
repubblica tedesca
att
ibm corporation
microft crpoation
goog

Tabella 2: Documenti target

Il calcolo della distanza combinata è suddiviso in tre fasi. La prima fase consiste nel *tokenizzare* e normalizzare (rendendo *case insensitive*) tutti i documenti di reference e target. Viene poi calcolata la distanza di edit di ogni token dei documenti di reference da ogni token dei documenti target i cui risultati sono riportati in tabella 3.

Dopo aver trovato la distanza per ogni coppia di token, viene calcolato il valore della funzione *tf-idf* per ogni token di ogni documento di reference (tabella 1) utilizzando la seguente formula:

$$(tf\text{-}idf_{i,j}) = tf_{i,j} * idf_i \tag{1}$$

Dove $tf_{i,j}$ è calcolato dividendo il numero di volte in cui il token i compare nel documento

	kongo	korea	milna	intrnazionale	torino	repubblica
ibm	5.0	5.0	4.0	12.0	5.0	9.0
corporation	9.0	8.0	10.0	10.0	8.0	9.0
at&t	5.0	5.0	5.0	12.0	6.0	10.0
corporation	9.0	8.0	10.0	10.0	8.0	9.0
microsoft	8.0	8.0	7.0	11.0	7.0	10.0
corporation	9.0	8.0	10.0	10.0	8.0	9.0
google	4.0	5.0	6.0	10.0	5.0	9.0
inc	4.0	5.0	3.0	11.0	4.0	9.0
repubblica	10.0	9.0	8.0	11.0	9.0	0.0

Tabella 3: Estratto dell'output del calcolo della distanza di edit

	0	1	2	3	4	5
ibm	1.66	0.00	0.00	0.00	0.00	0.00
corporation	0.87	0.87	0.87	0.00	0.00	0.00
at&t	0.00	1.66	0.00	0.00	0.00	0.00
corporation	0.87	0.87	0.87	0.00	0.00	0.00
microsoft	0.00	0.00	1.66	0.00	0.00	0.00
corporation	0.87	0.87	0.87	0.00	0.00	0.00
google	0.00	0.00	0.00	1.66	0.00	0.00
inc	0.00	0.00	0.00	1.66	0.00	0.00
repubblica	0.00	0.00	0.00	0.00	0.43	0.43

Tabella 4: Estratto dell'output del calcolo della funzione tf-idf

j per la lunghezza in token del documento j:

$$tf_{i,j} = \frac{n_{i,j}}{|d_j|} \tag{2}$$

Il valore idf_i è pari al logarimo del rapporto tra il numero di documenti di reference e il numero di documenti di reference che contengono il token i:

$$idf_i = \log_2 \frac{|D|}{|\{d : i \in d\}|} \tag{3}$$

I risultati del calcolo della funzione *tf-idf* per i documenti reference della tabella 1 sono riportati in tabella 4.

Infine, calcoliamo il valore della distanza di ogni documento target da ogni documento di reference facendo una somma pesata tra la distanza di edit di ogni token del target dal token di reference più vicino (escludendo i token reference già scelti in precedenza) moltiplicata per il valore di *tf-idf* del token di reference scelto.

$$combined\_distance_{i,j} = \sum_{h \in tokens(i)} edit\_distance(h,k) * tf\text{-}idf_{k,i} \tag{4}$$

Dove k è il token reference di distanza minima dal token h senza considerare tutti i token di reference già utilizzati dai token $t < h$. Se il documento i ha più token del documento di reference j, tutti i token in eccesso hanno peso 0.

La distanza così calcolata è riportata in tabella 5.

	kongo	korea	milna	intrnazionale	torino	repubblica tedesca	atet	ibm corporation	microft crpoation	googe
IBM Corporation	8.30	8.30	6.64	8.68	8.30	23.63	6.64	**1.74**	11.70	8.30
AT&T Corporation	8.30	8.30	8.30	8.68	9.97	19.44	**1.66**	8.38	11.70	8.30
Microsoft Corporation	13.29	13.29	11.63	8.68	11.63	21.10	13.29	15.02	**5.06**	11.63
Google Inc	6.64	8.30	4.98	16.61	6.64	24.91	6.64	14.95	23.25	**1.66**
Repubblica Democratica del Congo	**0.83**	3.32	3.32	4.78	3.32	4.15	2.49	7.47	8.02	2.49
Repubblica Democratica di Corea	3.32	**0.83**	3.32	4.78	3.32	4.15	3.32	6.64	8.02	3.32
Repubblica Democratica Tedesca	7.75	5.54	6.64	6.37	6.64	**0.00**	6.64	12.38	11.80	7.75
Associazione Calcio Milan	5.54	3.87	**1.55**	8.86	5.54	15.38	5.54	11.95	10.51	3.87
Torino Football Club	4.43	4.43	3.10	11.07	**0.00**	12.84	3.10	10.85	12.84	5.54
Football Club Internazionale Milano	2.90	2.90	1.66	**0.83**	3.32	8.71	2.32	6.39	8.80	2.90

Tabella 5: Distanza combinata

Pseudocodice

$R \leftarrow$ *documenti reference*

$T \leftarrow$ *documenti target*

$combined\_distance \leftarrow matrix(|R|, |T|)$

for all $i \in R$ **do**

  for all $j \in T$ **do**

    $ref\_tokens \leftarrow tokens(i)$

    $distance \leftarrow 0$

    for all $k \in tokens(j)$ **do**

      $h \leftarrow$ *token meno distante da* k *in* ref_tokens

      Rimuovi h *da* ref_tokens

      $distance \leftarrow distance + edit\_distance(h, k) * tdidf[h, j]$

    end for

    $combined\_distance[i, j] \leftarrow distance$

  end for

end for