Valutazione qualità

Gabriele Pisciotta

10/12/2020

In questa sperimentazione vogliamo valutare la qualità dei match ottenuti attraverso le query per bibref, sia verso l'indice locale sia verso Crossref. Qui estendiamo la precedente (05/12/2020) di 5 documenti, quindi i risultati ottenuti tengono in considerazione anche i documenti precedentemente elaborati, al fine di avere un risultato su un campione di esempi più ampio e significativo possibile.

Configurazione della sperimentazione

- Selezionato 10 documenti citanti e relative citazioni (output di BEE)
- Creato un CSV estraendo ogni coppia (DOI, bibref) presente nei 5 documenti citanti
- Per ogni riga del CSV, prendo la bibref ed effettuo la query sia verso l'indice locale, sia verso Crossref, salvando i risultati in altre due colonne del CSV.
- Annoto manualmente il DOI nei casi:
  - DOI esplicito mancante
  - DOI discordanti tra locale e remoto
  - DOI discordanti tra locale ed esplicito
  - DOI discordanti tra remoto ed esplicito.
- Se non è stato possibile trovare un DOI manualmente, inserisco un "???".
- Escludo dal file CSV le righe appartenenti ai seguenti casi:
  - il doi manuale non è stato possibile trovarlo (valore: ???)
  - il doi trovato manualmente non è incluso nel dump di Crossref (e quindi non sarebbe possibile trovarlo!)

Esempio di tabella risultante

A	B	C	D	E	F
explicit_doi	**bibref**	**retrieved_local_doi**	**retrieved_remote_doi**	**manual_doi**	
	Mendoza, L, Vilela, R, Voelz, I	10.1101/cshperspect.a019562	10.1101/cshperspect.a019562	10.1101/cshperspect.a019562	
	Prabhu, RM, and Patel, R. Mu	10.1111/j.1470-9465.2004.00843.x	10.1111/j.1470-9465.2004.00843.:	10.1111/j.1470-9465.2004.00843.x	
	Bittencourt, AL, Marback, R, M	10.4269/ajtmh.2006.75.936	10.4269/ajtmh.2006.75.936	10.4269/ajtmh.2006.75.936	
10.1128/cmr.18.3.556-569.2005	Spellberg, B, Edwards, J, Jr, II	10.1128/cmr.18.3.556-569.2005	10.1128/cmr.18.3.556-569.2005	10.1128/cmr.18.3.556-569.2005	
10.1371/journal.pntd.0003984	Blumentrath, CG, Grobusch, M	10.1371/journal.pntd.0003984	10.1371/journal.pntd.0003984	10.1371/journal.pntd.0003984	
	Gugnani, HC. Entomophthoro	10.1007/bf00158574	10.1007/bf00158574	10.1007/bf00158574	
10.1037/0033-2909.133.3.482	Levin, R, Nielsen, T.A. Disturb	10.1037/0033-2909.133.3.482	10.1037/0033-2909.133.3.482	10.1037/0033-2909.133.3.482	
	Schredl, M. Researching Drea	10.1007/978-3-319-95453-0	10.1007/978-3-319-95453-0	10.1007/978-3-319-95453-0	
10.1037/a0037749	Böckermann, M, Gieselmann,	10.1037/a0037749	10.1037/a0037749	10.1037/a0037749	

La tabella completa e tutte quelle generate nei passaggi intermedi, è possibile visionarli (qui)

Elementi valutati

Saranno valutati i seguenti valori al fine di caratterizzare la qualità dei match:

- numero di volte in cui il DOI locale è corretto e quello remoto no
- numero di volte in cui il DOI remoto è corretto e quello locale no
- numero di volte in cui sono entrambi errati e identici
- numero di volte in cui sono entrambi errati ma differenti
- numero di volte in cui sono entrambi corretti e identici
- numero di volte in cui troviamo un supplemento del doi corretto come risultato (usando l'indice)
- numero di volte in cui troviamo un supplemento del doi corretto come risultato (usando le API)
- percentuale dei locali sbagliati che sono supplementi sul totale
- percentuale dei remoti sbagliati che sono supplementi sul totale
- percentuale dei locali sbagliati che sono supplementi sugli errori usando l'indice
- percentuale dei remoti sbagliati che sono supplementi sugli errori usando le API
- percentuale di corrispondenza tra DOI locale e DOI manuale (quanta correttezza usando l'indice)
- percentuale di corrispondenza tra DOI remoto e DOI manuale (quanta correttezza usando le API remote)
- percentuale di corrispondenza tra DOI remoto e DOI locale (quanta corrispondenza tra l'indice e le API)
- percentuale entrambi corretti
- percentuale entrambi errati e identici

- percentuale entrambi errati ma differenti

Riguardo il caso in cui sono entrambi errati siamo infine interessati a:

- percentuale entrambi errati e identici (su tutti gli errati)
- percentuale entrambi errati ma differenti (su tutti gli errati)

Risultati

I seguenti risultati sono stati valutati su un numero di 413 record a seguito di un preprocessing volto ad escludere le righe appartenenti a casistiche indesiderate, come definito nella configurazione della sperimentazione, dal totale 17 sono stati esclusi perchè non posseduti dal dump di Crossref.

- Numero di volte in cui il DOI locale è corretto e quello remoto no: 17
- Numero di volte in cui il DOI remoto è corretto e quello locale no: 5
- Numero di volte in cui sono entrambi errati e identici: 6
- Numero di volte in cui sono entrambi errati ma differenti: 2
- Numero di volte in cui sono entrambi corretti e identici: 383
- Numero di volte in cui vengono restituiti supplementi dall'indice locale: 5
- Numero di volte in cui vengono restituiti supplementi dalle API: 11
- Percentuale locali sbagliati che sono supplementi (sul totale): 1.21%
- Percentuale remoti sbagliati che sono supplementi (sul totale): 2.66%
- Percentuale locali sbagliati che sono supplementi (sui locali sbagliati): 38.46%
- Percentuale remoti sbagliati che sono supplementi (sui remoti sbagliati): 44.0%
- Percentuale di corrispondenza tra DOI locale e DOI manuale (quanta - correttezza usando l'indice): 96.85%
- Percentuale di corrispondenza tra DOI remoto e DOI manuale (quanta correttezza usando le API remote): 93.95%
- Percentuale di corrispondenza tra DOI remoto e DOI locale (quanta corrispondenza tra l'indice e le API): 94.19%
- Percentuale entrambi corretti: 92.74%
- Percentuale entrambi errati e identici: 1.45%
- Percentuale entrambi errati ma differenti: 0.48%

Circa gli errati:

- Percentuale entrambi errati e identici su tutti gli errati: 75.0%
- Percentuale entrambi errati ma differenti su tutti gli errati: 25.0%