

Compare Linking across linking runs

Christoph, Flavio, Mona

2023-07-23

Linking Advisors and Graduates from ProQuest to MAG AuthorIds

To check whether the linking makes sense we compare the links across several independent runs of the labelling.

Which fields and which linking runs to compare?

Comparison for Graduates

```
fields_to_process <- c("biology", "business", "chemistry", "computer science", "economics", "engineering", "environmental science", "geology", "history", "mathematics", "materials science", "philosophy", "physics", "political science", "psychology")
fields_to_process <- c("biology", "business", "computer science", "economics", "engineering", "environmental science", "geology", "history", "mathematics", "materials science", "philosophy", "physics", "political science", "psychology")

#fields_to_process <- c("mathematics", "economics", "physics")

linker1 = "mona_degree0"
linker2 = "christoph_degree0"

res_combined_graduates <- reduce(res, rbind)

res_combined_graduates %>%
  mutate(across(where(is.numeric), \(x) round(x, digits = 2))) %>% # this is an updated form of mutate
  kable(format = "latex", digits = 2, booktabs = TRUE) %>%
  kable_styling(latex_options = c("striped", "scale_down"))

#res_combined_graduates %>% mutate(across(where(is.numeric), \(x) round(x, digits = 2))) %>% View()
```

Comparison for Advisors

```
fields_to_process <- c("biology", "business", "chemistry", "computer science", "economics", "engineering", "environmental science", "geology", "history", "mathematics", "materials science", "philosophy", "physics", "political science", "psychology")
fields_to_process <- c("economics")

linker1 = "mona_degree0" # flavio_with_protocol_cleaninst
linker2 = "christoph_degree0" # christoph_degree0_with_protocol_updated
```

field	same	only1	only2	diff	diff_rell	nlink1	nlink2	namedist_pq_1	namedist_pq_2	namedist_diff	namedist_diff_pq_1	namedist_diff_pq_2	namedist_pq_only1	namedist_pq_only2
biology	0.45	0.03	0.51	0.01	0.02	40569	79517	0.07	0.08	0.05	0.07	0.08	0.07	0.01
business	0.24	0.19	0.56	0.01	0.01	5872	10683	0.08	0.08	0.27	0.08	0.08	0.08	0.03
computer science	0.52	0.08	0.40	0.00	0.01	12169	18599	0.07	0.05	0.17	0.07	0.05	0.07	0.01
economics	0.20	0.01	0.78	0.00	0.02	2048	9367	0.10	0.08	0.25	0.10	0.08	0.10	0.00
engineering	0.68	0.25	0.07	0.00	0.00	23776	19220	0.06	0.06	0.07	0.06	0.06	0.06	0.23
environmental science	0.43	0.01	0.55	0.01	0.01	4944	10843	0.09	0.09	0.06	0.09	0.09	0.09	0.00
geology	0.29	0.00	0.69	0.02	0.07	2401	7666	0.09	0.12	0.16	0.09	0.12	0.09	0.00
history	0.42	0.09	0.48	0.01	0.02	4744	8277	0.10	0.10	0.11	0.10	0.10	0.10	0.02
mathematics	0.20	0.00	0.79	0.00	0.00	3829	18520	0.06	0.06	0.17	0.06	0.06	0.06	0.00
materials science	0.38	0.07	0.54	0.01	0.02	7975	16115	0.08	0.08	0.07	0.08	0.08	0.08	0.01
philosophy	0.11	0.06	0.83	0.00	0.00	614	3301	0.08	0.08	NaN	NaN	NaN	0.08	0.01
physics	0.06	0.39	0.18	0.36	0.44	28191	20980	0.32	0.08	0.32	0.08	0.08	0.32	0.16
political science	0.20	0.04	0.75	0.01	0.06	3566	13506	0.12	0.08	0.25	0.12	0.08	0.12	0.00
psychology	0.24	0.01	0.74	0.01	0.05	15897	61227	0.08	0.08	0.10	0.08	0.08	0.08	0.00

field	same	only1	only2	diff	diff_rell	nlink1	nlink2	namedist_pq_1	namedist_pq_2	namedist_diff	namedist_diff_pq_1	namedist_diff_pq_2	namedist_pq_only1	namedist_pq_only2
biology	0.00	0.00	1.00	0.00	NaN	1	29998	NaN	0.02	NaN	NaN	NaN	NaN	0.00
business	0.69	0.00	0.31	0.00	0.00	8336	12213	0.02	0.02	0.10	0.02	0.02	0.02	0.00
chemistry	0.03	0.10	0.10	0.78	0.86	7749	12816	0.34	0.03	0.35	0.34	0.03	0.34	0.03
computer science	0.64	0.20	0.04	0.13	0.13	18989	17488	0.02	0.07	0.27	0.02	0.07	0.02	0.35
economics	0.65	0.30	0.01	0.03	0.03	8803	6024	0.02	0.02	0.06	0.02	0.02	0.02	0.71
engineering	0.15	0.11	0.19	0.56	0.69	22962	21225	0.05	0.27	0.31	0.05	0.27	0.05	0.15
environmental science	0.66	0.32	0.01	0.00	0.00	10343	7205	0.03	0.03	0.16	0.03	0.03	0.03	0.83
geology	0.28	0.35	0.00	0.37	0.37	6236	3665	0.21	0.03	0.29	0.21	0.03	0.21	3.89
history	0.67	0.02	0.30	0.02	0.03	7704	10681	0.03	0.02	0.11	0.03	0.02	0.03	0.00
mathematics	0.97	0.01	0.01	0.01	0.01	12943	13008	0.03	0.03	0.20	0.03	0.03	0.03	0.05
materials science	0.12	0.14	0.00	0.73	0.74	11244	12510	0.29	0.04	0.33	0.29	0.04	0.29	5.54
philosophy	0.92	0.00	0.07	0.01	0.01	4424	4830	0.02	0.02	0.05	0.02	0.02	0.02	0.00
physics	0.49	0.50	0.01	0.00	0.00	12611	6821	0.08	0.07	0.14	0.08	0.07	0.08	4.78
political science	0.94	0.04	0.00	0.02	0.02	13594	13160	0.02	0.02	0.04	0.02	0.02	0.02	0.60
psychology	0.89	0.02	0.08	0.01	0.01	29902	32473	0.02	0.02	0.14	0.02	0.02	0.02	0.00

```

res_combined_advisors <- reduce(res, rbind)

res_combined_advisors %>%
  mutate(across(where(is.numeric), \(x) round(x, digits = 2))) %>% # this is an updated form of mutate
  kable(format = "latex", digits = 2, booktabs = TRUE) %>%
  kable_styling(latex_options = c("striped", "scale_down"))

#res_combined_advisors %>%
# mutate(across(where(is.numeric), \(x) round(x, digits = 2))) %>% View()

```

Compare names manually

```

#
# field="economics"
# linktype="advisors"
# years="19902015"
#
# links <- compare(field, linktype, linker1, linker2, "19902015", inspect=TRUE)
#
#

```