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
fields_to_process <- c("biology", "business", "computer science", "economics", "engineering", "environme

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

linker1 = "mona_degree0"
linker2 = "christoph_degree0"
```

```
res_combined_graduates <- reduce(res, inner_join, by = "variable")

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 %>%
 #s 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

#fields_to_process <- c("economics")

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

variable	biology	business	computer science	economics	engineering	environmental science	geology	history	mathematics	materials science	philosophy	physics	political science	psychology
same	0.45	0.24	0.52	0.20	0.68	0.43	0.29	0.42	0.20	0.38	0.11	0.06	0.20	0.24
only1	0.03	0.19	0.08	0.01	0.25	0.01	0.00	0.09	0.00	0.07	0.06	0.39	0.04	0.01
only2	0.51	0.56	0.40	0.78	0.07	0.55	0.69	0.48	0.79	0.54	0.83	0.18	0.75	0.74
diff	0.01	0.01	0.00	0.00	0.00	0.01	0.02	0.01	0.00	0.01	0.00	0.36	0.01	0.01
diff_rel1	0.02	0.01	0.01	0.02	0.00	0.01	0.07	0.02	0.00	0.02	0.00	0.44	0.06	0.05
nlink1	40569.00	5872.00	12169.00	2048.00	23776.00	4944.00	2401.00	4744.00	3829.00	7975.00	614.00	28191.00	3566.00	15897.00
nlink2	79517.00	10683.00	18599.00	9367.00	19220.00	10843.00	7666.00	8277.00	18520.00	16115.00	3301.00	20980.00	13506.00	61227.00
namedist	0.10	1.09	0.32	1.23	0.11	0.14	0.52	0.25	0.83	0.18	NaN	0.75	1.17	0.38
namedist_pq_1	0.15	0.44	0.33	1.33	0.12	0.31	0.31	0.27	0.48	0.25	NaN	0.41	1.07	0.46
namedist_pq_2	0.11	0.18	0.12	0.11	0.08	0.11	0.22	0.15	0.10	0.10	NaN	0.12	0.08	0.12

variable	biology	business	chemistry	computer science	economics	engineering	environmental science	geology	history	mathematics	materials science	philosophy	physics	political science	psychology
same	0	0.69	0.03	0.64	0.65	0.15	0.66	0.28	0.67	0.97	0.12	0.92	0.49	0.94	0.89
only1	0	0.00	0.10	0.20	0.30	0.11	0.32	0.35	0.02	0.01	0.14	0.00	0.50	0.04	0.02
only2	1	0.31	0.10	0.04	0.01	0.19	0.01	0.00	0.30	0.01	0.00	0.07	0.01	0.00	0.08
diff	0	0.00	0.78	0.13	0.03	0.56	0.00	0.37	0.02	0.01	0.73	0.01	0.00	0.02	0.01
diff_rel1	NaN	0.00	0.86	0.13	0.03	0.69	0.00	0.37	0.03	0.01	0.74	0.01	0.00	0.02	0.01
nlink1	1	8336.00	7749.00	18989.00	8803.00	22962.00	10343.00	6236.00	7704.00	12943.00	11244.00	4424.00	12611.00	13594.00	29902.00
nlink2	29998	12213.00	12816.00	17488.00	6024.00	21225.00	7205.00	3665.00	10681.00	13008.00	12510.00	4830.00	6821.00	13160.00	32473.00
namedist	NaN	0.15	0.44	0.35	0.09	0.45	0.24	0.45	0.16	0.21	0.38	0.05	0.28	0.04	0.15
namedist_pq_1	NaN	0.12	0.39	0.03	0.02	0.06	0.09	0.29	0.17	0.18	0.33	0.04	0.11	0.05	0.09
namedist_pq_2	NaN	0.06	0.04	0.33	0.10	0.35	0.17	0.05	0.03	0.12	0.05	0.06	0.17	0.02	0.11

```
res_combined_advisors <- reduce(res, inner_join, by = "variable")

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)
```