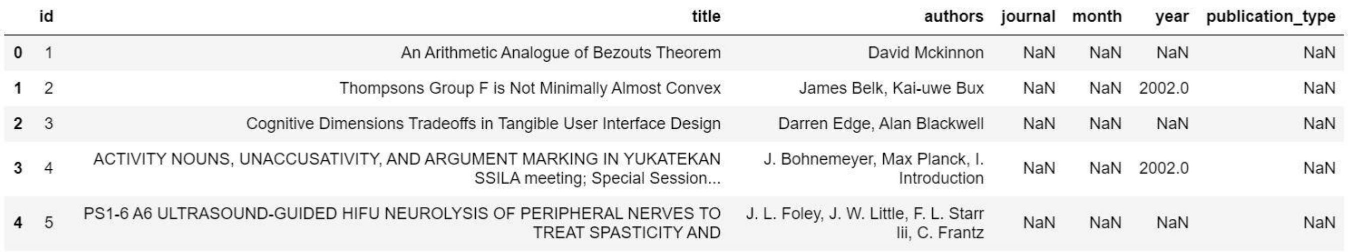
ASSIGNMENT 2

The dedupe part

https://docs.dedupe.io/en/latest/API-documentation.html

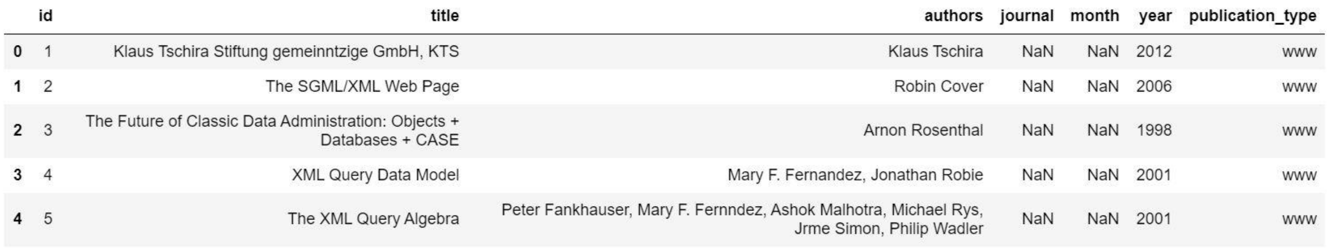
Two datasets

- `citeseer.csv -> 1823978 rows`



	id	title	authors	journal	month	year	publication_type
0	1	An Arithmetic Analogue of Bezouts Theorem	David Mckinnon	NaN	NaN	NaN	NaN
1	2	Thompsons Group F is Not Minimally Almost Convex	James Belk, Kai-uwe Bux	NaN	NaN	2002.0	NaN
2	3	Cognitive Dimensions Tradeoffs in Tangible User Interface Design	Darren Edge, Alan Blackwell	NaN	NaN	NaN	NaN
3	4	ACTIVITY NOUNS, UNACCUSATIVITY, AND ARGUMENT MARKING IN YUKATEKAN SSILA meeting; Special Session...	J. Bohnemeyer, Max Planck, I. Introduction	NaN	NaN	2002.0	NaN
4	5	PS1-6 A6 ULTRASOUND-GUIDED HIFU NEUROLYSIS OF PERIPHERAL NERVES TO TREAT SPASTICITY AND	J. L. Foley, J. W. Little, F. L. Starr Iii, C. Frantz	NaN	NaN	NaN	NaN

- `dblp.csv -> 2512927 rows`



	id	title	authors	journal	month	year	publication_type
0	1	Klaus Tschira Stiftung gemeinntzige GmbH, KTS	Klaus Tschira	NaN	NaN	2012	www
1	2	The SGML/XML Web Page	Robin Cover	NaN	NaN	2006	www
2	3	The Future of Classic Data Administration: Objects + Databases + CASE	Arnon Rosenthal	NaN	NaN	1998	www
3	4	XML Query Data Model	Mary F. Fernandez, Jonathan Robie	NaN	NaN	2001	www
4	5	The XML Query Algebra	Peter Fankhauser, Mary F. Fernndez, Ashok Malhotra, Michael Rys, Jrme Simon, Philip Wadler	NaN	NaN	2001	www

GOAL

Find entities in one table that match entities in the other.

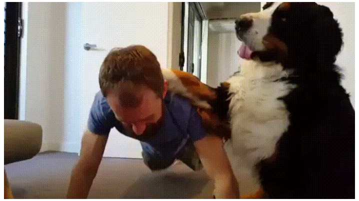
How to do it

- **RecordLink:** A dedupe object that can join two datasets.
- **Steps:**
  - Read the two CSVs (more on this later)
  - Feed the read structures two RecordLink
  - Use prepare_training() to -> set up data in dedupe and also run entity matching measures
  - Use console_label() to ask user for input (active learning part)
  - Call train() to train the classifier
  - Now use join(), pairs(), score(), one_to_one(), many_to_one() to get matching pairs

Reading CSV

- Look at the csv example: https://github.com/dedupeio/dedupe-examples/blob/master/csv_example/csv_example.py
- readData() can take a csv file and give a structure that can be used by dedupe.

PREPARE TRAINING



- This is the part that looks at the data sources, uses blocking and entity matching techniques (non-ML related) to find potential pairs.
- **Problem:** Since the two datasets are in millions, matching across them will consume too much resources.
- **Proper Solution:** Use downsampling (with filtering) so that only entities with matching tokens get sampled.
    - dedupe doesn’t offer this
- **Alright Solution:** Just use the 10000 rows or so - you can do this by tweaking the readData() function (even this takes 5 minutes).

OUTPUT

- The output should be entities that match and their confidence scores.
  - An example format: ***[Row from citeseer, Row from dblp, Confidence score]*** - this demonstrates one pair in your list of matching pairs.
- join() is one useful method
- Other useful methods:
  - https://docs.dedupe.io/en/latest/API-documentation.html#id4
  - pairs() -> gives pairs that are similar
  - score() -> scores the pairs for similarity
  - one_to_one(), many_to_one() -> use scores to get matching entities

What else you can try

- StaticRecordLink ->
  - Requires an already trained model
  - One can train multiple times using this approach
  - One can store a trained model using write_settings()
- Labeled data -> You can also re-use your labeled data (ones you label in the active learning phase) by storing them locally using write_training()
- The csv_example.py file (https://github.com/dedupeio/dedupe-examples/blob/master/csv_example/csv_example.py) is very useful here.

END.