Semester Project - Final Defence

A Semi-Supervised Approach to Citation Matching in the Humanities

Student: Tao Sun

Supervisor: Matteo Romanello

Digital Humanities Laboratory, EPFL

Problem Defintion

- **Scientific Citation**
  - ***Well-structured***
  - Straightforward citation matching
- **Humanities Citation**
  - Old papers are of great importance
  - ***Various*** citation styles throughout history
  - So, how to ***match*** them?

Problem Defintion

- **Problem**:
  - How to match citations in the humanities publications?
  - **One seed ref -> Find all other refs pointing to the same doc**
- **Data**:
  - Citation data from **Linked Books** project
  - Large scale: 4 million references —> reduce search space
  - Lack of Ground Truth —> semi-supervised learning

Approach

Step by Step Approach

- **Local Clustering (1 step)**
  - Group references in one document
- **Global Clustering (2 steps)**
  - Group references among documents

- Do Classification in the ***reduced search space***

Step 1: Local Clustering

- **Abbreviation - Partial Reference**
  - e.g. ibid., voi. i, pp. 167-169.
  - Formulate rules for various situations

Abbreviation	Meaning (to the one right before)
idem / id / eadem / ead	Same author
ibidem / ibid	Same reference
ivi	
op. cit. / op. ctt.	Same source
…	…

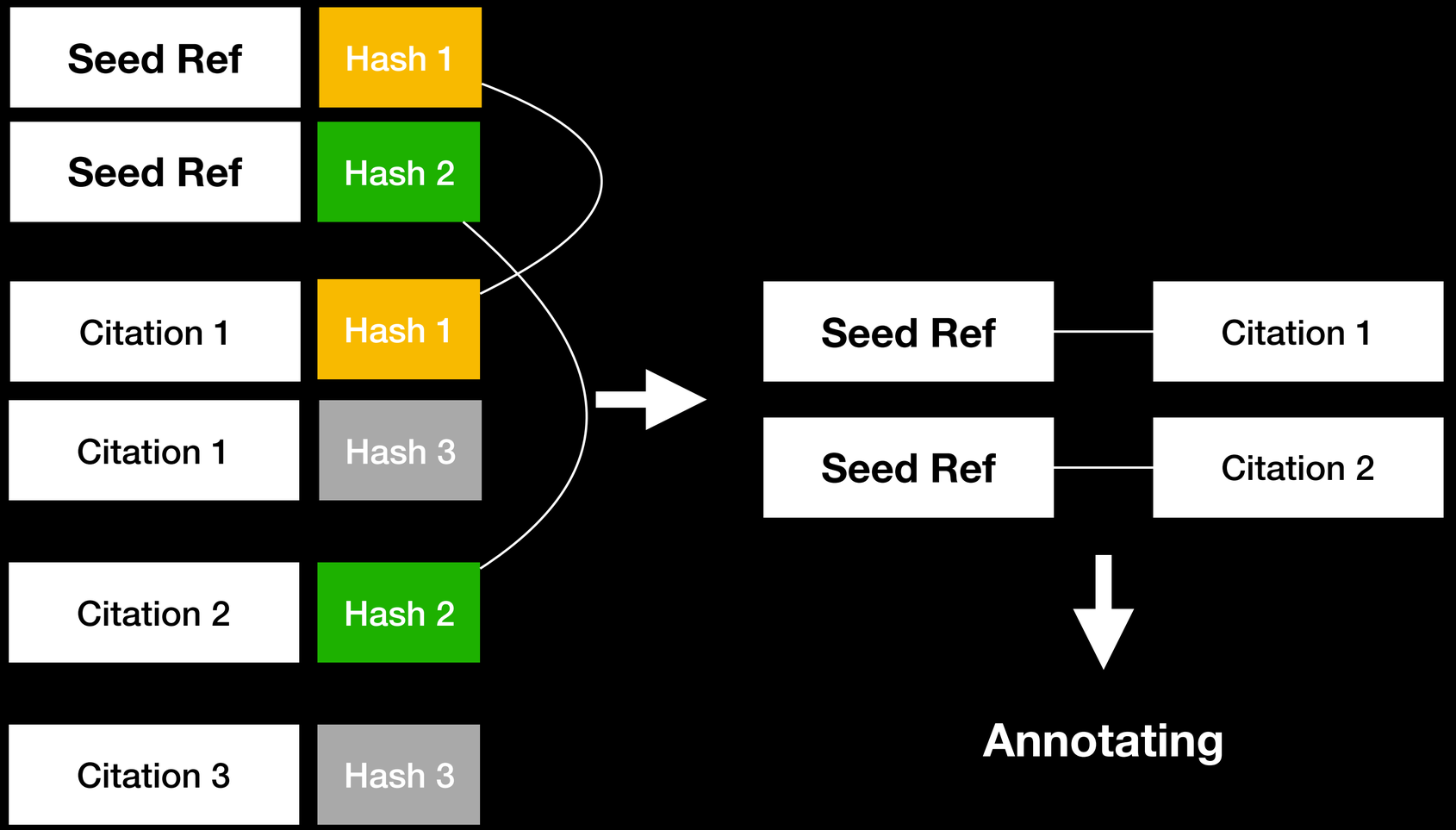
Step 1: Local Clustering

- **Rule-Based Matching**
  - Order in a document
  - Rules of partial references
  - Iteration through all for several times

Step 2: Hash Matching

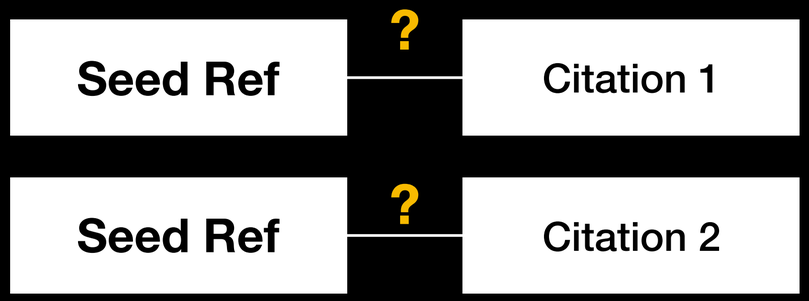
- **Hash Function**
  - Consider citation matching as ***deduplicated*** task
  - Generate keys/hashes for each citation (local cluster)
  - Bigrams of Title, Author Names, {Year-1, Year, Year+1}
    - *Morassi, Novità e precisazioni sul Tiepolo in « Le Arti », 1942, p. 91.*
    - Hashes: morassi #1942, morassi # novita # precisazioni, …

Step 2: Hash Matching



Step 3: Do Classification

- **Build Binary Classifier**
- Logistic Regression
- Feature vector
  - Token/Ngrams Similarity of Author and Title
  - Exactly Year Match (0/1)
  - LCS (Longest Common Sequence) Similarity of Publisher



Step 3: Do Classification

- **Semi-supervised Learning**
- Disambiguation Dataset
  - Ref pointing to doc in the catalogue of Italian libraries
- Pos / Neg Ex. (**50** / **50**) -> Boost classifier
- Run on unlabelled data -> high confidence (0.9)
  -> Enrich training set (**6000** / **6000**)
- Re-train on all

Evaluation

Discussion

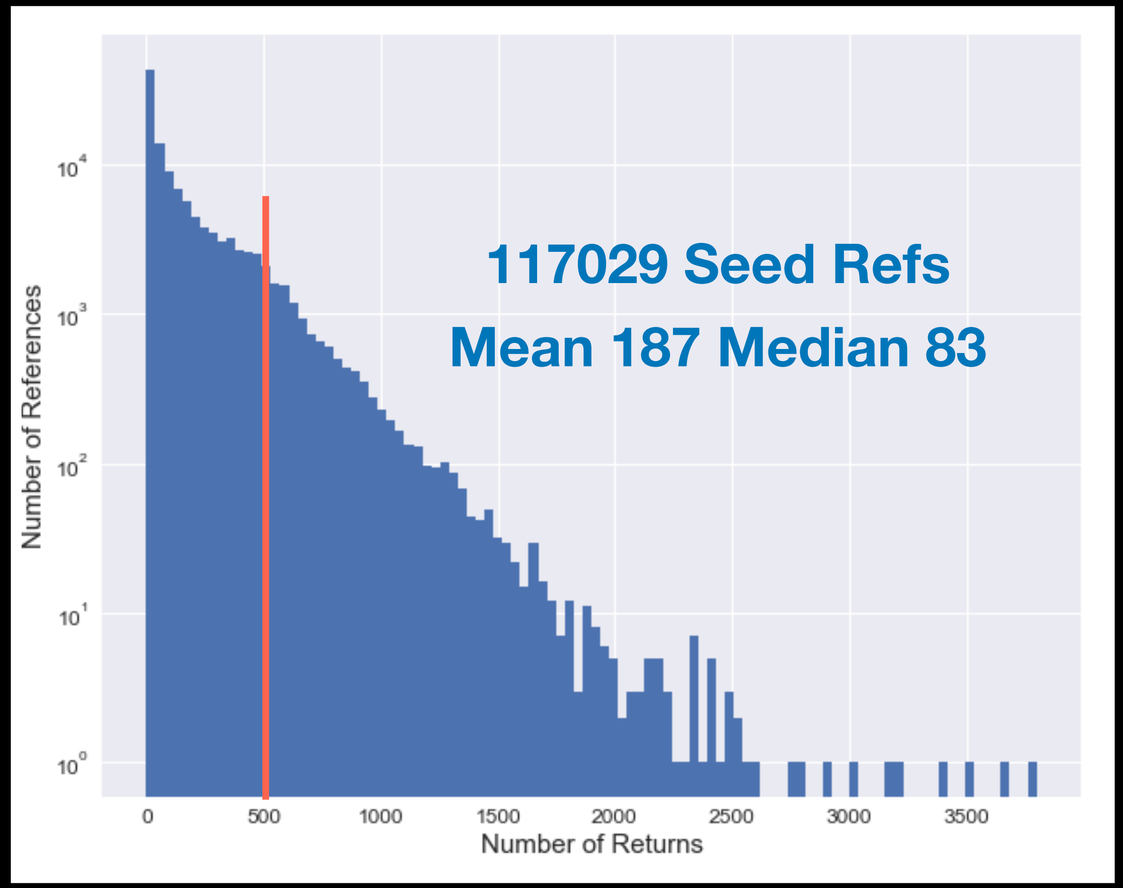
Local Evaluation

- Pair of full ref and partial refs (collected by 2 annotators)

- Hard to generate a **single metric** to evaluate
- Main problems:
  - Highly dependent on extracting orders
  - Highly dependent on parsing results

Hash Evaluation

- Refs with title/author/year fields as “Full” refs (495339 in total)
- Use Black List to ban hashes with returned refs > 500
- 30 Samples for annotating
  - 10 of 0-10 / 10-100 / 100-500 returns



Hash Evaluation

Hash Function	
Baseline	15
Title	3031
Author-Title	1026
Author-Year	1011
Author-Year (Blur)	1350
Author-Year-Page	2
Author-Year-Page (Blur)	3
Author-Year-Num^3	217
Combined	3583

Hash Function	Recall	Precision
Baseline	6.48	27.78
Title	82.44	33.59
Author-Title	73.89	54.74
Author-Year	84.52	62.87
Author-Year (Blur)	86.25	43.82
Author-Year-Page	12.50	12.50
Author-Year-Page (Blur)	25.00	25.00
Author-Year-Num^3	20.75	31.65
Combined	84.47	25.47

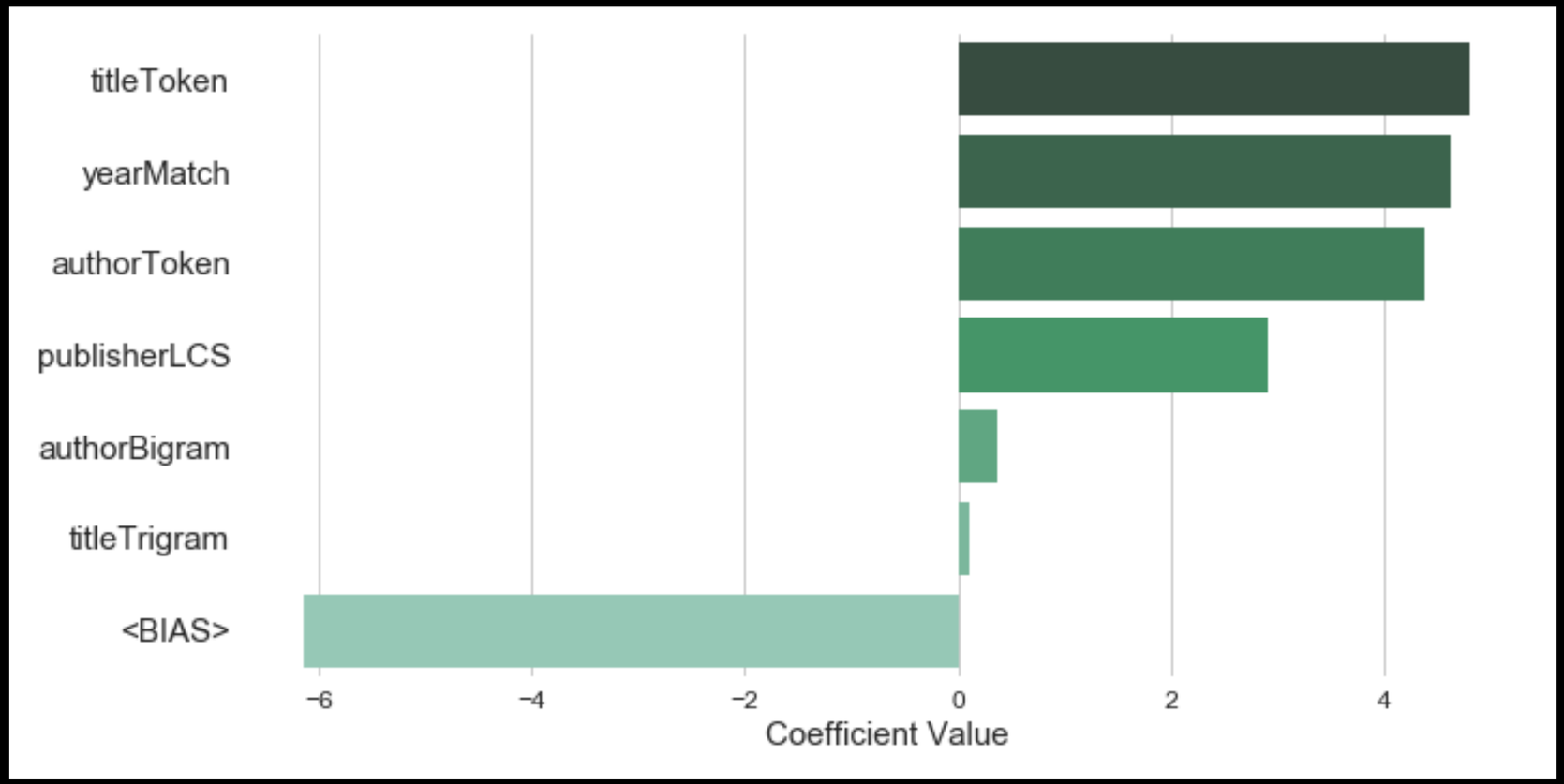
Average of all 30 Seed Refs

Hash Evaluation

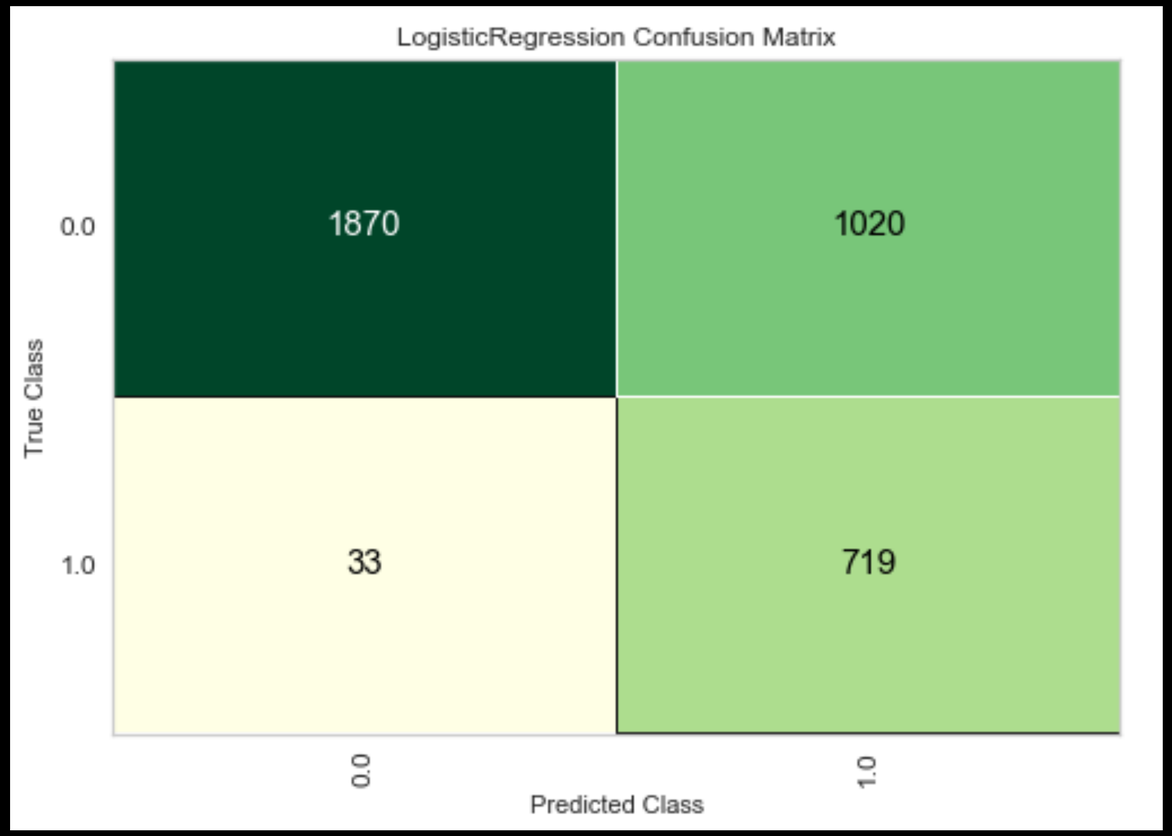
- **False Negative**:
  - 54 citations -> Not in our samples of “**Full**” refs
  - Not contain *year* or *author* or *title* field

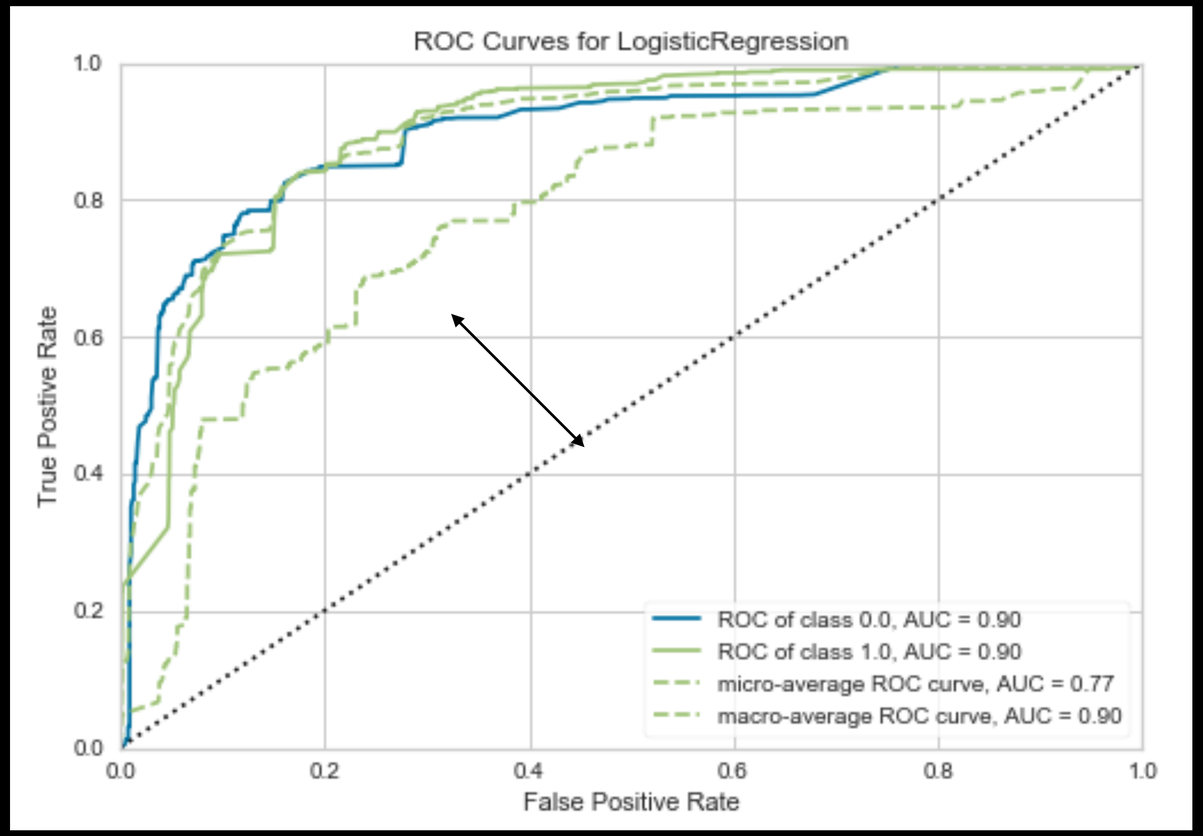
- **Robust on Extracting and Parsing Results**

Classification Evaluation



Classification Evaluation





Recall: 96% **Precision: 41%**

Classification Evaluation

- Difference in the distribution of training and test set
- Training: most of Negative Pairs have zero vector
- Test: each pair share same hashes
  - fewer with zero vector

- Get more hash samples for training?

Reference

- Colavizza, Giovanni, Matteo Romanello, and Frédéric Kaplan. "The references of references: a method to enrich humanities library catalogs with citation data." International Journal on Digital Libraries (2017): 1-11.

- Fedoryszak, Mateusz, and Łukasz Bolikowski. "Efficient Blocking Method for a Large Scale Citation Matching." D-Lib Magazine 20, no. 11/12 (2014).

- Fedoryszak, Mateusz, Dominika Tkaczyk, and Łukasz Bolikowski. "Large scale citation matching using Apache Hadoop." International Conference on Theory and Practice of Digital Libraries. Springer, Berlin, Heidelberg, 2013.

“Thank you!”

– Tao Sun