Text Matching based on Document Embeddings with Application to Measuring Gender Bias in Scientific Community

Kojin Oshiba & Wenshuo Wang & Han Yan

Harvard University

December 11, 2017

Overview

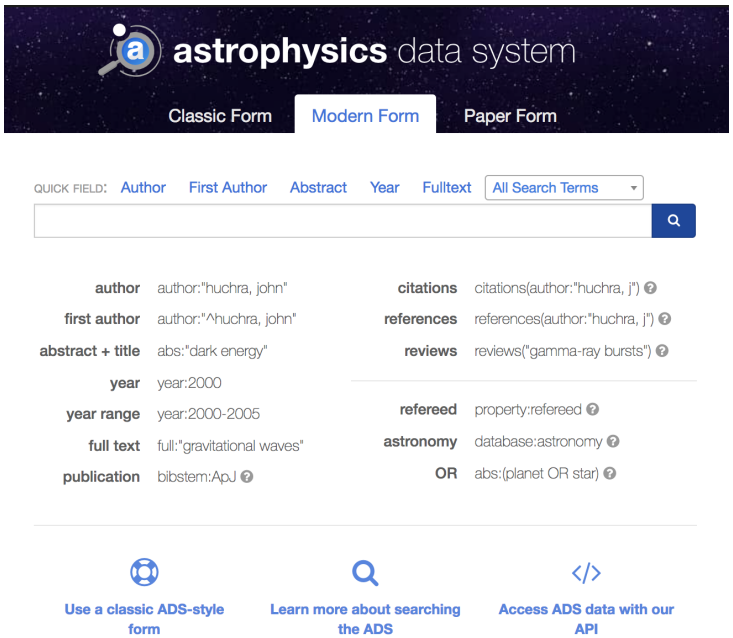
Introduction

Gender bias in science publications

- Question: Do people cite papers less often if they were written by females?
- Covariates: Summary vector obtained by applying Doc2Vec or STM to abstract/full text; number of months elapsed since publication; number of authors.
- Treatment: Readers *perceive* first authors as male (control) vs female (treatment).
- Response: Citation counts

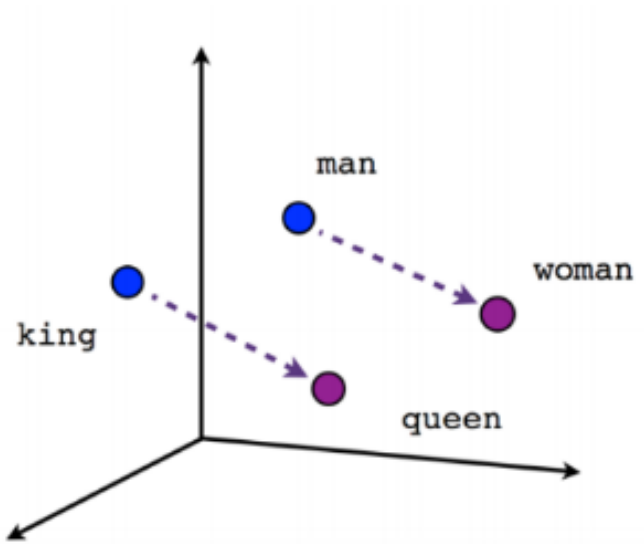
Astronomy Data

- We focus on publications in astronomy because of full access to abstract, author names and citation counts.
- Supplement full texts by finding the same papers on arxiv.



Word Embeddings

- Converts a word to a vector (often times, using neural networks).

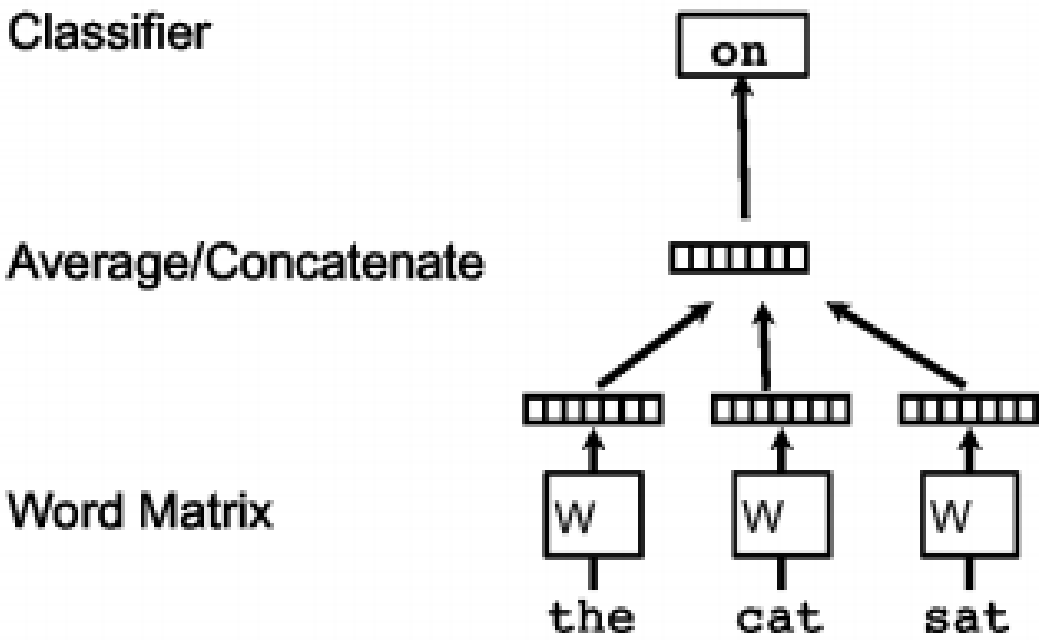


Male-Female

Example: Word2Vec

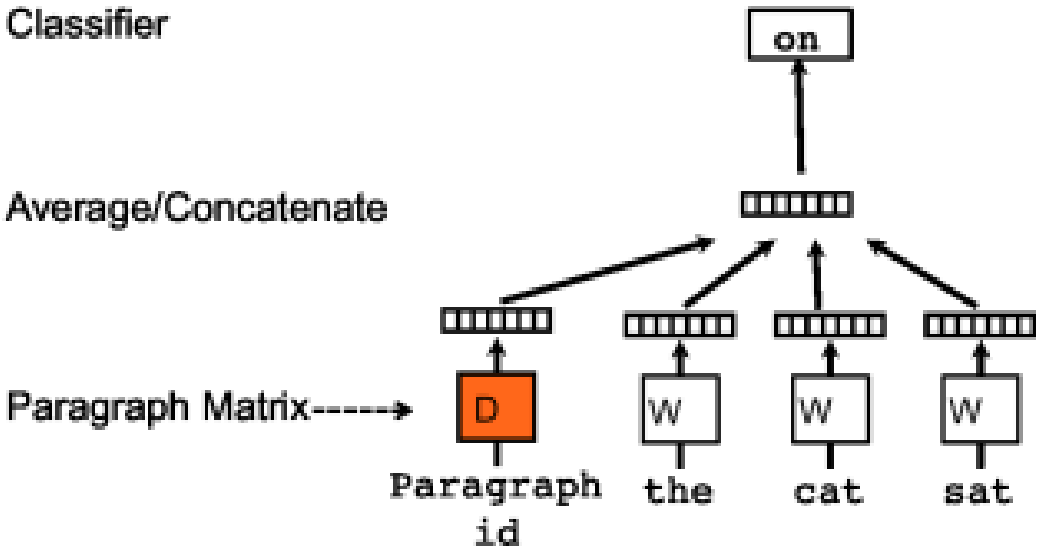
- (One type of) Word2Vec: Predict a word from surrounding words

- Word2Vec is a two layer neural network.



Word2Vec → Doc2Vec

- Simply, add a paragraph ID to the input.



Matching based on Document Embeddings

Cosine Similarity

- Doc2Vec assigns each unit i an m-dimensional non-zero vector X_i.
- A unit's treatment assignment is defined by the perceived gender of the first author's given name.
- For unit i in the treatment group (female) and unit j in the control group (male), define their cosine similarity to be $\frac{X_i \cdot X_j}{\|X_i\|\|X_j\|} \in [-1, 1]$.
- Cosine similarity close to 1 indicates high similarity between the two documents.

Matching Method

- For each treatment unit, want to find the closed match in the control group in terms of cosine similarity.
- Matching without replacement: use each control unit no more than once.
- Determine the matching order.
- Check covariate balance after matching.

Matching Order

- Treatment units with few good matches have high priority.
- Prefix an integer k; find the k closet matches in the control group for each treatment unit (with replacement); for each treatment unit, calculate the range of its cosine similarities with the k control units, and use that to determine the matching order.

	1	2	3	4	5	range ($k = 3$)
1	0.98	0.99	0.96	0.95	0.93	0.99-0.96=0.03
2	0.7	0.98	0.8	0.94	0.85	0.98-0.85=0.13
3	0.92	0.4	0.6	0.3	0.66	0.92-0.6=0.32

Results and Discussion

Results and Discussion

- Assess Cosine Matching Result

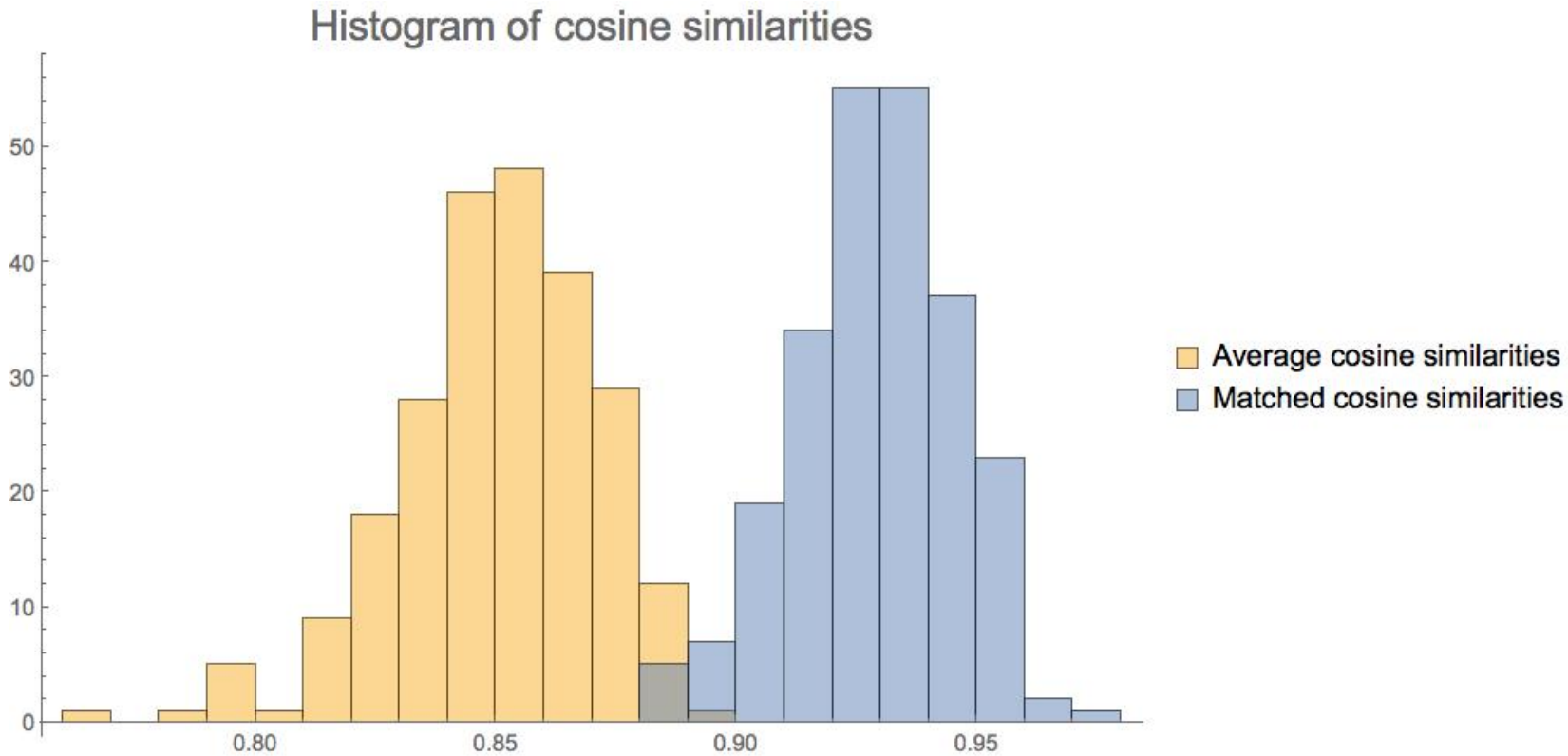


Figure: Histogram of cosine similarities.

Results and Discussion

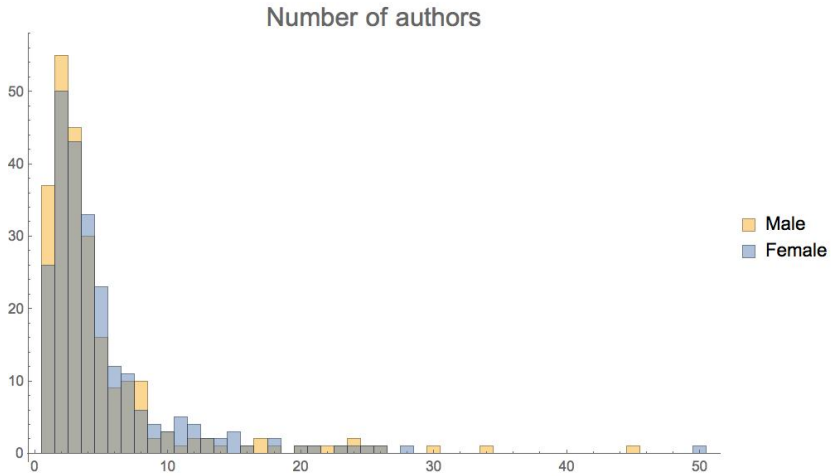
- Covariate Balance check for "Number of authors": Summary Statistics

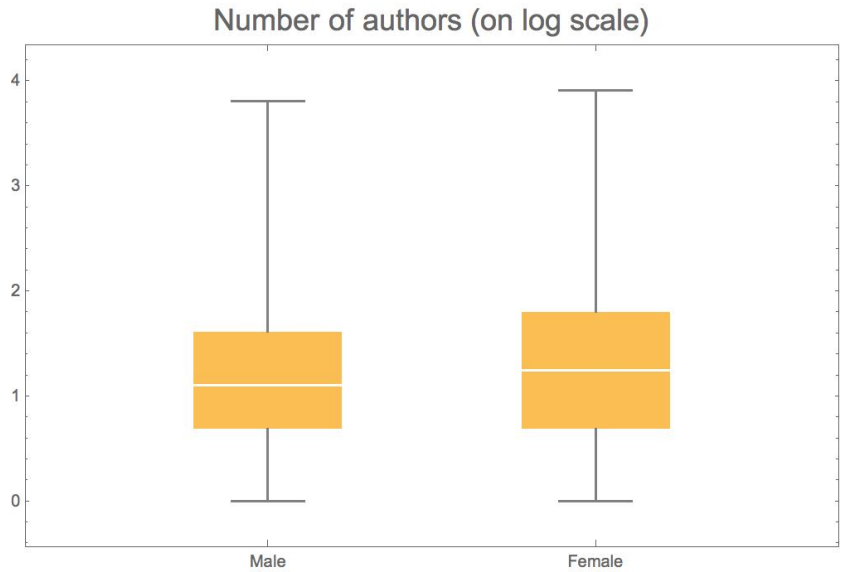
	Female	Matched Male	Male
Mean	5.1932	4.9201	4.3145
Variance	30.7895	34.5379	21.2832

Table: Means and variances of the number of authors.

Results and Discussion

- Covariate Balance check for "Number of authors": Histograms and Box-plots





Results and Discussion

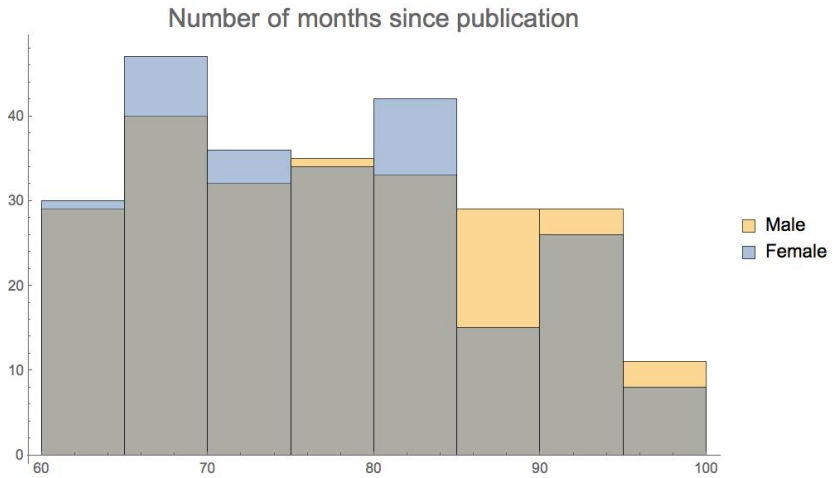
- Covariate Balance check for "Number of Months elapsed since publication": Summary Statistics

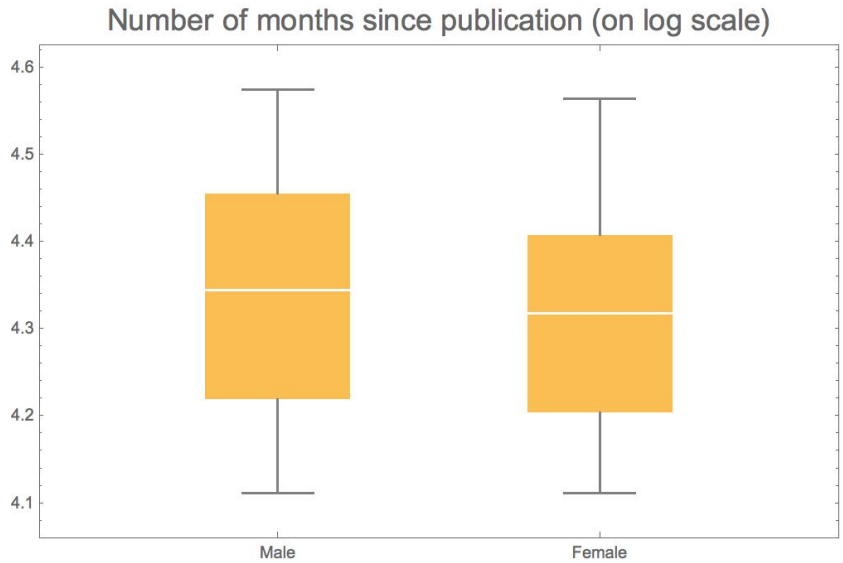
	Female	Matched Male	Male
Mean	76.042	77.3487	76.6183
Variance	98.1754	105.84	102.512

Table: Means and variances of the number of months since publication.

Results and Discussion

- Covariate Balance check for "Number of Months elapsed since publication": Histograms and Box-plots





Results and Discussion

- For i^{th} matched pair, index its treated and control unit by (i, m_i^c) respectively.
- Unit-level treatment effect: $\tau_i = Y_i(1) - Y_i(0)$
- An estimator for τ_i: $\hat{\tau}_i = Y_i^{obs} - Y_{m_i^c}^{obs}$
- Estimator for ATT:
  $\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}(Y_i^{obs} - Y_{m_i^c}^{obs}) = \frac{1}{N}\sum_{i}^{N}(Y_i(1) - Y_{m_i^c}(0)) = -7.9$
- "Female" first-authored paper on average get 8 fewer citations than an otherwise similar paper published with a "male" first author.

Results and Discussion

	Matched Control Group	All Control Group
Aver. Citation Counts	40.08	35.85

Table: Citation counts in control group pre- and post- matching

Potential Extensions

- Adding in more relevant covariates and gathering more data by extending to more earlier years to arrive at more accurate estimate.
- Comparing our method to other methods for high-dimensional text matching.
- Seeking alternative methods for high-dimensional data.

Summary

- Apply Doc2Vec to convert long text documents into vector representation.
- Use Cosine Similarity to match high-dimensional text data.
- Apply high-dimensional matching method to analyzing the gender bias in citation counts of academic publications.

The End