Computational Social Science Course
(Final Report)

Conference Speaker Biography Analysis

July 2018

Submitted by

- Shiau Chu Heng
- Shide Adibi
- Chuyi Sun
- Md Kamal Hossain

Under Supervision of:

- JProf. Dr. Claudia Wagner

Table of Contents

1. Basic Methodology

- Do the self-presentation of male and female researchers differ? In what ways?

- **Data: Biographies of speakers in computer science conferences**
  - A concise self-introduction to be read by fellow academics.
  - Should contain information the speaker considers to be important about their career.
  - No unified format across conferences.

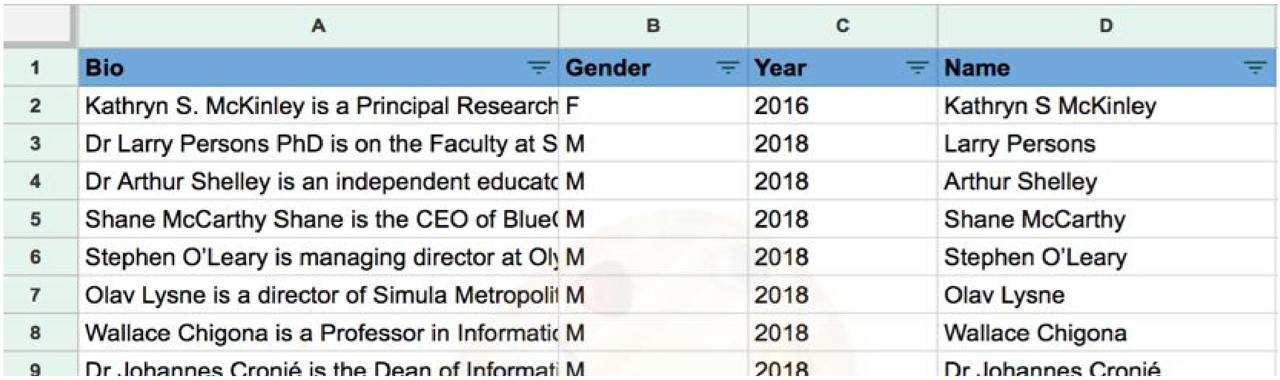
- **Metrics:**
  - **Cosine similarity:** simple way to quantify difference between texts.
  - **Term frequency:** what information gets emphasized over and over again?
  - Most research into gender representation in computer science focuses on the male-to-female ratio, and the larger societal trends that causes less women entering the field.
    - Many compiled statistical reports are also available from government organizations (Bureau of Labor etc...)

  - Most research are also focused on the computer science workforce instead of academia.

  - Much less research on how women already in computer science view their roles.
  - “Women in Technology” report:highlights the unique barriers facing women working in the technology sector.
    - Do the same barriers exist in academia?
    - Does this affect how women researchers presents themselves and how they are presented by others?

2. Data Collection

- For collecting data at the beginning we started to to write a crawler for different conference website, but then we realize not all the web sites have the same format and we must have to write a different crawler for each one of the website hence we decided to do collect our data manually. First We Collects biography for each scientist from the conference websites and then Extract all the significant information out of each conference such as **Biography**, **Gender**, **Publish year**, **Name of speaker.**



	A	B	C	D
1	Bio	Gender	Year	Name
2	Kathryn S. McKinley is a Principal Research	F	2016	Kathryn S McKinley
3	Dr Larry Persons PhD is on the Faculty at S	M	2018	Larry Persons
4	Dr Arthur Shelley is an independent educatc	M	2018	Arthur Shelley
5	Shane McCarthy Shane is the CEO of Blue(	M	2018	Shane McCarthy
6	Stephen O'Leary is managing director at Oly	M	2018	Stephen O'Leary
7	Olav Lysne is a director of Simula Metropolit	M	2018	Olav Lysne
8	Wallace Chigona is a Professor in Informatic	M	2018	Wallace Chigona
9	Dr Johannes Cronié is the Dean of Informati	M	2018	Dr Johannes Cronié

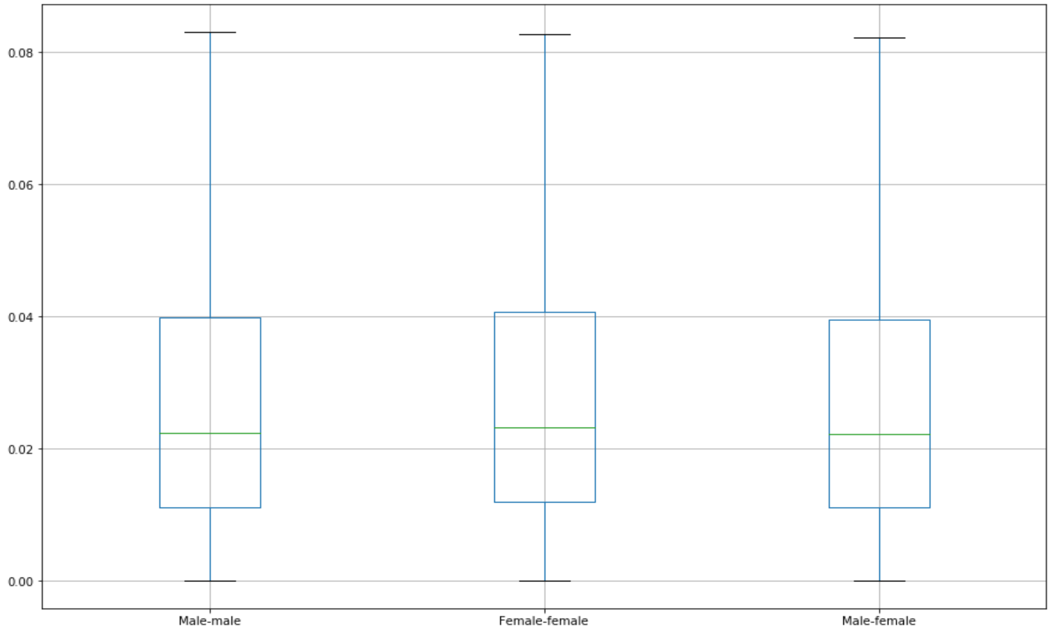
3. Preliminary Results

1. Calculated cosine similarity on a small data set of 94 biographies (74 male, 20 female), with primitive preprocessing (small list of stop words, no stemming)
2. Average cosine similarity of male-male pairs: 0.0331
3. Average cosine similarity of female-female pairs: 0.0588
   - High similarity due to small sample size?
4. Average cosine similarity of male-female pairs: 0.0255
5. There does seem to be a slight difference between male and female biographies

4. Result

4.1 Cosine Similarity

- Cosine similarity calculated over a set of 191 bios (153 male, 38 female).
- Improved preprocessing via NLTK.



- Average cosine similarity of male-male pairs: 0.031
- Average cosine similarity of female-female pairs: 0.032
- Average cosine similarity of male-female pairs: 0.030
- Similar distribution
- **No significant difference!**

4.2 Term Frequency

- Frequency of word stems in male and female bios

research	84
comput	71
scienc	66
univers	64
award	44
professor	42
engin	27
receiv	25
learn	24
data	23

Female

comput	353
research	297
univers	280
scienc	242
professor	130
system	114
award	110
algorithm	98
work	93
includ	89

Male

- Term frequency separated by part-of-speech tags (noun, adjective, verb, adverb) showed similar results: significant overlap of most common terms across genders.
- Calculated without removing stop words and stemming to ensure accuracy of tagger.

4.3 Document Frequency

- Frequency of word stems by documents in male and female bios

research	33
univers	30
professor	28
comput	27
scienc	25
receiv	19
associ	16
engin	16
fellow	15
interest	14

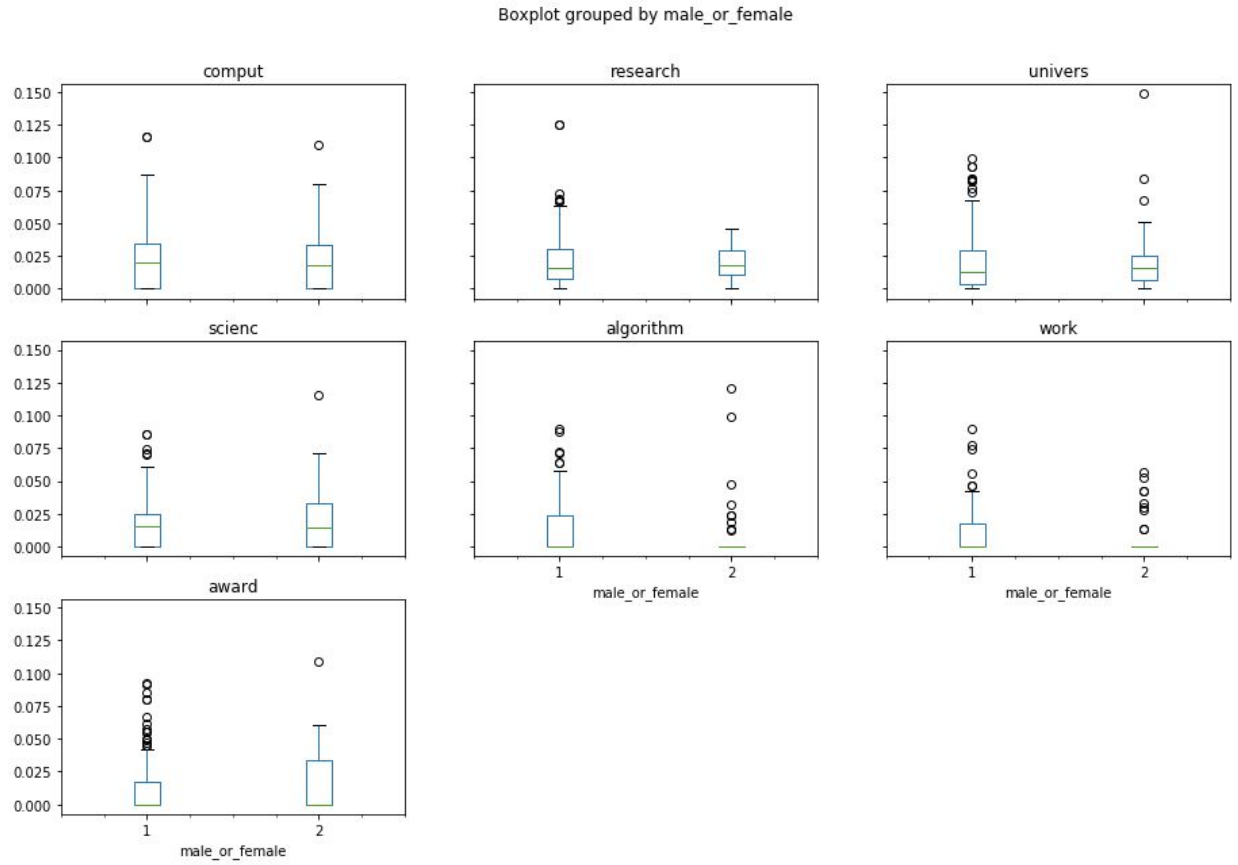
Female

research	121
univers	115
scienc	112
comput	109
professor	93
includ	65
work	60
receiv	58
institut	55
award	54

Male

- Document frequency separated by part-of-speech tags (noun, adjective, verb, adverb) showed similar results: significant overlap of most common terms across genders
- Again, calculated with minimal preprocessing

4.4 TFIDF


Boxplot grouped by male_or_female
comput
research
univers
scienc
algorithm
work
award
0.150
0.125
0.100
0.075
0.050
0.025
0.000
1
2
male_or_female

4.5 PMI(Pointwise Mutual Information)

- After extracting 4375 unique terms, We then calculated a word-by-word Pointwise Mutual Information (PMI) for each of the categories of female and male. We then compared both results in order to see if there are any differences.
- Out of around 19 million pair of terms about 3% had different PMIs comparing female and male.
- Here is the ranked list based on the difference of pair of terms.

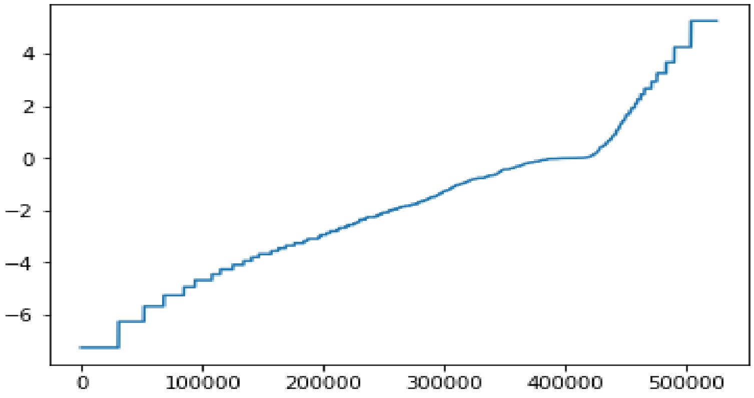
  - **Top pairs of terms representing Male class**

Term 1	Term 2
'aachen'	'cross'
'aachen'	'envoy'
'aberration'	'array'
'ababa'	'age'
'abstract'	'deeper'
'abstract'	'pure'
'academics'	'brothers'
'accenture'	'watch'
'adapted',	'serious'
address'	'bacterial'

- **Top pairs of terms representing Female class**

Term 1	Term 2
'ability',	'alzheimers'
'ability'	'anita',
'ability'	'disease'
'about'	'failure'
'about'	'smart'
'accounting'	‘assist’
'accounting'	'algorithm'
'activities'	'it'
'activities',	'leadership'

- **Graph for Pair of term with different PMIs for two class of male and female:**



5.Conclusion

- No significant difference exists in the language used by male and female computer scientists in their conference biographies.
- Related research shows large gender differences in the tech industry:
  - Academic advancement (based on blind peer reviews) more objective than metrics for climbing the “corporate ladder”?
- More significant differences across conferences than across gender?
  - Individual “conference traditions”, since no unified format across conferences exist?
- Does the result hold for older conferences?
- Would the presumably larger gender gap make this harder to calculate?
- The importance of having a large enough data set
- Format standards (or lack thereof) sometimes have a larger influence than the data they carry!
- Different preprocessing procedures for different calculations

6. Appendix

- **Group tasks for the project:**

1- Data Collection :

✓ Shide Adibi-Md Kamal Hossain-Chuyi Sun

1- Mutual Information Method & Crawler :

✓ Shide Adibi-Md Kamal Hossain

2-TFIDF Method & Crawler :

✓ Chuyi Sun

3-Cosine Similarity, Term Frequency and Document Frequency:

✓ Shiau Chu Heng