Does Location and Fame affect your citability?

Mukund Rungta, Janvijay Singh
Georgia Institute of Technology

Problem Statement

Number of Citation is the only metric capturing impact of a researcher’s work.

Importance of Citation as metric:
- Position and career of the researcher in the field
- Funding, ranking of the institute, global collaboration

Common Perception:
- Research from popular labs gets more cited
- Authors working with popular researchers gets more cited
- Overall, researcher from specific geographic location gets more cited

Proposal:
- Research in NLP is conducted across multiple countries, but only some countries get more cited.

Power Law Curve Analysis:
- x-axis: log of number of citations
- United states has the longest tail curve
- High disparity exists in the citation pattern of countries

Research Question 1

How has the citation count for countries changed over the years?

For a country- j and year-k, mean-citation is given by:

$$ MC_{lk} = \frac{1}{n} \sum_{i=1}^{n} C_l(i)_{lk} $$

here $C_l(i)$: citations of paper- i until year-k; $P_l$: all papers until year-k.

Discussion:
- Top-3 countries (US, UK & Canada) dominate the metric for past 20 years.
- Growth-rate (slope) for top-3 countries is remarkably higher compared to others.

Research Question 2

How do countries cite each other? What contributes to the higher citation count of some countries?

Score $F$ of country- j (row) & country- j (col) is given by:

$$ F_{ij} = \frac{\sum R(i)_{kj}}{\sum P_{kj}} $$

here $R(i)$: references of paper- i; $P$: all papers until 2021. Intuitively, $F(i,j)$: fraction of references from country- j in an average paper from country- k.

Discussion:
- Everyone heavily cites top-3 countries.
- Intra-country citations are primary source.

Research Question 3

Does there exist a closed group of researchers (community) that reinforces the higher citation count of some countries?

- Community (clique) : Closed group of researchers citing each other
- Clique of size $k$ is considered for the analysis

$$ S_{kj} = \frac{5^{k}(\text{size}(G_{k}))}{V(G_{k})} $$

here $G_{k}$ is subgraph with authors from country- k.

Discussion:
- US researchers form the largest number of cliques with self and researchers from other countries

Research Question 4

How do the citation statistics of different countries vary across venues? Is higher citation a side effect of a country publishing in a highly cited venue?

- Mean citations for United States, United Kingdom and Canada are mostly above the average citation for all the considered conferences.
- Even though significant number of papers are published by countries from Eastern World across all conferences, their mean citation is significantly below average citation.

Research Question 5

Is disparity in citation statistics consistent across areas of research within NLP? Or is the gap simply because some countries work in areas that receive low numbers of citations (overall)?

- Considered word bigrams from title to represent area of research
- Top 10 bigrams based on number of papers are considered for analysis
- For Sentiment analysis (100,74), dependency parsing (108,80) & relation extraction (119,85) - although number of papers for United States and China are comparable, there is huge disparity in mean citation

Dataset

- As of January 2022, ACL Anthology had 71,568 papers.
- Semantic Scholar API: get Semantic Scholar ID(SSID) for the paper’s BibTEX
- For papers whose SSID cannot be retrieved, we perform fuzzy string matching score between BibTEX’s PDFMiner
- There are 98.63% of total papers retrieved
- Citation graph: Only referenced papers

Country Extraction

- Crawled pdf of papers using PDFMiner
- Extract text before Abstract
- Use Country list, University to Country list
- Crawl papers using ACL Anthology

Count of papers based on number of affiliated countries

<table>
<thead>
<tr>
<th>Paper Title</th>
<th>Number of Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIKE: Re-training of Deep Multilingual Transformers for Language Understanding</td>
<td>2785</td>
</tr>
<tr>
<td>Hidden Global Vectors for Representa-</td>
<td>4407</td>
</tr>
<tr>
<td>tions: Open Source Toolkit for Statistical Machine</td>
<td>2805</td>
</tr>
<tr>
<td>Translation: Deep Contextualized Word Representations</td>
<td>3353</td>
</tr>
</tbody>
</table>

Top 5 papers from ACL Anthology with number of citations

- Mean citations for United States, United Kingdom and Canada are mostly above the average citation for all the considered areas of NLP.