Does Location and Fame affect your citability?

Mukund Rungta, Janvijay Singh

Georgia Institute of Technology

Problem Statement

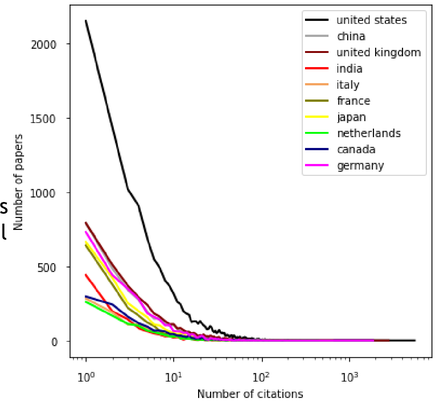
Number of Citation is the only metric capturing impact of a researcher's work.

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
- Huge disparity exists in the citation pattern of countries



Power law curve for the citation received per paper varied across countries

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 title and Semantic Scholar title
- **98.63%** of total papers retrieved
- Citation graph: Only referenced papers in ACL Anthology is considered

Country Extraction
- Crawl pdf of papers using PDFMinner
- Extract text before Abstract
- Use Country list, University to Country Map and popular cities to Country Map

Details	Number of papers
One Country	42972
More than One Country	13399
Zero Country	15197

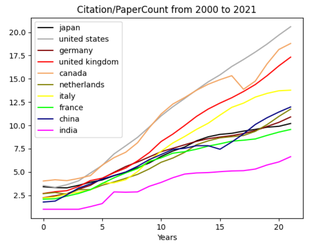
Count of papers based on number of affiliated countries

Paper Title	Number of Citations
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	7265
Bleu: a Method for Automatic Evaluation of Machine Translation	5580
GloVe: Global Vectors for Word Representation	4467
Moses: Open Source Toolkit for Statistical Machine Translation	2803
Deep Contextualized Word Representations	2353

Top 5 papers from ACL Anthology with number of citations

Research Question 1

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



Curves showing variation of mean citations (per publication) across time for different countries

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

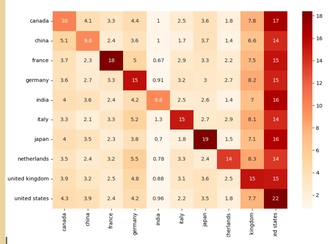
$$MC_{(j,k)} = \frac{\sum_{i \in P_k} C_k(i) 1_{i \in j}}{\sum_{i \in P_k} 1_{i \in j}}$$

here $C_k(i)$: citations of paper-*i* until year-*k*; P_k: 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?



Heat-map depicting inter-country citation trends (cells denote %age).

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

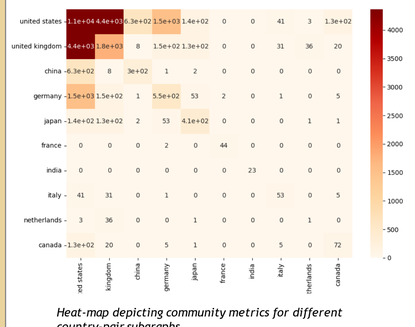
$$C_j(i) = \frac{\sum_{r \in R(i)} 1_{r \in j}}{|R(i)|} \quad F(k,j) = \frac{\sum_{i \in P} C_j(i) 1_{i \in k}}{\sum_{i \in P} 1_{i \in k}}$$

here R(i): references of paper-*i;* P: all papers until 2021. Intuitively, F(k,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?



Heat-map depicting community metrics for different country-pair subgraphs.

- Community (clique) : Closed group of researchers citing each other
- Clique of size 5 is considered for the analysis

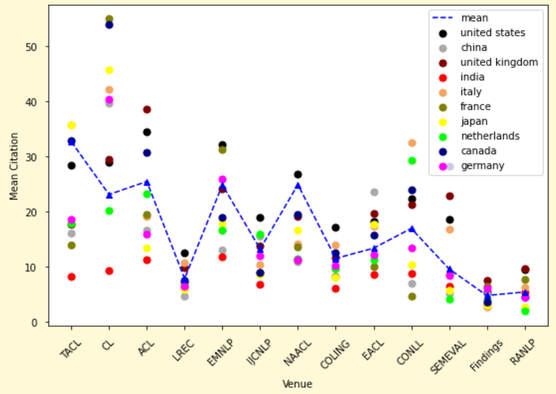
$$S(k,j) = \frac{5\text{-}clique(G_{k \cup j})}{V(G_{k \cup j})}$$

here $G_{(k \cup j)}$ is subgraph with authors from country-k and country-j.

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?

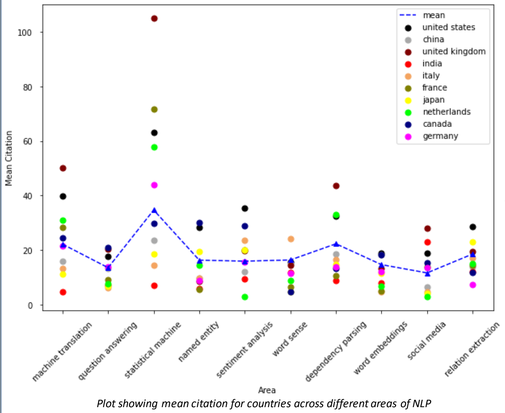


- Mean citations for United States, United Kingdom and Canada are mostly above the average citation for all the considered conferences.
- Even though significant number of papers are published by countries from Eastern World across all conferences, their mean citation is significantly below average citation.

Conference	Scope	Dominating Country
ACL	Worldwide	United States
NAACL	North America	United States
EACL	Europe	China

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)?



Plot showing mean citation for countries across different areas of NLP

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