



Big Data Analytics – Fall, 2018

Assignment 1


You make work in teams of up to 2 persons

dblp – Computer Science Bibliography

The dblp website provides a search tool for finding authors, journals/conferences/workshops, series and monographs in the area of Computer Science. The screenshot below shows the [home page](#). I would recommend using Chrome instead of IE to view.

maintained by  SCHLOSS DAGSTUHL at  Universität Trier

home | browse | search | about



dblp
computer science bibliography

[+] Welcome to dblp

[-]

> Home

■ browse authors | editors

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

■ browse journals

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z by publisher

■ browse conferences | workshops

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

■ browse series

CoRR LNCS CEUR-WS LNEE IFIP LNI EPTCS LIPICS other

■ browse monographs

books & theses reference works edited collections



[-] News and announcements

2018-07-13: ORCID state of dblp

Since more than a year now, dblp has intensified its efforts to link dblp bibliographies to ORCID iDs used by that author. ORCID information are now added regularly to the dblp data set. The primary sources for ORCID iDs are: first, the annual ORCID open data dump and, second, metadata directly provided by publishers who have started to increasingly label author signatures with ORCID information. Neither of those data sources are free of errors and data hick-ups, so we are still manually cleaning the ORCID data prior to adding them to the corpus. But overall, ORCID iDs have helped us to correct numerous cases of homonymous and synonymous bibliographies in dblp, so it is absolutely worth our time.

show more

[-] About dblp

This service provides open bibliographic information on major computer science journals and proceedings. dblp is a joint service of the  University of Trier and  Schloss Dagstuhl. For more information check out our F.A.Q.

[-] dblp statistics

■ # of publications: 4,320,213

■ # of authors: 2,160,223

■ # of conferences: 5,511

■ # of journals: 1,593

■ % of publication types in dblp:

Conference papers: 51.1%

Journal articles: 31.1%

Books: 1.1%

Other: 16.7%

more statistics

[-] dblp tweets


load tweets from twitter.com

1

The screenshot below shows the results for searching on *Boetticher*.

maintained by SCHLOSS DAGSTUHL at Universität Trier

home | browse | search | about



dblp
computer science bibliography

Boetticher

[+] Search dblp

[+] powered by CompleteSearch, courtesy of Hannah Bast, University of Freiburg

> Home

[+] Author search results




Likely matches

- Gary D. Boetticher
- H. v. Boetticher
- Nicolai von Bötticher




[+] Publication search results

found 13 matches

2014

   Gary D. Boetticher:
Engineering Financial Engineering. CIFEr 2014: 231-238

2010

   Gary D. Boetticher, Günther Ruhe:
Special Section on Best Papers PROMISE 2009. Information & Software Technology 52(11): 1229 (2010)

2009

[+] Refine list

refine by author
Gary D. Boetticher (11)
Günther Ruhe (2)
Nicolai von Bötticher (1)
I. Schäfer (1)
Thomas J. Ostrand (1)
Lotfi A. Zadeh (1)
Hans-Heinrich Bothe (1)
Shu-Ching Chen (1)
Stuart Harvey Rubin (1)
W. Hoffmann (1)
17 more options

Goal

For this assignment, you will apply the MapReduce algorithm to the Citation Network Dataset (A dblp dataset.).

Tasks

- 1) Go to the [Citation Network Dataset](#) website and review the details of the *DBLP-Citation-network V10* dataset in terms of number of records and attributes. Please note, there are many versions. We will be working with version 10. Please download version 10. The zip file is about 1.8 gigabytes. I have the zip file on a pen drive. If you want, see me during office hours or at break in class to get a copy of the zip file. Within the zip file is a folder which contains 4 json files (dblp-ref-0.json, dblp-ref-1.json, dblp-ref-2.json, and dblp-ref-3.json). The first 3 json files are about 1.4 gigabytes. The last one is 93 megabytes.
- 2) Included in the homework assignment is a very small json file called *dblp-ref-veryshort.json*. For this task make sure you can read the attributes correctly. To verify, print out the records. The data scheme is as follows:

Field Name	Field Type	Description	Example
id	string	paper ID	013ea675-bb58-42f8-a423-f5534546b2b1
title	string	paper title	Prediction of consensus binding mode geometries for related chemical series of positive allosteric modulators of adenosine and muscarinic acetylcholine receptors
authors	list of strings	paper authors	["Leon A. Sakkal", "Kyle Z. Rajkowski", "Roger S. Armen"]
venue	string	paper venue	Journal of Computational Chemistry
year	int	published year	2017
n_citation	int	citation number	0
references	list of strings	citing papers' ID	["4f4f200c-0764-4fef-9718-b8bccf303dba", "aa699fbf-fabe-40e4-bd68-46eaf333f7b1"]
abstract	string	abstract	This paper studies ...

Below is a sample record:

```
{
  "authors": [
    "Leon A. Sakkal",
    "Kyle Z. Rajkowski",
    "Roger S. Armen"
  ],
  "n_citation": 0,
  "references": [
    "4f4f200c-0764-4fef-9718-b8bccf303dba",
    "aa699fbf-fabe-40e4-bd68-46eaf333f7b1"
  ],
  "title": "Prediction of consensus binding mode geometries for related chemical
series of positive allosteric modulators of adenosine and muscarinic acetylcholine
receptors",
  "venue": "Journal of Computational Chemistry",
  "year": 2017,
  "id": "013ea675-bb58-42f8-a423-f5534546b2b1"
}
```

- 3) **References and citations.** In the sample record on the previous page. The paper has 3 authors who reference 2 papers. We may view this as follows:

Original Paper ID	→	Referenced Paper ID
013ea675-bb58-42f8-a423-f5534546b2b1	→	4f4f200c-0764-4fef-9718-b8bccf303dba
013ea675-bb58-42f8-a423-f5534546b2b1	→	aa699fbf-fabe-40e4-bd68-46eaf333f7b1

It is possible to “invert” the relationship. We call this **cited by**.

Paper ID	→	“Cited by” Paper ID
4f4f200c-0764-4fef-9718-b8bccf303dba	→	013ea675-bb58-42f8-a423-f5534546b2b1
aa699fbf-fabe-40e4-bd68-46eaf333f7b1	→	013ea675-bb58-42f8-a423-f5534546b2b1

Write a Python (or Java) program (or script) to write out two columns. The paperID along with the CitedBy PaperID for the *dblp-ref-veryshort.json* dataset.

- 4) **MapReduce problem 1.** For any given paper, we want to know how many times it was cited. Initially, apply your code to the *dblp-ref-veryshort.json* dataset. I have modified the data so that it does give back some results. Your output should include a paperID and a count of the number of times it was cited.

Modify your code so that it returns the most influential paper (which is the one most cited) only.

- 5) Modify your program to determine the 5 most influential papers. These are papers that are most cited.

Run this program against the *dblp-ref-veryshort.json* dataset.

Next, run this program against the *dblp-ref-3.json* dataset (See step 1).

Bonus: See if you can run it against all 4 data sets from step 1.

- 6) **MapReduce problem 2.** Write a mapReduce program that will determine how many papers an author wrote. Any paper may have one or more authors. In this case, each author gets credit for having written one paper. Run this against the *dblp-ref-veryshort.json* dataset.

We will ignore the following two issues:

- A) An author’s name may vary from paper to paper. For example, G. Boetticher, Gary Boetticher, or GD Boetticher may refer to the same author. Count this as 3 separate authors.
- B) There may be two or more authors with the exact same name. To simplify the problem, assume that it is always the same person.

- 7) Modify your program to determine the 5 most influential authors. These are authors that have written the most papers.

Run this program against the *dblp-ref-veryshort.json* dataset.

Next, run this program against the *dblp-ref-3.json* dataset (See step 1).

Bonus: See if you can run it against all 4 data sets from step 1.

- 8) **MapReduce problem 3.** MapReduce problem 1 determined the 5 most influential papers. The MapReduce problem 2 determined the 5 most influential authors. This problem builds on those first two mapReduce problems. For each author, determine the average citation of their 3 was most influential papers. Output the top ten authors (the ones with the highest citation average).

Run this program against the *dblp-ref-veryshort.json* dataset.

Next, run this program against the *dblp-ref-3.json* dataset (See step 1).

Bonus: See if you can run it against all 4 data sets from step 1.

Deliverables

- 1) Use the following naming convention for the zip file:
 - If you work alone:
Last name_FirstName_HW1.ZIP
 - If you work as a group of two:
LastName1FirstName1LastName2FirstName2_HW3.ZIP
 - For example: Rajiv Gandhi and Shriya Saran would be:
GandhiRajivSaranShriya_HW1.ZIP
- 2) If you work in a group, make sure to CC your partner in your submission.
- 3) If you work in a group, make only one submission.
- 4) What to place in the zip file?
 - A) Documented Java or Python code.
 - B) A ReadMe file that explains how to run your code.
 - C) An MS-Word or PDF file that contains screenshots showing your results for most influential paper, most influential author, and highest citation average. Also, specify the dataset used for each screenshot. Include your names at the beginning of the document.
 - D) Do not include any datasets.**

Due Date: Wednesday, October 24th, 7 PM via email.

Email it to boetticher@uhcl.edu

To make life simple, do not email your solution to the TA.