

KORONA – Overview

1. Splitting of DBLP-data

Create directory „nt-files“

```
split -l 100000 dblp-2017-04-18.nt nt-files/
for file in *; do mv "$file" "${file%}.nt"; done
```

Input

DBLP NT-Triples dump file

dblp-2017-04-18.nt

```
<http://dblp.org/rec/journals/amco/WangG13>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dblp.org/rdf/schema-2017-04-18#Publication> .
[...]
```

Output

620 split NT-Triples files containing max. 100,000 lines of the original file

nt-files/...nt

```
<http://dblp.org/rec/journals/amco/WangG13>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dblp.org/rdf/schema-2017-04-18#Publication> .
[...]
```

2. Filtering and reduction of DBLP-data

Install libraries nose / tornado / rdflib / openpyxl

```
sudo python 1.\ filter-nt.py
```

Input

620 split NT-Triples files

nt-files/...nt

```
<http://dblp.org/rec/journals/amco/WangG13>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dblp.org/rdf/schema-2017-04-18#Publication> .
[...]
```

Output

A single accumulated NT-Triples file containing only triples with the subject prefix

ISWC.nt

“http://dblp.org/rec/conf/semweb/”

```
<http://dblp.org/rec/conf/semweb/0001CDB0VA16>
  <http://dblp.org/rdf/schema-2017-04-18#publishedInBook>
    "International Semantic Web Conference (2)" .
[...] (44855 Lines)
```

3. Feature selection

```
sudo python 2.\ rdflib2excel.py
```

Input

NT-Triples file

ISWC.nt

```
<http://dblp.org/rec/conf/semweb/0001CDB0VA16>  
  <http://dblp.org/rdf/schema-2017-04-18#publishedInBook>  
    "International Semantic Web Conference (2)"    .
```

[...]

Output

Excel spreadsheet containing information filtered on the predicates title of paper, author name, and year of publication. Each row represents one paper

metis.xlsx

[Paper Number] [Title] [Number of Authors] [Year]

```
[1] [TripleWave: Spreading RDF Streams on the Web.] [7] [2016]  
[http://dblp.org/pers/c/Calbimonte:Jean=Paul]  
[http://dblp.org/pers/d/Dell=Aglio:Daniele]  
[http://dblp.org/pers/b/Brambilla_0001:Marco]  
[http://dblp.org/pers/a/Aberer:Karl]  
[http://dblp.org/pers/v/Valle:Emanuele_Della]  
[http://dblp.org/pers/b/Balduini:Marco]  
[http://dblp.org/pers/m/Mauri_0001:Andrea]
```

[...] (3139 lines)

4. Generation of Conference similarity matrix and bipartite graph

Install library bs4 and create directory "output"

Replace wb.get_active_sheet() function in source code with wb.active

```
sudo python 3.\ similarities.py
```

Input

Excel spreadsheet

metis.xlsx

[Paper Number] [Title] [Number of Authors] [Year]

```
[1] [TripleWave: Spreading RDF Streams on the Web.] [7] [2016]  
[http://dblp.org/pers/c/Calbimonte:Jean=Paul]  
[http://dblp.org/pers/d/Dell=Aglio:Daniele]  
[http://dblp.org/pers/b/Brambilla_0001:Marco]  
[http://dblp.org/pers/a/Aberer:Karl]  
[http://dblp.org/pers/v/Valle:Emanuele_Della]  
[http://dblp.org/pers/b/Balduini:Marco]  
[http://dblp.org/pers/m/Mauri_0001:Andrea]
```

[...]

Output

Output file for indexing authors

output/author-key-map.txt

```
A1    http://dblp.org/pers/c/Calbimonte:Jean=Paul  
A2    http://dblp.org/pers/d/Dell=Aglio:Daniele  
[...] (4918 lines)
```

List of authors	output/author-list.txt
http://dblp.org/pers/c/Calbimonte:Jean=Paul http://dblp.org/pers/d/Dell=Aglione:Daniele [...] (4918 lines)	
Author vertices file	output/Author.txt
4918 A1 A2 [...] (4919 lines)	
Conference vertices file	output/Conf.txt
16 C2001 C2002 [...] (17 lines)	
Conference similarity matrix file	output/Conf_matrix.txt
16 1.0 0.128205128205 0.0637254901961 0.0524861878453 0.0498866213152 0.0329457364341 0.0314569536424 0.021613832853 0.0289115646259 0.0267295597484 0.0201863354037 0.0132352941176 0.0147895335609 0.0126467931346 0.0147213459516 0.0143027413588 [...] (17 lines)	
Bipartite graph with weighted edges from authors to conferences (matrix)	output/Auth-Conf_graph.txt
8214 A1 C2010 edge 0.0714285714286 [...] (8215 lines)	

5. Conversion of author URIs to names

Install library lxml

Create directory "Problems" and file "404.txt" in this directory

```
sudo python 5.\ authornam4mlink.py
```

Input

File for indexing authors with URIs	output/author-key-map.txt
-------------------------------------	---------------------------

A1 http://dblp.org/pers/c/Calbimonte:Jean=Paul
A2 http://dblp.org/pers/d/Dell=Aglione:Daniele
[...]

Output

File for indexing authors with names	output/author-key-map.txt
--------------------------------------	---------------------------

A1 Marco Brambilla
A2 Daniele Dell'Aglione
[...]

File with URIs	output/Problems/404.txt
----------------	-------------------------

http://dblp.org/pers/c/Cecconi:Cecile
http://dblp.org/pers/l/Lefort:Laurent
http://dblp.org/pers/k/Kontokostas:Dimitris
[...] (56 lines)

6. Identification of duplicates

```
sudo python3 find-duplicates.py
```

Input

File for indexing authors	output/author-key-map.txt
---------------------------	---------------------------

A1 <http://dblp.org/pers/c/Calbimonte:Jean=Paul>

A2 <http://dblp.org/pers/d/Dell=Aglio:Daniele>

[...]

List of authors	output/author-list.txt
-----------------	------------------------

<http://dblp.org/pers/c/Calbimonte:Jean=Paul>

<http://dblp.org/pers/d/Dell=Aglio:Daniele>

[...]

Output

File containing authors occurring more than once	output/Problems/duplicate_links.txt
--------------------------------------------------	-------------------------------------

1.Duplicated Author Number

Author Name: Maarten Menken

Duplicated Links:

http://dblp.org/pers/m/Menken:Maarten_R=

<http://dblp.org/pers/m/Menken:Maarten>

File containing URIs corresponding to duplicate author names	output/de-duplicate.txt
--------------------------------------------------------------	-------------------------

http://dblp.org/pers/m/Menken:Maarten_R=

<http://dblp.org/pers/m/Menken:Maarten>

7. Elimination of duplicates

```
sudo python3 remove_duplicates.py
```

Input

File containing URIs corresponding to duplicate author names	output/de-duplicate.txt
--------------------------------------------------------------	-------------------------

http://dblp.org/pers/m/Menken:Maarten_R=

<http://dblp.org/pers/m/Menken:Maarten>

NT-Triples file	ISWC.nt
-----------------	---------

<<http://dblp.org/rec/conf/semweb/0001CDB0VA16>>

<<http://dblp.org/rdf/schema-2017-04-18#publishedInBook>>

"International Semantic Web Conference (2)" .

[...]

Output

NT-Triples file without duplicate authors	output/Filtered-ISWC.nt
-------------------------------------------	-------------------------

<<http://dblp.org/rec/conf/semweb/0001CDB0VA16>>

<<http://dblp.org/rdf/schema-2017-04-18#publishedInBook>>

"International Semantic Web Conference (2)" .

[...]

8. Replace URIs for duplicate authors in NT-Files

```
javac RemoveDuplicatesNTFiles.java
java RemoveDuplicatesNTFiles
```

Input

620 NT-Triples files	nt-files/...nt
----------------------	----------------

```
<http://dblp.org/rec/journals/amco/WangG13>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dblp.org/rdf/schema-2017-04-18#Publication> .
[...]
```

List of URIs for duplicate authors	output/de-duplicate.txt
------------------------------------	-------------------------

```
http://dblp.org/pers/m/Menken:Maarten_R=
http://dblp.org/pers/m/Menken:Maarten
```

Output

620 NT-Triples files with redundant URIs replaced	f_nt-files/...nt
---------------------------------------------------	------------------

```
<http://dblp.org/rec/journals/amco/WangG13>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dblp.org/rdf/schema-2017-04-18#Publication> .
[...]
```

Repeat steps 3 to 5 with new ISWC-file.

9. Generation of Author similarity matrix

Remove DBLP from the path in the source code file

```
sudo python 4.\ author\ similarity.py
```

Input

620 split NT-Triples files	nt-files/...nt
----------------------------	----------------

```
<http://dblp.org/rec/journals/amco/WangG13>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dblp.org/rdf/schema-2017-04-18#Publication> .
[...]
```

List of authors	output/author-list.txt
-----------------	------------------------

```
http://dblp.org/pers/c/Calbimonte:Jean=Paul
http://dblp.org/pers/d/Dell=Aglio:Daniele
[...] (4918 lines)
```

Output

Author similarity matrix file	output/Auth_matrix.txt
-------------------------------	------------------------

```
4918
1.0 0.352112676056 0.0289256198347 0.0543293718166 0.180124223602 [...]
[...] (4919 lines)
```

10. Calculation of percentiles

```
sudo python 6.\ get_percentiles.py output/Conf_matrix.txt  
sudo python 6.\ get_percentiles.py output/Auth_matrix.txt
```

Input

Conference similarity matrix	output/Conf_matrix.txt
------------------------------	------------------------

16

```
1.0 0.128205128205 0.0637254901961 0.0524861878453 0.0498866213152  
0.0329457364341 0.0314569536424 0.021613832853 0.0289115646259  
0.0267295597484 0.0201863354037 0.0132352941176 0.0147895335609  
0.0126467931346 0.0147213459516 0.0143027413588
```

[...] (17 lines)

Output

Min: 0.0108

Max: 0.1743

Average: 0.0691

Median: 0.0616

Percentile	Similarity
------------	------------

10	0.0199
15	0.0266
20	0.0296
25	0.0317
30	0.0401
35	0.0496
40	0.0513
45	0.0554
50	0.0616
55	0.0673
60	0.0717
65	0.0807
70	0.0866
75	0.0981
80	0.1057
85	0.1211
90	0.1290
95	0.1479
98	0.1586

Input

Author similarity matrix	output/Auth_matrix.txt
--------------------------	------------------------

4918

```
1.0 0.352112676056 0.0289256198347 0.0543293718166 0.180124223602 [...]
```

[...] (4919 lines)

Output

Min: 0.0007

Max: 1.0000

Average: 0.0697

Median: 0.0396

Percentile	Similarity
------------	------------

10	0.0105
----	--------

15	0.0137
20	0.0169
25	0.0202
30	0.0237
35	0.0272
40	0.0311
45	0.0351
50	0.0396
55	0.0444
60	0.0500
65	0.0571
70	0.0652
75	0.0750
80	0.0882
85	0.1061
90	0.1379
95	0.2000
98	0.3333

11. Application of semEP

```
./semEP -p <-l left threshold> <-r right threshold>
testdblp/Auth_matrix.txt testdblp/Author.txt
testdblp/Conf_matrix.txt testdblp/Conf.txt testdblp/Auth-
Conf_graph.txt
```

Output

Folder containing computed clusters	nr_drug-target_graph-0.3061-0.1614-Clusters
-------------------------------------	---------------------------------------------

```
[...]
A1853 C2011 0.0714      edge
A2188 C2011 0.0714      edge
A2185 C2011 0.0714      edge
A2186 C2011 0.0714      edge
A2189 C2011 0.0714      edge
[...] (different numbers of lines)
```

Text file containing predictions	nr_drug-target_graph-0.3061-0.1614-Predictions
----------------------------------	------------------------------------------------

```
Cluster      1051
A218  C2015 0.5000
A2325 C2014 0.5000
Cluster      1056
A245  C2015 0.5000
A1431 C2016 0.5000
Cluster      1061
A266  C2015 0.5000
A3898 C2014 0.5000
[...] (3008 lines)
```

12. Generation of similarities matrix

Create directory "simrelations"

```
sudo python 7.\ sim_matrix_with_rel_constraints.py <threshold_1>  
<threshold_2> output/Auth_matrix.txt output/Author.txt  
output/Conf_matrix.txt output/Conf.txt output/Auth-Conf_graph.txt  
simrelations/<output_file>
```

Output

Text file containing the matrix with similarities between all pairs or relations

output.txt

```
1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0, [...]  
0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0, [...]  
0.0,0.0,1.0,0.151231945624,0.0,0.0,0.0,0.352112676056, [...]  
[...] (8214 lines + columns)
```

13. Computation of clustering measures

```
./cma ../<semEP clusters directory> ../output/Auth-Conf_graph.txt  
../simrelations/<simrel_file>
```

Output

```
Starting the application  
Cluster files folder: Auth-Conf_graph-0.2000-0.1479-Clusters  
Number of cluster: 3291  
Number of edges: 8214  
Similarity matrix loaded!  
Computing measures.....
```

```
*****
```

```
Clustering measures
```

```
*****
```

```
#Cluster Conductance
```

```
0 0.000000000000
```

```
Starting the application
```

```
Cluster files folder: Auth-Conf_graph-0.2000-0.1479-Clusters
```

```
Number of cluster: 3291
```

```
Number of edges: 8214
```

```
Similarity matrix loaded!
```

```
Computing measures.....
```

```
*****
```

```
Clustering measures
```

```
*****
```

```
#Cluster Conductance
```

```
0 0.000000000000
```

```
[...]
```

```
1228 0.896971921922  
1229 0.306936798062  
1230 1.000000000000  
1231 0.333950046254  
1232 0.357992311410  
1233 0.000000000000
```



```
[...]

3288 0.864894706763
3289 0.959921001461
3290 0.934109856227
*****
Max conductance: 1.000000000000
Min conductance: 0.000000000000
Average conductance: 0.523881036852
Coverage: 0.109763276452
Modularity: 0.099878732594
Total cut: 48010.683951266088
*****

Total time 18.037 secs
```

14. Generation of METIS graph

```
sudo python3 10.generate_metis_graph.py <number of columns sim-
matrix> <similarity matrix of relations> <output file name>
```

Input

Text file containing the matrix with similarities between all pairs of relations	simrel.txt
----------------------------------------------------------------------------------	-------------------

```
1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0, [...]  
0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0, [...]  
0.0,0.0,1.0,0.151231945624,0.0,0.0,0.0,0.352112676056, [...]  
[...] (8214 lines + columns)
```

Output

Text file containing the METIS graph	metisgraph.txt
--------------------------------------	-----------------------

```
8214 298122 001  
74 339 75 2177 160 2178 305 455 306 2921 326 344 327 2209 [...]  
[...] (8215 lines)
```

15. Application of METIS

In folder /metisinstall/bin

```
./gpmmetis <filename> <nparts>  
./gpmmetis ../graphs/memis85.txt 1391
```

Input

File description	filename
File content	

Output

File description	metis85.txt.part.1391
File content	

16. Convert METIS-output to semEP-output

Create folder "metis2semep/85/"

```
sudo python 11.\ metis2semEP.py output/Auth-Conf_graph.txt  
graphs/memis85.txt.part.1391 metis2semep/85/
```

Input

File description	filename
------------------	----------

File content

Output

File description	filename
------------------	----------

File content

17. Filter clusters for visualization

```
sudo python3 12.\ Filter-visualization.py clusters/Clusters98/  
filtervis_semep/98/  
sudo pyhton3 12.\ Filter-visualization.py metis2semep/85/  
filtervis_memis/85/
```

Input

File description	filename
------------------	----------

File content

Output

File description	filename
------------------	----------

File content

18. Generation of predictions

Move cluster-files to "output/author-clusters/"

```
sudo pip3 install openpyxl
```

```
sudo python3 13.\ Filter-predictions.py output/author-key-map.txt  
output/author-list.txt output/author-clusters/
```

Input

File description	filename
------------------	----------

File content

Output

File description	author-clusters- predictions.txt
------------------	-------------------------------------

File content

19. Verification of predictions

Change path to prediction-file in source code file + correct split statement

```
python3 14.\ Verify-prediction.py
```

Input

File description	filename
------------------	----------

File content

Output

File description	filename
------------------	----------

File content
