The ACL Anthology Network Release 2014

We are happy to announce that the 2014 release is now available for download.

For more information, please visit:

http://tangra.cs.yale.edu/newaan

Contents of this README:

---

1. USAGE INSTRUCTIONS

To use this data, please follow the following guidlines:

1. For research only.
2. Do not re-distribute.
3. If you decide to use this work in your publication, Please cite one of the following papers.

```
@article{,
       year = {2013},
       issn = {1574-020X},
       journal = {Language Resources and Evaluation},
       doi = {10.1007/s10579-012-9211-2},
       title = {The ACL anthology network corpus},
       url = {http://dx.doi.org/10.1007/
s10579-012-9211-2},
       publisher = {Springer Netherlands},
       keywords = {ACL Anthology Network; Bibliometrics;
Scientometrics; Citation analysis; Citation summaries},
       author = {Radev, DragomirR. and Muthukrishnan,
Pradeep and Qazvinian, Vahed and Abu-Jbara, Amjad},
       pages = {1-26},
       language = {English}
}


@techreport{Joseph&Radev07,
   author = {"Joseph, Mark T. and Radev, Dragomir R.",
```

```
              title = {"Citation Analysis, Centrality, and the ACL
Anthology"},
              institution = {"University of Michigan. Department of
Electrical Engineering and Computer Science"},
              pdf = {"http://clair.si.umich.edu/\~radev/papers/
csetr535-07.pdf"},
              postscript = {"http://clair.si.umich.edu/\~radev/
papers/csetr535-07.ps"},
              year =  {"2007"},
              number = {"CSE-TR-535-07"},
              x-category = {"CLAIR,RADEV,MISC"}
           }

           @inproceedings{Radev&al.09a,
              author = {Radev, Dragomir R. and Muthukrishnan, Pradeep
and Qazvinian, Vahed},
              title = {The {ACL} Anthology Network Corpus},
              year = {"2009"},
              address = {"Singapore"},
              booktitle = {"Proceedings, ACL Workshop on Natural
Language Processing and Information Retrieval for Digital
Libraries"},
              x-category ="CLAIR,RADEV,CONFERENCE"
           }


           @article{Radev&al.09b,
              author = {Dragomir R. Radev, Mark Thomas Joseph, Bryan
Gibson, Pradeep Muthukrishnan},
              year   = "2009",
              title  = {{A} {B}ibliometric and {N}etwork {A}nalysis
of the field of {C}omputational {L}inguistics},
              journal = {Journal of the American Society for
Information Science and Technology},
              publisher = {John Wiley & Sons},
              pdf="http://tangra.si.umich.edu/\~radev/papers/
biblio.pdf",
              x-category = "CLAIR,RADEV,JOURNAL"
           }
```

   4. Please inform us if you publish as we are interested in the
   output of this work.

---

2. FILES INCLUDED

All the statistics and metadata are stored in release/2014/ The exact command used for creating the statistics is also mentioned below. More details about the scripts and how to use them is given in Section 3.

acl-metadata.txt
Contains the metadata associated with each paper id. The metadata associated with every paper consists of the paper id, title, year, venue.

论文ID，标题，年代，发表的地点

acl.txt
This is the paper citation network formatted as "paper_id1 ==> paper_id2".
This file consists of all the citations in AAN.

论文引用网络

The above two files are the only canonical files. From the above two files, we have created the different networks and statistics using in-house scripts.

作者之间的引用

author_citations.txt
Contains the number of citations for every author

COMMAND: bin/aan_author_citations.pl

author_citations_nonself.txt
Contains the number of citations for every author excluding self citations.

每一个作者引用的除了自己的作者数量

COMMAND: bin/aan_author_citations.pl --nonself

authorhindex.txt
Contains the H-Index score of every author

?

COMMAND: bin/aan_hindex.pl

authorhindex_nonself.txt
Contains the H-Index score of every author excluding self citations

COMMAND: bin/aan_hindex.pl --nonself

paper_citations.txt
Contains the number of citations for every paper

COMMAND: bin/aan_paper_citations.pl

paper_citations_nonself.txt
Contains the number of citations for every paper excluding self citations

COMMAND: bin/aan_paper_citations.pl --nonself

author_collaborations.txt
Contains the number of collaborations for every author.

COMMAND: bin/aan_author_collaborations.pl

author_citation_network_stats.txt
Contains some basic statistics of the author citation network.

COMMAND: bin/aan_network_stats.pl -input="acit" --stats

author_collaboration_stats.txt
Contains some basic statistics of the author collaboration
network.

COMMAND: bin/aan_network_stats.pl -input="acoll" --stats

paper_citation_network_stats.txt
Contains some basic statistics of the paper citation network.

COMMAND: bin/aan_network_stats.pl -input="pcit" --stats

paper_pageranks.txt
Contains the PageRank scores of every paper. The PageRank scores
were
computed using the paper citation network.

COMMAND: bin/aan_pageranks.pl -input="pcit"

author_pageranks.txt
Contains the PageRank scores of every author. The PageRank scores
were
computed using the author citation network.

COMMAND: bin/aan_pageranks.pl -input="acit"

author-citation-network.txt.*-centrality
Contains the betweeness,degree and closeness centrality scores for
every author based on the author citation network.

COMMAND: bin/aan_network_stats.pl --input="acit"
--degree-centrality --betweenness-centrality --closeness-centrality

author-collaboration-network.txt.*-centrality
Contains the betweeness, degree and closeness centrality scores for
every author based on the author collaboration network.

COMMAND: bin/aan_network_stats.pl --input="acoll"
--degree-centrality --betweenness-centrality --closeness-centrality

paper-citation-network.txt.*-centrality
Contains the betweeness, degree and closeness centrality scores for
every paper based on the paper citation network.

COMMAND: bin/aan_network_stats.pl --input="pcit"
--degree-centrality --betweenness-centrality --closeness-centrality

There are five different networks stored in /release/2011/networks/

1. paper-citation-network.txt
Paper Citation Network

COMMAND: bin/aan_make_paper_citations.pl

2. paper-citation-network-nonself.txt
Paper Citation Network excluding self citations

COMMAND: bin/aan_make_paper_citations.pl --nonself

3. author-citation-network.txt
Author Citation Network

COMMAND: bin/aan_make_author_citation.pl

4. author-citation-network-nonself.txt
Author Citation Network excluding self citations

COMMAND: bin/aan_make_author_citation.pl -nonself

5. author-collaboration-network.txt
Author Collaboration Network

COMMAND: bin/aan_make_author_collaboration.pl

All the networks are formatted using the Edgelist format, which lists
a single edge per line. An edge is formatted as
"Node1_label ==> Node2_label".

The AAN corpus includes five networks, paper citation, paper citation
network without self citations, author citation, author citation
network without self citations and author collaboration. The paper citation network (paper-citation-network.txt) is a directed network composed of nodes labeled with paper ids which correspond to individual papers (acl-metadata.txt). The author citation network (author-citation-network.txt), a directed network, is compiled from the paper network and the metadata file. For each citation in the paper network, where paper A cites paper B, and for each author in paper A, an edge is created for that author to each author in paper B.
Both the paper citation network and the author citation network have a
nonself version, i.e, the self citations are excluded. If paper A cites paper B and there is a common author between the two papers, then this citation is termed as a self citation. The author collaboration network (author-collaboration-network.txt), an undirected network, is composed of authors where, for each paper in the paper citation network, an edge is created between each collaborator for that paper.

---

3. SCRIPTS

There are a few scripts which compute the different networks, network
statistics, etc in bin/

1. aan_author_citations.pl

Outputs the number of citations for every author in AAN.

Usage: bin/aan_author_citations.pl [-year=to_year] [-incites]
[-outcites] [-nonself] [-help]

-year=to_year
when specified, only citations which are older than the year mentioned
are included. Can be any year greater than 1965, defaults to 2011.

-incites
prints out the number of incoming citations for every author
in the author citation network. By default it prints out the number
of incoming citations.

-outcites
prints out the number of outgoing citations for every author
in the author citation network

-nonself
when specified, self citations are excluded. By default self citations
are NOT excluded.

-help
prints out the different options available

Example: bin/aan_author_citations.pl -year=2011

2. aan_author_collaborations.pl

Outputs the number of collaborations for every author in AAN.

Usage: bin/aan_author_collaborations.pl [-year=to_year] [-help]

-year=to_year
when specified, only citations which are older than the year mentioned
are included. Can be any year greater than 1965, defaults to 2011.

-help
prints out the different options available

Example: bin/aan_author_collaborations.pl -year=2011

3. aan_hindex.pl

Outputs the H-Index for every author in AAN.

Usage: bin/aan_hindex.pl [-year=to_year] [-nonself] [-help]

-year=to_year
when specified, only citations which are older than the year mentioned
are included. Can be any year greater than 1965, defaults to 2011.

-nonself when specified, self citations are excluded. By default self citations are included.

-help
prints out the different options available

Example: bin/aan_hindex.pl

4. aan_make_author_citation.pl

Outputs the author citation graph.

Usage: bin/aan_make_author_citation.pl [-year=to_year] [-nonself] [-help]

-year=to_year
when specified, only citations which are older than the year mentioned
are included. Can be any year greater than 1965, defaults to 2011.

-nonself
when specified, self citations are excluded. By default self citations
are included.
-help
prints out the different options available

Example: bin/aan_make_author_citation.pl

5. aan_make_author_collaboration.pl

Outputs the author collaboration graph.

Usage: bin/aan_make_author_collaboration.pl [-year=to_year] [-help]

-year=to_year

when specified, only citations which are older than the year mentioned
are included. Can be any year greater than 1965, defaults to 2011.

-help prints out the different options available

Example: bin/aan_make_author_collaboration.pl

6. aan_network_stats.pl

Outputs network statistics about the network specified.

Usage: bin/aan_network_stats.pl -i=acit|acoll|pcit
[--delimout=output_delimiter] [-output=output_file]
[-pajek=pajek_file] [-stats] [-graphml=graphml_file]
[-sample=sample_size] [-sampletype=sample_type] [-extract]
[-components] [-undirected] [-paths] [-wcc] [-cc] [-scc] [-
triangles]
[-assortativity] [-verbose] [-localcc] [-all] [betweenness-
centrality]
[-degree-centrality] [-closeness-centrality] [-lexrank-centrality]
[-force] [graph-class=graph_class] [-filebased] [-help]

--input=acit|acoll|pcit

Input network

--delimout output_delimiter Vertices in output are delimited by

delimiter (can be printf format string)

--sample sample_size

Calculate statistics for a sample of the network The sample_size
parameter is interpreted differently for each sampling algorithm

--sampletype sampletype

Change the sampling algorithm, one of:

randomnode, randomedge, forestfire

        randomnode: Pick sample_size
        nodes randomly from the original network

        randomedge: Pick sample_size edges randomly from the
original
        network

        forestfire: Pick sample_size nodes randomly from the
original
        network using ForestFire sampling (see the tutorial for
more
        information) By default uses random edge sampling

--output out_file

If the network is modified (sampled, etc.) you can optionally write
it
out to another file

--pajek pajek_file Write output in Pajek compatible format

--extract, -e
Extract largest connected component before analyzing.

--undirected, -u
Treat graph as an undirected graph

--scc
Print strongly connected components

--wcc
Print weakly connected components

--components
Print components (for undirected graph)

--paths, -p
Print shortest path matrix for all vertices

--triangles, -t
Print all triangles in graph

-assortativity, -a
Print the network assortativty coefficient

--localcc, -l
Print the local clustering coefficient of each
vertex

--degree-centrality
Print the degree centrality of each vertex

--closeness-centrality
Print the closeness centrality of each vertex

--betweenness-centrality
Print the betweenness centrality of each vertex

--lexrank-centrality
Print the LexRank centrality of each vertex

example: bin/aan_network_stats.pl -input="author-citation"

Example with sampling: bin/aan_network_stats.pl -input="acit" --
sample
100 --sampletype randomnode -all

7. aan_paper_citations.pl

Outputs the number of citations for every paper.

Usage: bin/aan_paper_citations.pl [-year=to_year] [-incites] [-outcites]
[-nonself] [-help]

-year=to_year
when specified, only citations which are older than the year mentioned
are included. Can be any year greater than 1965, defaults to 2011.

-incites
prints out the number of incoming citations for every paper in the paper
citation network. By default it prints out the number of incoming citations.

-outcites

prints out the number of outgoing citations for every paper in the paper citation network -nonself when specified, self citations are excluded. By default self citations are NOT excluded.

-help
prints out the different options available

Example: bin/aan_paper_citations.pl -year=2011 -incites

8. aan_pageranks.pl

Outputs the PageRank score for every node in the network specified.

Usage: bin/aan_pageranks.pl -input=[acit|pcit] [-help]

--input=acit|pcit
Input network

Example: bin/aan_pageranks.pl -input=acit

---

4. INSTALLATION INSTRUCTIONS

The whole release is packaged together as aanrelease2011.tar.gz. First
untar the release. Let the release be untarred in a directory which we will refer to as $HOME

The release comes with a set of scripts that can be used to process the data.
The scripts can be found under the bin subdirectory. All the scripts are in
Perl and will work perfectly mostly out of the box.

The Perl scripts assume that Perl is installed in /usr/local/bin/

If this is not the case, then change the first line of the Perl scripts
to the directory where Perl is installed.

The aan_network_stats.pl makes use of scripts in clairlib (www.clairlib.org). To get
aan_network_stats.pl working, you need to install clairlib, which can
be found www.clairlib.org. The website also contains installation instructions for installing clairlib. Suppose clairlib is installed in
$CLAIRLIB_HOME, then you need to add the lib directory of clairlib to
PERL5LIB variable as

$HOME>PERL5LIB=$PERL5LIB:$CLAIRLIB_HOME/lib

Once this is done, all the scripts should work perfectly as shown in the examples.

---

5. KNOWN ISSUES

There are some minor issues with the AAN data, specifically paper IDs and
author names. The most updated and correct release is the 2012 release.
Specifically, we have merged some duplicate authors in the 2011 release
for the 2012 release. Also, due to changes in the parent data stored in
aclweb.org, we have mapped paper ids W04-99** to W04-32**. Therefore,
you will find W04-99**'s text in the fulltext directory but they truly
correspond to W04-32**'s fulltext.

---

6. ABOUT THE PROJECT

The ACL Anthology Network was built from the original pdf files available from the ACL Anthology (http://acl.ldc.upenn.edu/) as it stood in July 2011. Using open source OCR technologies, in-house clean-up scripts, and often tedious manual labor, a web interface was
developed that allowed for the annotation of individual references from each paper. A team of student research assistants manually matched references to existing ACL ID's returned using a keyword matching algorithm. Those citations deemed to refer to ACL papers but
which were not automatically matched were marked for post-processing.

Using this paper-id network (paper-citation-network) and the metadata
(acl-metadata), the author citation and author collaboration networks
were then created.

---

7. ACKNOWLEDGEMENTS

The current version is being maintained by Yale's LILY lab. Specifically
we would like to thank the following for their work with this website:

* Pong Trairatvorakul
* Daniel Keller
* Alexander Strzalkowski
* Sydney Young
* Jungo Kasai
* Aaron Pang
* Clark Xie
* Dan Friedman

Previously, a number of students from the University of Michigan's CLAIR
Group helped with the work involved to create the data, network, and webpages.
We would like to thank:

* YoungJoo (Grace) Jeon
* Mark Schaller
* Ben Nash
* John Umbaugh
* Tunay Gur
* Jahna Otterbacher
* Arzucan Ozgur
* Li Yang
* Anthony Fader
* Joshua Gerrish
* Stephen Hufnagel
* Dr. Igor Markov
* Nayeoung Kim
* Paul Hartzog
* Chen Huang
* Pradeep Muthukrishnan
* Vahed Qazvinian
* Ahmed Hassan
* Prem Ganeshkumar
* Amjad Abu Jbara
* Matt Simmons

The previous version of this work has been partially supported by the
National Science Foundation grant "Collaborative Research:

BlogoCenter
- Infrastructure for Collecting, Mining and Accessing Blogs", jointly
awarded to UCLA and UMich as IIS 0534323 to UMich and IIS 0534784 to
UCLA and by the National Science Foundation grant "iOPENER: A Flexible
Framework to Support Rapid Learning in Unfamiliar Research Domains",
jointly awarded to UMd and UMich as IIS 0705832.

For more information, please visit:

http://tangra.cs.yale.edu/newaan