Tutorial: Large Scale Network Analytics with SNAP

http://snap.stanford.edu/proj/snap-icwsm

Rok Sosič, Jure Leskovec
Stanford University

ICWSM-14, Ann Arbor, MI **June, 2014**

SNAP Hands-on Exercise

Rok Sosič, Jure Leskovec
Stanford University

ICWSM-14, Ann Arbor, MI **June, 2014**

Stack Overflow Dataset

- Publicly available by Stack Overflow

  https://archive.org/download/stackexchange/stackoverflow.com-Posts.7z

- 5.2GB compressed, 26GB uncompressed

- 19,881,020 posts from Jul 2008 to May 2014

Hands-on Exercise

- **Task:**
  - Find top Java experts on Stack Overflow
- **Possible approaches for finding experts:**
  - Use Stack Overflow reputation score:
    - Not Java specific
    - No control
  - Count the number of answers:
    - No measure of answer importance or usefulness
  - Create a social network and compute user centrality:
    - Pagerank

Finding Top Java Experts

- **Plan:**
  - Use node centrality measure, Pagerank
  - Need a graph
- **Constructing a graph:**
  - Nodes, each user a node
  - Edges, a question owner points to the owner of the accepted answer

Stack Overflow: Questions

- **Questions XML format in Posts.xml:**
  - Total 7,214,697 questions, Java 632,493

```
<row Id="4“ PostTypeId="1“
 OwnerUserId="8“ AcceptedAnswerId="7“
 Tags="&lt;c#&gt;&lt;winforms&gt;&lt;forms&gt;
   &lt;opacity&gt;" .. />
```

Field	Value
Question Id	4
Post Type	1 (question)
Question Owner	8
Accepted Answer	7
Tags	c#, winforms, forms, opacity

Stack Overflow: Answers

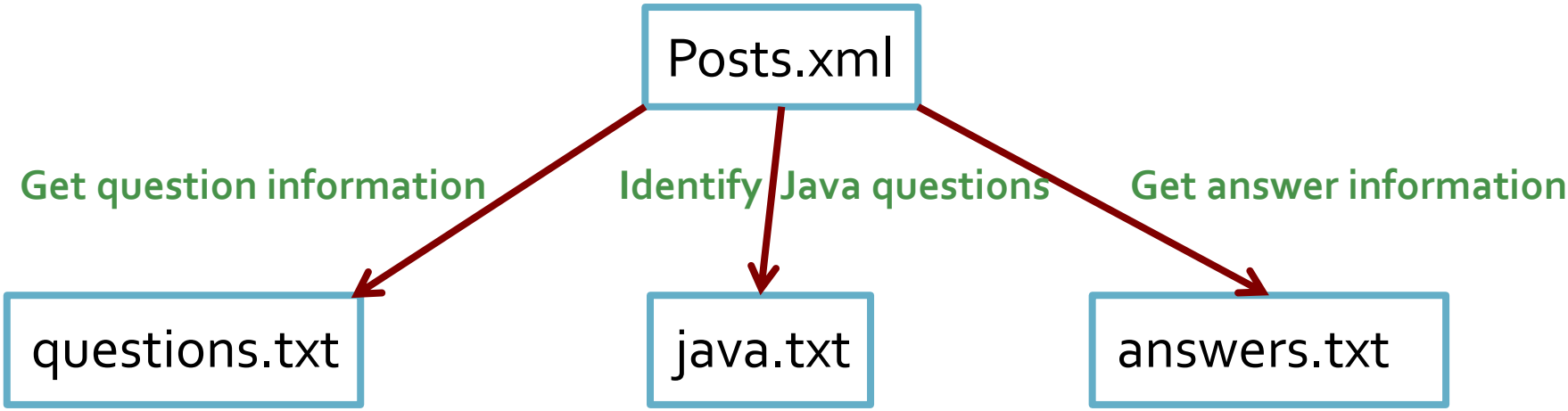
- **Answers XML format in Posts.xml:**
  - total 12,609,623

```
<row Id="12" PostTypeId="2" OwnerUserId="1" ... />
```

Field	Value
Answer Id	12
Post Type	2 (answer)
Answer Owner	1

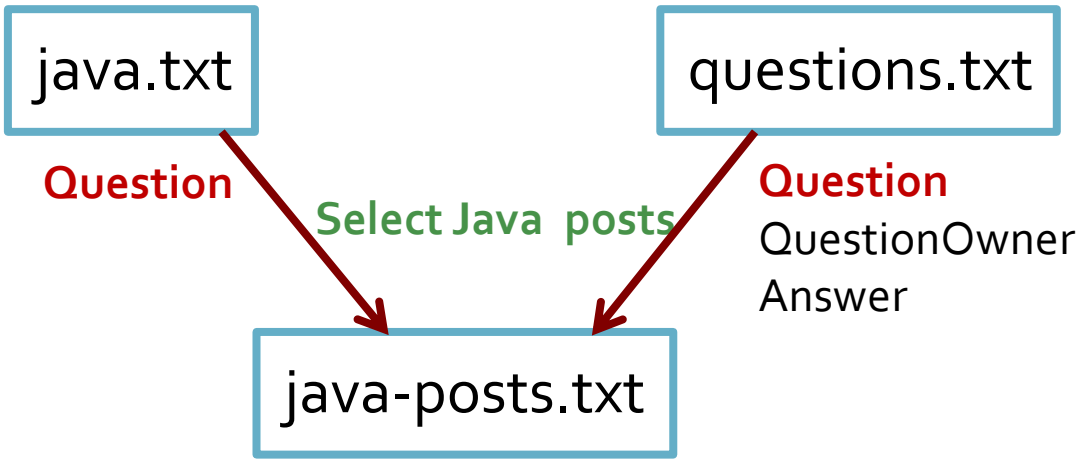
Workflow to Find Java Experts

- **Step 1, process input file, extract relevant fields**
  - Get lists of questions and answers, identify Java posts
  - Convert XML format to TSV (tab separated values)



Workflow to Find Java Experts

- **Step 2**, Select only questions about Java

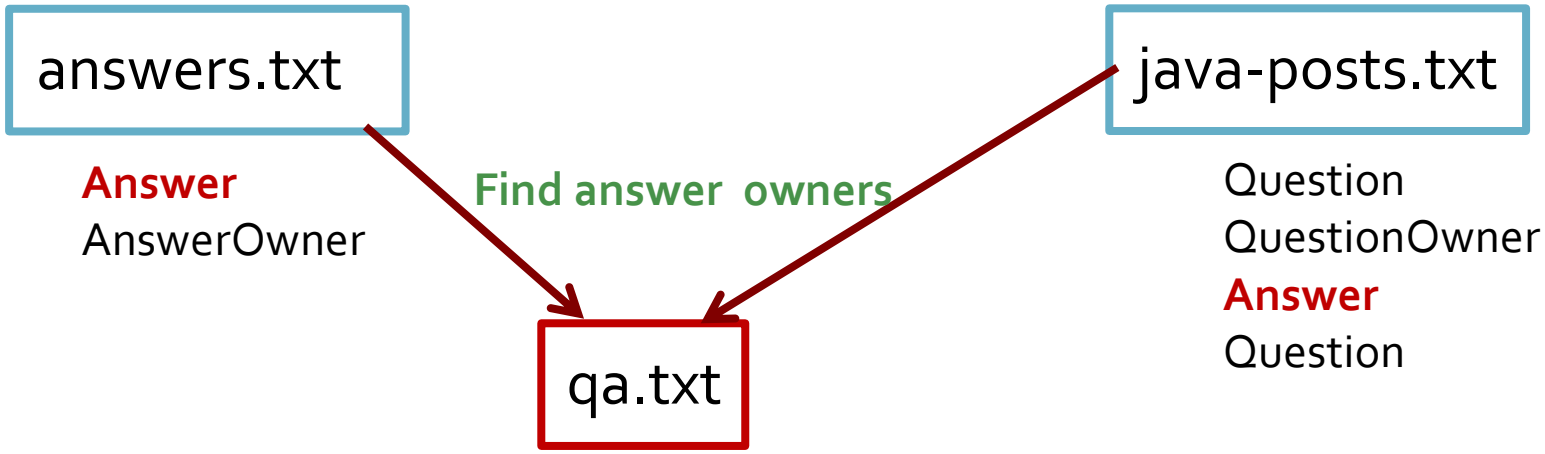


doJoin.py java.txt questions.txt 1 1

Question
QuestionOwner
Answer
Question

Workflow to Find Java Experts

- **Step 3, Find owners of accepted answers**

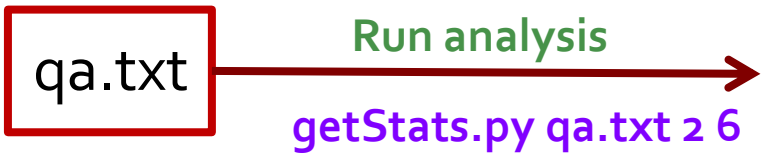


doJoin.py answers.txt java-posts.txt 1 3

Question
QuestionOwner
Answer
Question
Answer
AnswerOwner

Workflow to Find Java Experts

- **Step 4, analyze the graph**
  - Find top Java experts



Question
QuestionOwner
Answer
Question
Answer
AnswerOwner

- Program calculations
  - # of nodes, edges
  - Distribution of weakly connected components
  - In and out-degree distributions
  - Top 10 experts by Pagerank
  - Top 10 experts by Hits
  - Top 10 learners by HIts

```
top 10 experts by PageRank
id   992484, pagerank 0.013981
id   135152, pagerank 0.010006
id    22656, pagerank 0.007104
id   139985, pagerank 0.005521
id   157882, pagerank 0.004597
…
```

Find Java Experts: Hands-on Exercise

- **Download and install Snap.py**
  http://snap.stanford.edu/snappy/index.html
- **Download programs and data for the exercise: icwsm14-T4-code.zip** and **icwsm14-T4-data.zip**, for finding experts on Stack Overflow
  http://snap.stanford.edu/proj/snap-icwsm
- **Unpack** zip files icwsm14-T4-code.zip and icwsm14-T4-data.zip
- **Find experts** by executing the programs from command line
  - **stackoverflow.sh** on Mac OS X and Linux
  - **stack.bat** on Windows
- **Explore getStats.py**
  - Extend it with different graph analysis methods
- **Extra exercise**
  - Find Javascript experts
- Stack Overflow original data
  https://archive.org/download/stackexchange/stackoverflow.com-Posts.7z

Contact information: Rok Sosic, rok@cs.stanford.edu

Further SNAP Resources

Rok Sosič, Jure Leskovec
Stanford University

ICWSM-14, Ann Arbor, MI **June, 2014**

Snap.py Resources

- **Prebuilt packages** available for Mac OS X, Windows, Linux
  http://snap.stanford.edu/snappy/index.html
- **Snap.py documentation**:
  http://snap.stanford.edu/snappy/doc/index.html
  - Quick Introduction, Tutorial, Reference Manual
- **SNAP user mailing list**
  http://groups.google.com/group/snap-discuss
- **Developer resources**
  - Software available as open source under BSD license
  - GitHub repository
    https://github.com/snap-stanford/snap-python

SNAP C++ Resources

- **Source code** available for Mac OS X, Windows, Linux
  http://snap.stanford.edu/snap/download.html
- **SNAP documentation**
  http://snap.stanford.edu/snap/doc.html
  - Quick Introduction, User Reference Manual
  - Source code, see **tutorials**
- **SNAP user mailing list**
  http://groups.google.com/group/snap-discuss
- **Developer resources**
  - Software available as open source under BSD license
  - GitHub repository
    https://github.com/snap-stanford/snap
  - SNAP C++ Programming Guide

SNAP Network Datasets

Collection of over 70 social network datasets: http://snap.stanford.edu/data

Mailing list: http://groups.google.com/group/snap-datasets

- **Social networks:** online social networks, edges represent interactions between people
- **Twitter and Memetracker :** Memetracker phrases, links and 467 million Tweets
- **Citation networks:** nodes represent papers, edges represent citations
- **Collaboration networks:** nodes represent scientists, edges represent collaborations (co-authoring a paper)
- **Amazon networks :** nodes represent products and edges link commonly co-purchased products