Tutorial: Large Scale Network Analytics with SNAP

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

Rok Sosič, Jure Leskovec
Stanford University

WWW-15, Florence, Italy **May, 2015**

SNAP Hands-on Exercise

Rok Sosič, Jure Leskovec
Stanford University

WWW-15, Florence, Italy **May, 2015**

Stack Overflow Dataset

- Publicly available by Stack Overflow

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

- 6.6GB compressed, 33GB uncompressed

- From Jul 2008 to Apr 2015
  - 8,978,719 questions, 15,074,572 answers

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, HITS

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

Finding Top Java Experts

- **Method Overview:**
  - Step 1: Extract relevant fields from input
  - Step 2: Select questions about Java
  - Step 3: Build the graph
    - Find owners of accepted answers
  - Step 4: Analyze the graph

Stack Overflow: Questions

- **Questions XML format in Posts.xml:**
  - Total 8,978,719 questions, Java 810,071

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

Field	Value
Id	4
PostTypeId	1
OwnerUserId	8
Accepted AnswerId	7
Tags	c#, winforms, forms, opacity

Stack Overflow: Answers

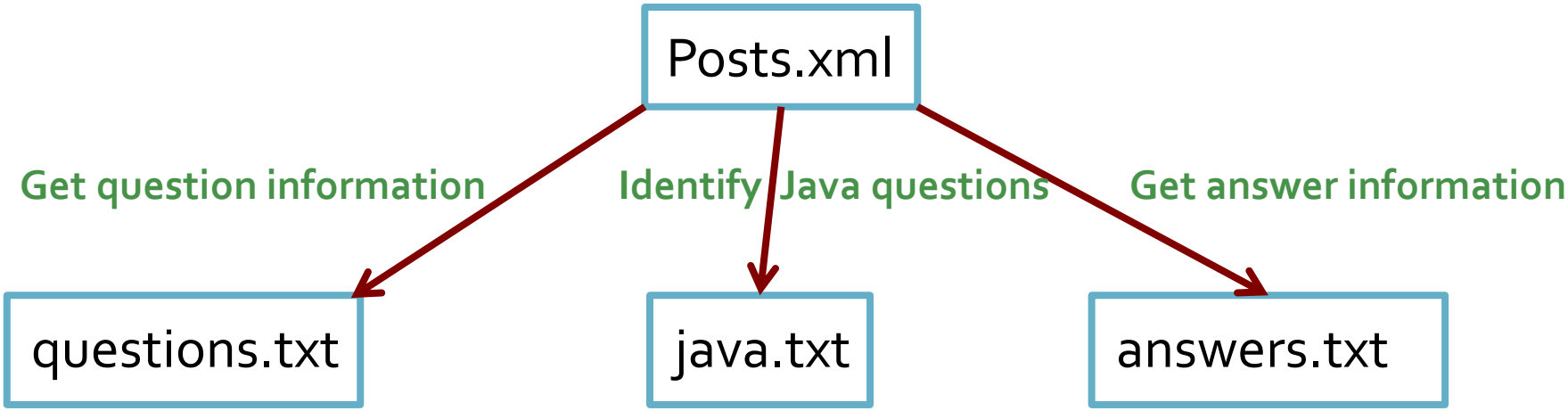
- **Answers XML format in Posts.xml:**
  - Total 15,074,572

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

Field	Value
Id	12
Post TypeId	2
OwnerUserId	1

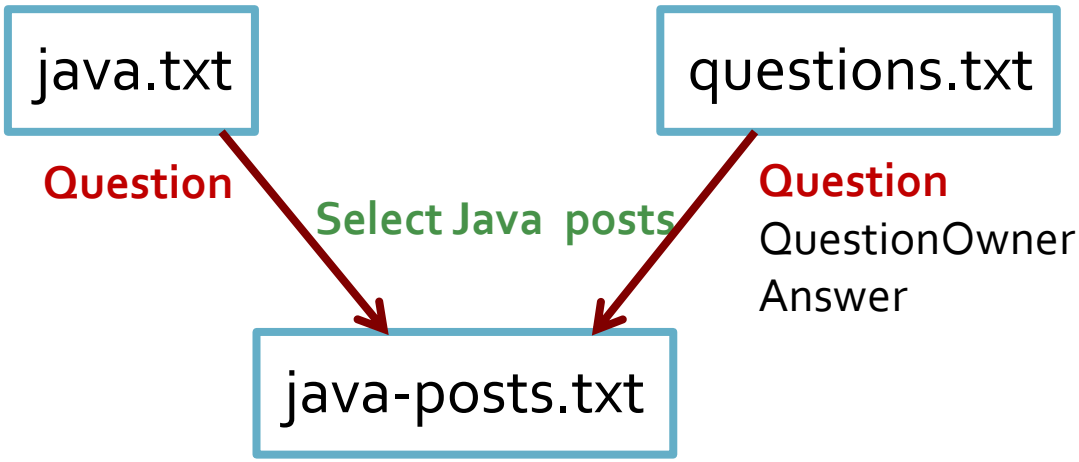
Workflow to Find Java Experts

- **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 Java related questions

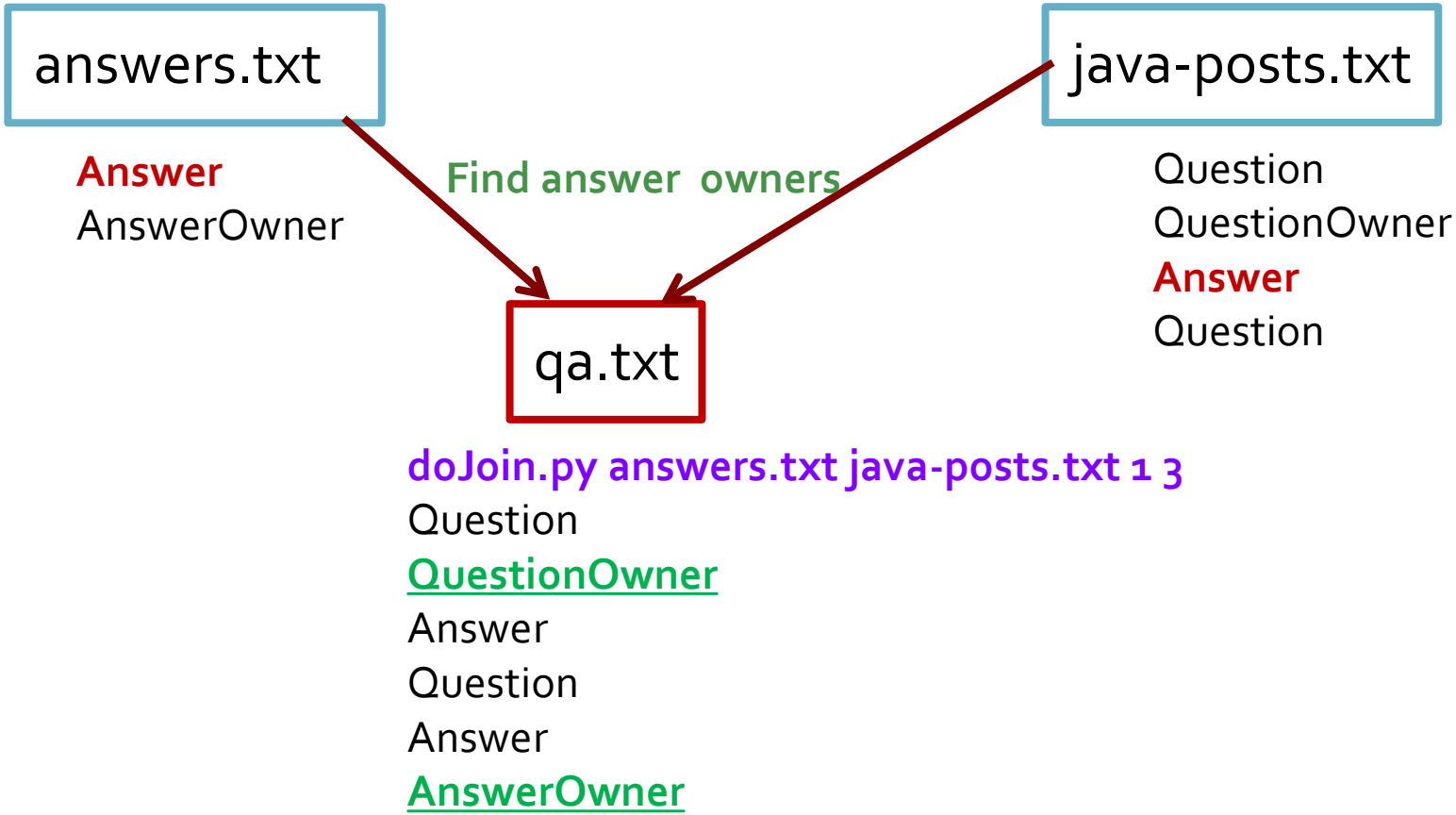


doJoin.py java.txt questions.txt 1 1

Question
QuestionOwner
Answer
Question

Workflow to Find Java Experts

- **Step 3, Build the graph by finding owners of accepted answers**



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   22656, pagerank 0.007056
id  139985, pagerank 0.005290
id  571407, pagerank 0.004348
id  992484, pagerank 0.003722
id  157882, pagerank 0.003628
...
```

Java Experts Graph

```
G = snap.LoadEdgeList(snap.PNGraph, "qa.txt", 1, 5)
snap.PrintInfo(G, "QA Stats", "qa-info.txt", False)
```

Output:

```
QA Stats: Directed
  Nodes:                     188406
  Edges:                     415174
  Zero Deg Nodes:            0
  Zero InDeg Nodes:          108618
  Zero OutDeg Nodes:         38319
  NonZero In-Out Deg Nodes:  41469
  Unique directed edges:     415174
  Unique undirected edges:   415027
  Self Edges:                26924
  BiDir Edges:               27218
  Closed triangles:          46992
  Open triangles:            69426319
  Frac. of closed triads:    0.000676
  Connected component size:  0.886745
  Strong conn. comp. size:   0.025758
  Approx. full diameter:     13
  90% effective diameter:    5.751723
```

Java Experts on Stack Overflow

- **Comparing methods on top 10 results:**

http://stackoverflow.com/users/<id>

In-degree	RageRank	HITS
22656	22656	22656
571407	139985	571407
992484	571407	57695
157882	992484	139985
57695	157882	157882
139985	57695	203907
522444	218978	992484
131872	70604	522444
438154	230513	131872
207421	438154	438154

Java Learners on Stack Overflow

- **Comparing methods on top 10 results:**

  http://stackoverflow.com/users/<id>

Out-degree	HITS
1194415	892029
892029	1194415
785349	359862
470184	648138
454049	470184
853836	802050
359862	384706
44330	225899
663148	454049
1379286	130758

Finding Top Java Experts

- **Solution:**
  - Step 1: Extract relevant fields from input

```
python getQuestions.py Posts.xml > questions.txt
python getAnswers.py Posts.xml > answers.txt
python getTag.py Posts.xml java > java.txt
```

  - Step 2: Select questions about Java

```
python doJoin.py java.txt questions.txt 1 1 > \
                java-posts.txt
```

  - Step 3: Build the graph

```
python doJoin.py answers.txt java-posts.txt 1 3 > \
                qa.txt
```

  - Step 4: Analyze the graph

```
python getStats.py qa.txt 2 6 > stats.txt
```

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: www15-code.zip** and **www15-data.zip**, for finding experts on Stack Overflow
  http://snap.stanford.edu/proj/snap-icwsm
- **Unpack** zip files www15-code.zip and www15-data.zip
- **Find experts** by executing the following programs from command line
  - **stackoverflow.sh** on Mac OS X and Linux
  - **stack.bat** on Windows
  - **stats.txt** contains the output
- **Explore getStats.py**
  - Extend it with different graph analysis methods
- **Extra exercise**
  - Find Javascript experts, change in experts over time
- Stack Overflow original data - 6.6GB!
  https://archive.org/download/stackexchange/stackoverflow.com-Posts.7z

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

Further SNAP Resources

Rok Sosič, Jure Leskovec
Stanford University

WWW-15, Florence, Italy **May, 2015**

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 web and 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