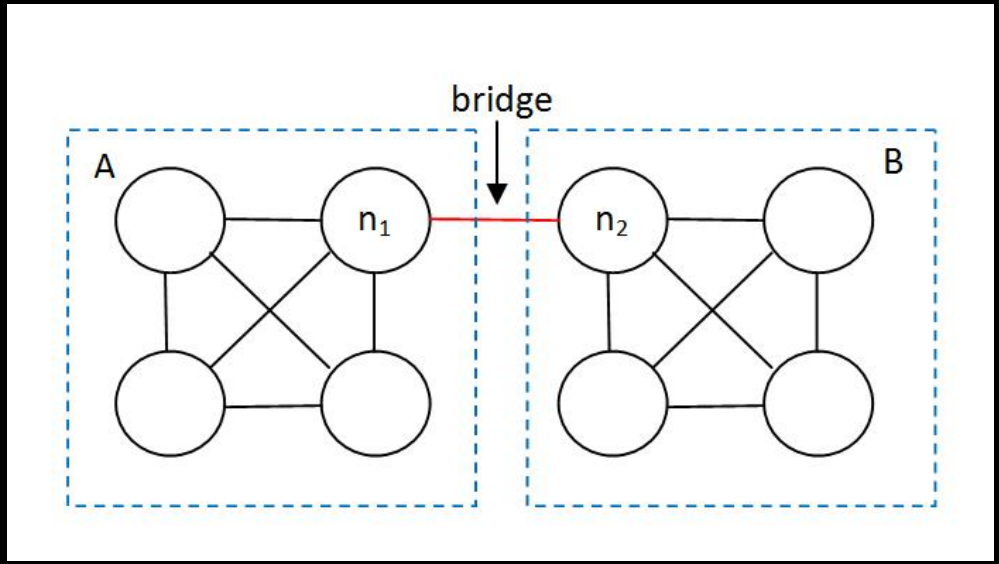

COIN7 CoolPeople
Final presentation
David Huser
Igor Bosnjak

Goals

Extract **names of persons** in **text sources** and generate a **network file** compatible with e.g. Gephi network visualization.

Two persons are linked if they appear in the same document.

If a person is in more than a document it becomes a bridge.



Methods used

1. Apache openNLP

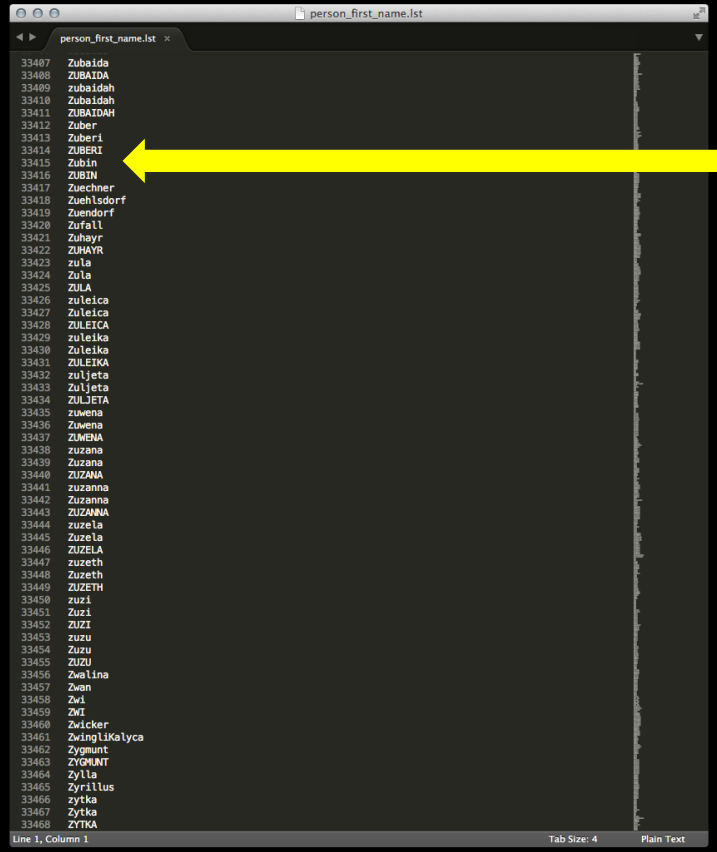
Natural Language Processing

Machine learning based toolkit for the processing of natural language text.

Supports the most common NLP tasks:

tokenization, sentence segmentation, part-of-speech tagging, **named entity extraction**, chunking, parsing, coreference resolution

2. Whitelist

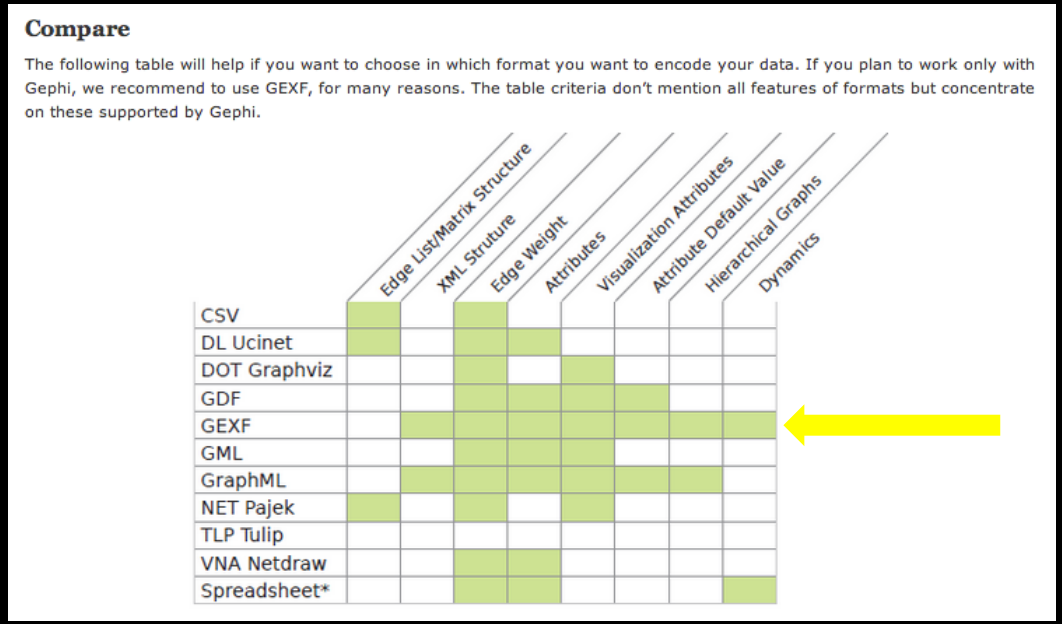


If a word in the input document contains one of the names:

Add the **following word** as a new person, <u>if the prename and the lastname are Uppercase.</u>

3. GEXF4J

A java library for the GEXF file format



Compare

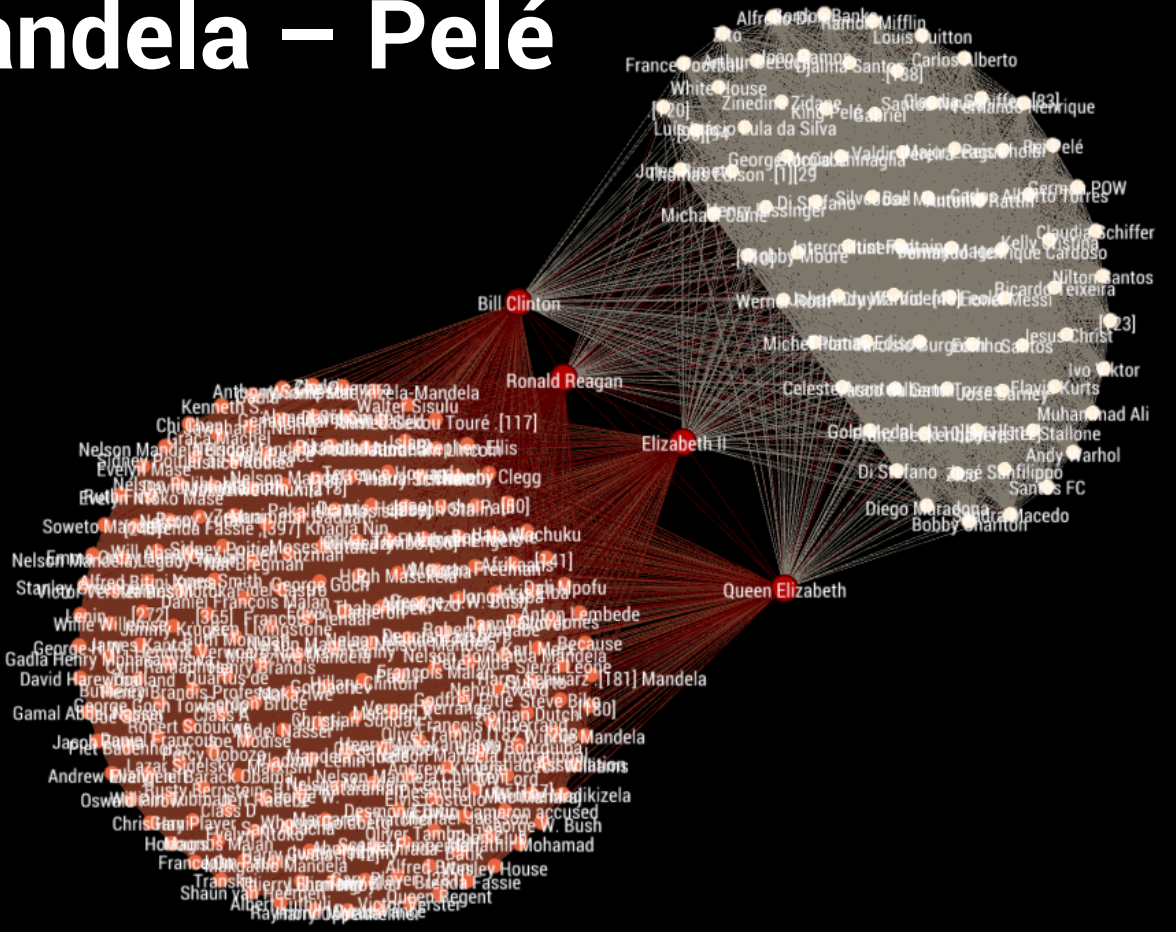
The following table will help if you want to choose in which format you want to encode your data. If you plan to work only with Gephi, we recommend to use GEXF, for many reasons. The table criteria don't mention all features of formats but concentrate on these supported by Gephi.

	Edge List/Matrix Structure	XML Structure	Edge Weight	Attributes	Visualization Attributes	Attribute Default Value	Hierarchical Graphs	Dynamics
CSV	■		■					
DL Ucinet	■		■	■				
DOT Graphviz			■		■			
GDF			■	■	■	■		
GEXF		■	■	■	■	■	■	■
GML			■	■	■			
GraphML		■	■	■	■	■	■	
NET Pajek	■		■		■			
TLP Tulip								
VNA Netdraw			■	■				
Spreadsheet*			■	■				■

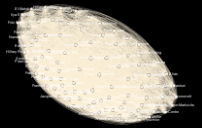
Main results

Nelson Mandela – Pelé

Nelson Mandela – Pelé



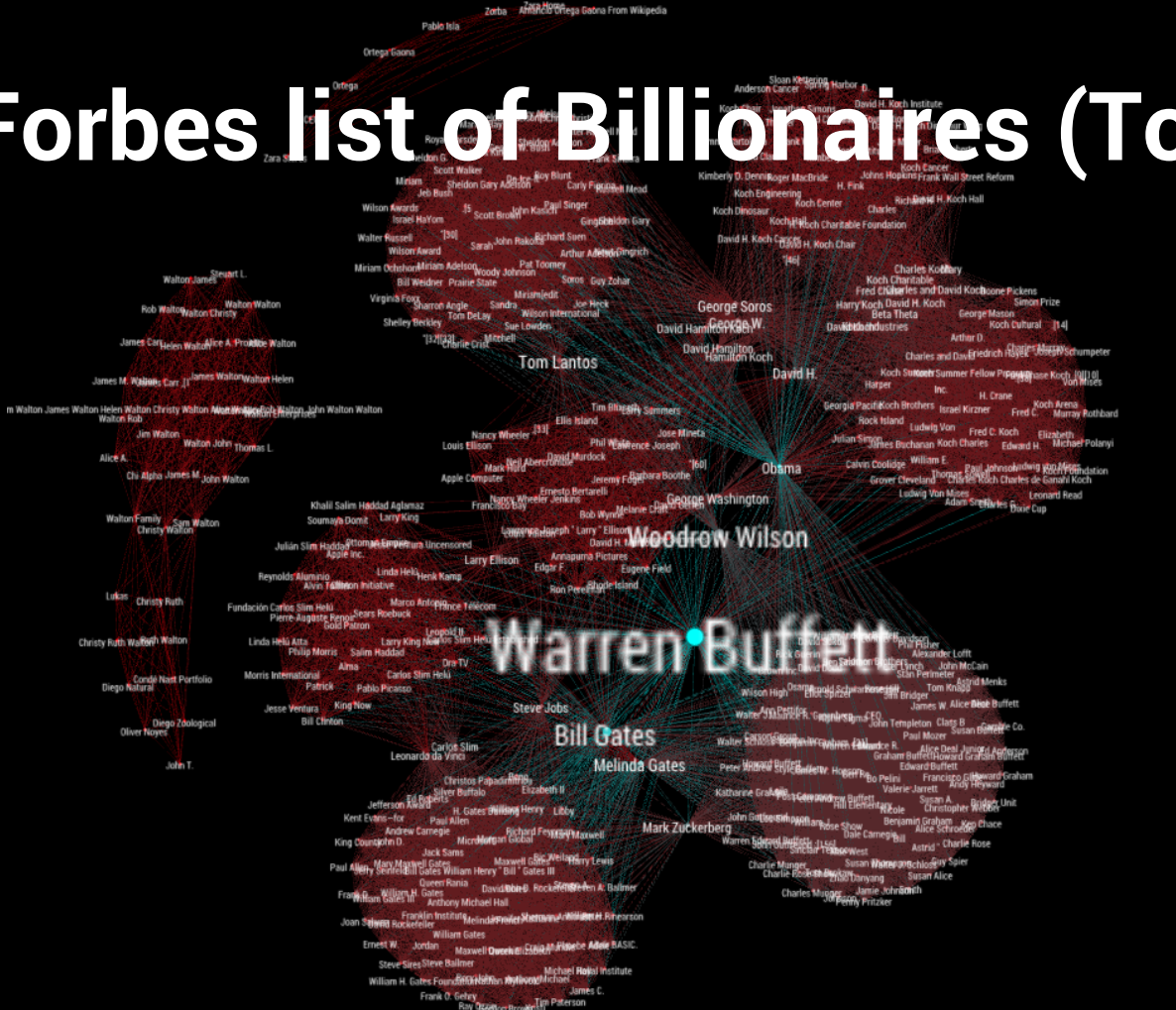
Anne Frank – Pablo Picasso



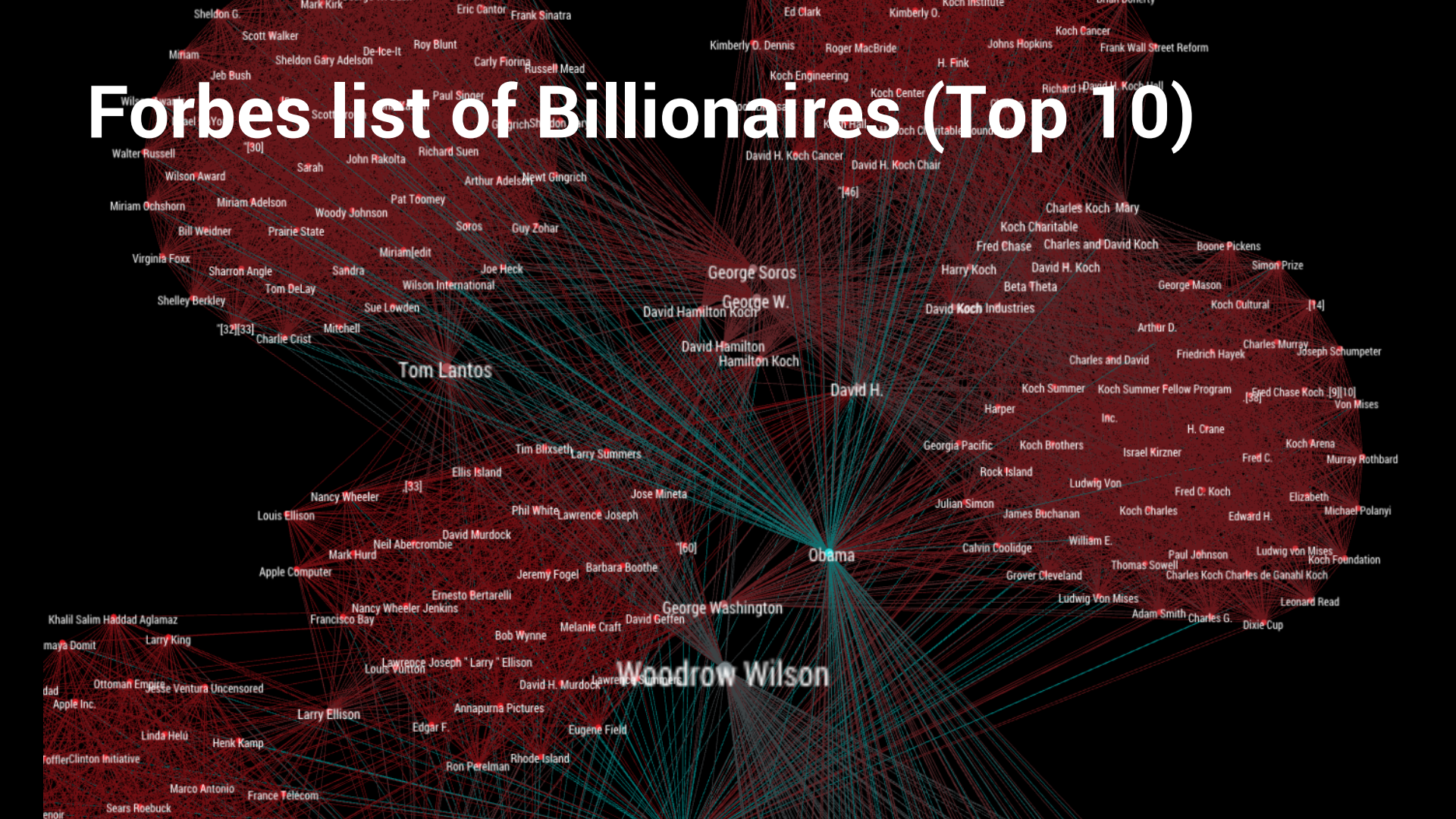
Forbes list of Billionaires (Top 10)

No.	Name	Net worth (USD)	Age	Citizenship	Source(s) of wealth
1	Bill Gates	$78.0 billion	58	United States	Microsoft
2	Carlos Slim & family	$71.2 billion	74	Mexico	Telmex, América Móvil, Grupo Carso
3	Warren Buffett	$65.5 billion	83	United States	Berkshire Hathaway
4	Amancio Ortega	$62.3 billion	78	Spain	Inditex Group
5	Larry Ellison	$51.4 billion	69	United States	Oracle Corporation
6	Charles Koch	$41.4 billion	78	United States	Koch Industries
6	David Koch	$41.4 billion	74	United States	Koch Industries
8	Christy Walton & family	$37.8 billion	58-59	United States	Wal-Mart
9	Sheldon Adelson	$37.7 billion	80	United States	Las Vegas Sands
10	Bernard Arnault & family	$36.3 billion	65	France	LVMH Moët Hennessy • Louis Vuitton

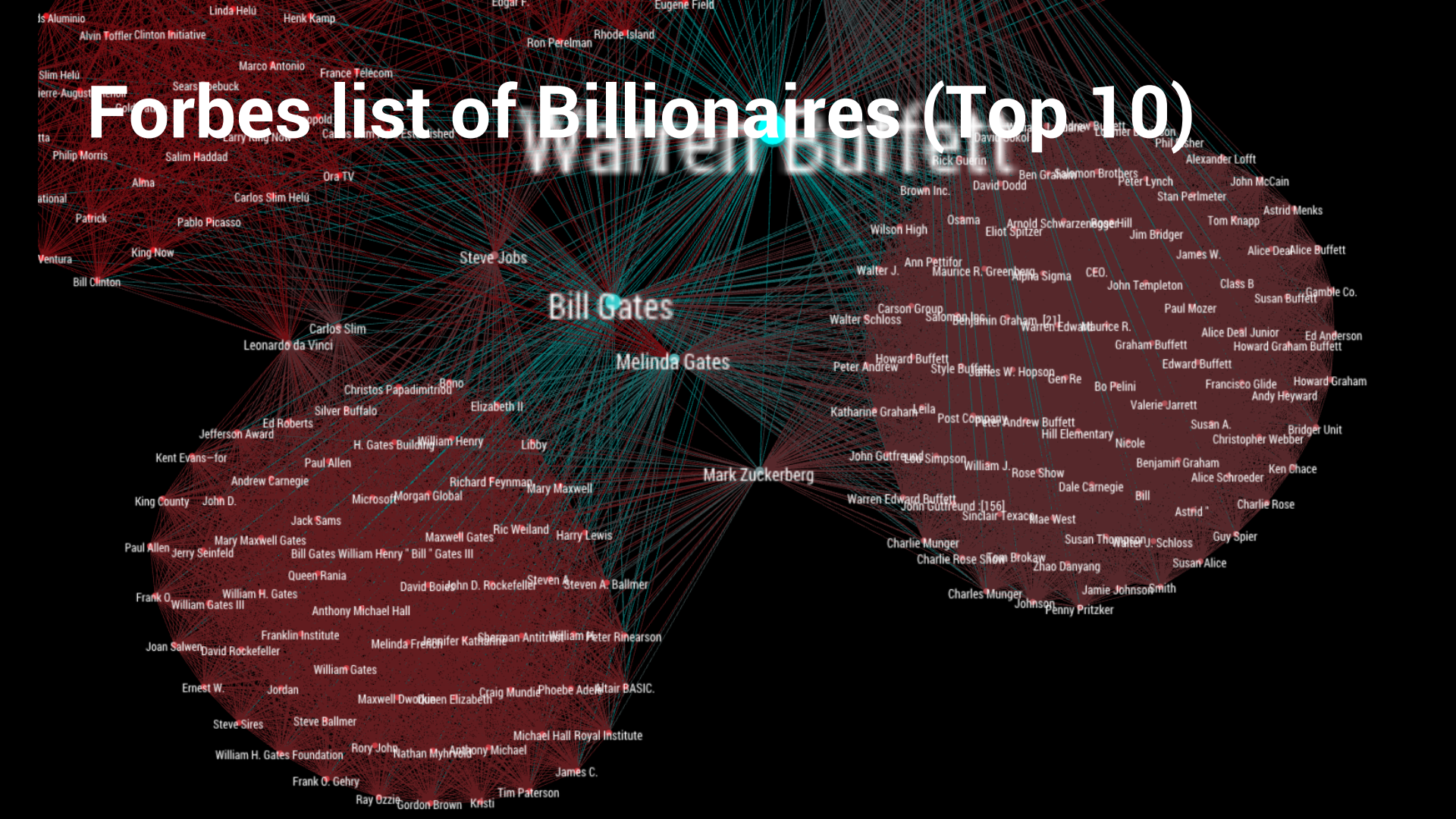
Forbes list of Billionaires (Top 10)



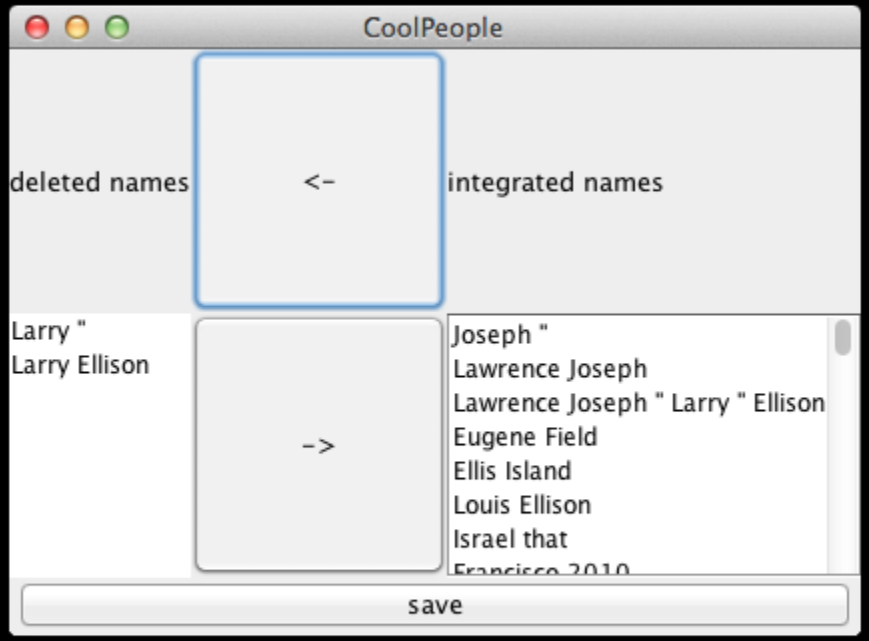
Forbes list of Billionaires (Top 10)



Forbes list of Billionaires (Top 10)



Filter GUI


CoolPeople
deleted names
<-
integrated names
Larry "
Larry Ellison
->
Joseph "
Lawrence Joseph
Lawrence Joseph " Larry " Ellison
Eugene Field
Ellis Island
Louis Ellison
Israel that
save

Open issues

Performance

- Multithreading
- Forbes Top 10: 30 seconds @ Intel i5
- GUI improvements: Filechooser
- Save button

Future work

Extendable architecture

A document object can be extended… all we need is a String to extract names.

- Web scraping
- MySQL Adapter
- ...

Extendable export file

The Gephi export format can be used for directed graphs, timelines, ...

Maven

Apache Maven build tool for dependencies, testing and resources

Discussion
Q&A

Links

Apache openNLP – https://opennlp.apache.org/

openNLP models – http://opennlp.sourceforge.net/models-1.5/

Gephi – https://gephi.org/

GEXF – http://gexf.net/format/index.html

Apache Maven – http://maven.apache.org/