KGTK:
Tools for Creating and Exploiting Large Knowledge Graphs

Filip Ilievski, Daniel Garijo*, Hans Chalupski and Pedro Szekely
USC Information Sciences Institute
*Universidad Politécnica de Madrid

Filip Ilievski
USC/ISI

Daniel Garijo
UP Madrid

Hans Chalupsky
USC/ISI

Pedro Szekely
USC/ISI

Presenters

Key Contributors, Collaborators and Students

Jay Pujara

Research Assistant Professor

Research Lead

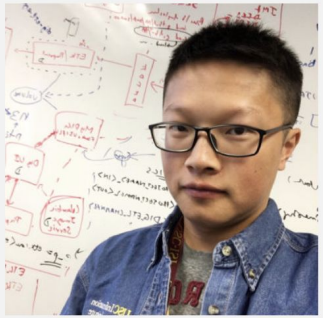
Director of the Center on Knowledge Graphs

Craig Rogers

Systems Programmer

Amandeep Singh

Research Programmer

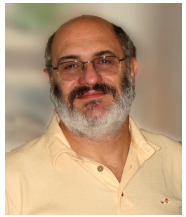


Yixiang Yao

Research Programmer

Gleb Satyukov

Research Programmer

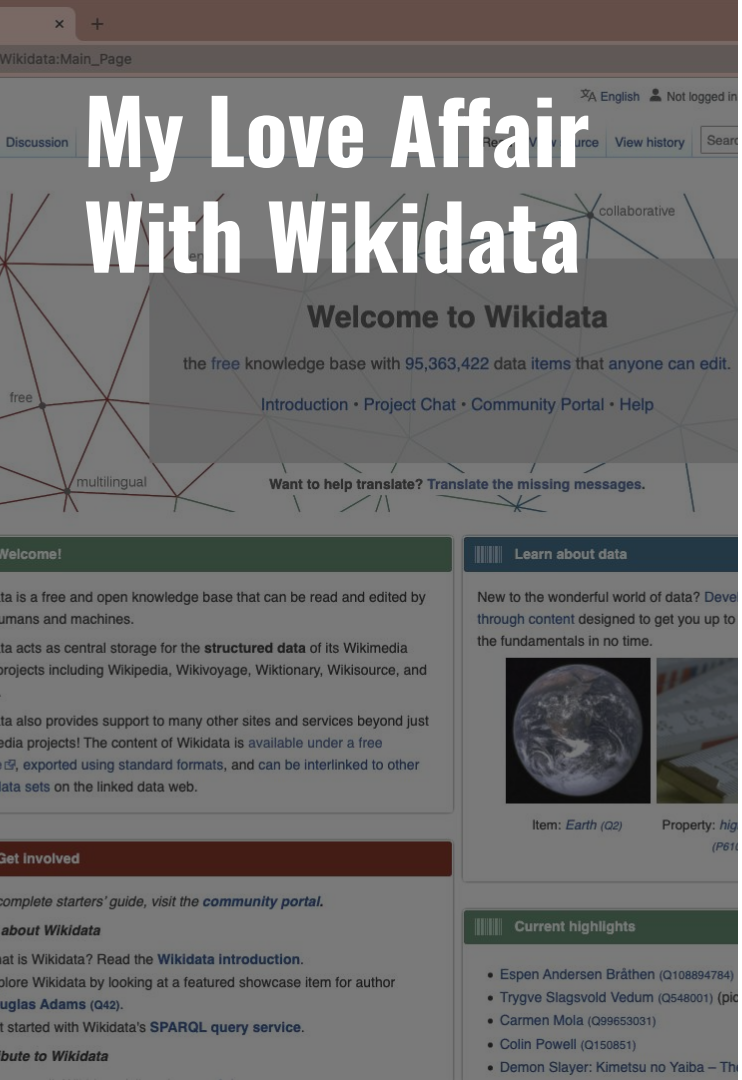


Daniel Schwabe

Ke-Thia Yao, Rongpen Li, Jun Liu, and our awesome student workers: Hardi Rathod, Shreya Anil Naik, Kartik Shenoy, Jiang Wang and Bohui Zhang

Introduction

My Love Affair With Wikidata



Met around 2017 and discovered

- high quality data
- lots of it
- actively maintained
- qualifiers and references

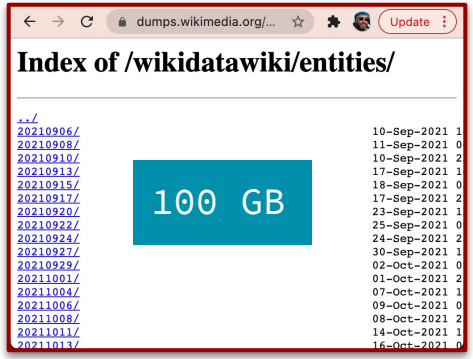
Wanted to use it in all my projects

- extract subsets
- extend them with custom data
- use it for analytics and ML

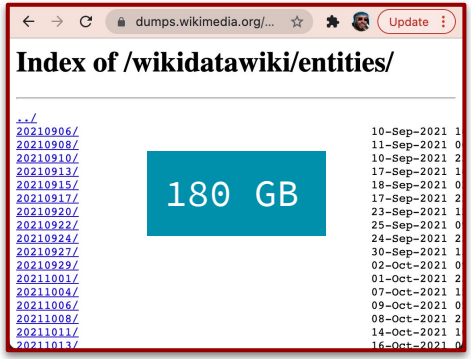
Using Wikidata

SPARQL

Wikidata JSON



RDF ntriples

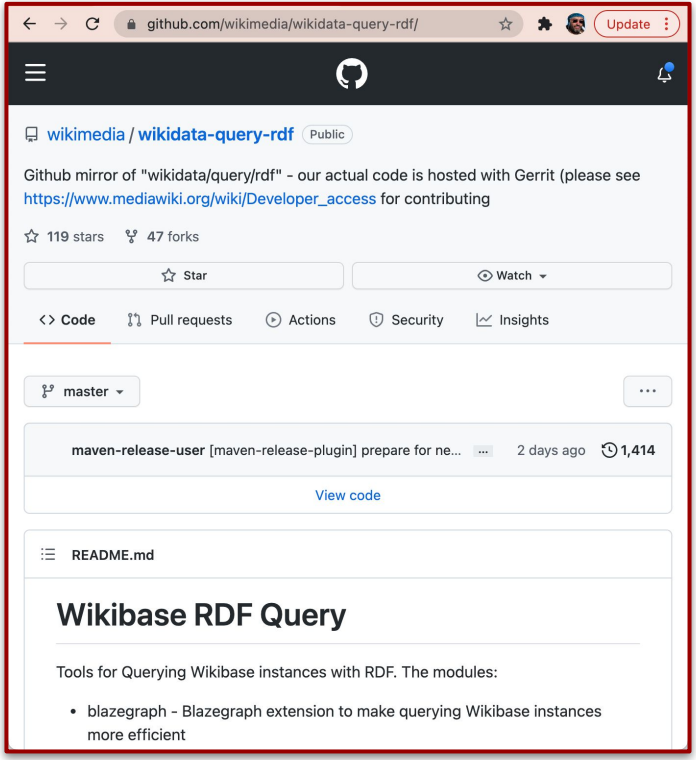


- I can use it online
- It can fit on my laptop

Reality Check

- My SPARQL queries time out
- No tools to work with the JSON dump
- 180 GB too large for RDF tools even on a server

I'll Install My Own SPARQL Endpoint



- > 10 days to load
- If I mess it up I get to do it all over again

Why KGTK?

Born out of frustration:

- Wanted to use Wikidata, but the tools were not up to it

Tasks I wanted to do:

- extract subsets
- do analytics and machine learning
- extend with new data
- work with embeddings

Deploy to customers:

- many industrial/government customers don't want RDF/SPARQL
- students cannot experiment with it (delete parts, add their own, ...

If I built my own KG tools, how would they work?

KGTK Tenets

Scale to billions of statements

Speed competitive with SQL databases

No library needed to read/write KGs

Support Wikidata data model: statements, qualifiers, references

Easy to export/import to state of the art tools (network analytics, embeddings, ElasticSearch, …)

Pandas-friendly because people love Pandas

Pipeline-friendly to create workflows with multiple components

KGTK Non-Tenets: Things I Didn't Care About

URIs: I want to work with one large KG

RDF: I didn't see a need to reuse any RDF tools, happy to import/export

OWL: reasoners don't scale to Wikidata scale

SPARQL: most of my customers don't like it, those who do can use it via RDF export

KGTK Design

One KG = one or more TSV files:

- Columns: <node1, label, node2, edge-id>

Toolkit commands:

- command_i(TSV_1, TSV_2, ...) → TSV_1, TSV_2, …
- input and output are sets of TSV files

Schema free:

- nodes can be anything (identifiers, strings, numbers, dates, …)
- edge labels can also be anything
- don't need to declare anything

Structured literals:

- commonly used literal types are represented as one symbol
- syntax designed for easy, efficient parsing

Why I like KGTK

I can work with Wikidata on my laptop

If I can think it, I can build it

It is fun to build KG pipelines in Jupyter

It is easy to select subsets and add new data

It is easy to combine queries, embeddings & network analytics

It runs faster on my laptop than SPARQL on a server

If I can think it, I can build it: The tutorial KG

The Tutorial KG

Based on Wikidata

Small enough so it runs on Google colab in real time

Large enough to do queries, network analytics & embeddings

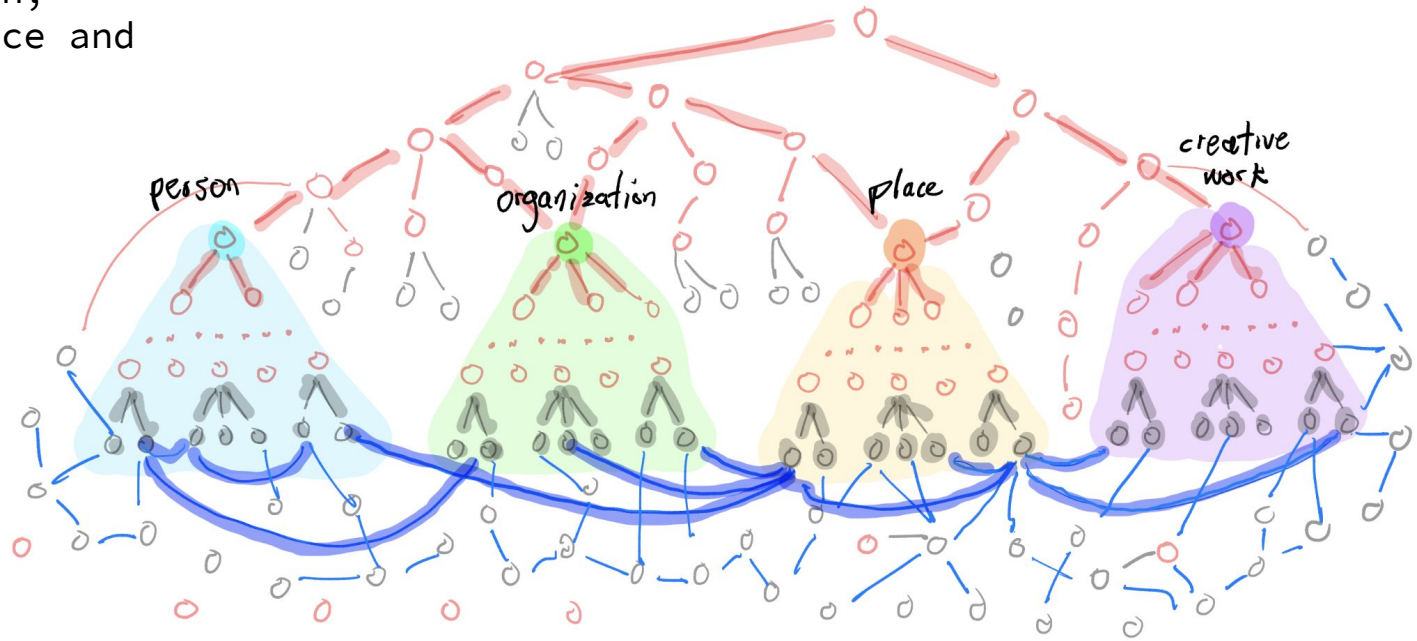
Interesting so it is fun to work with it

Natural so we can find external sources to augment it

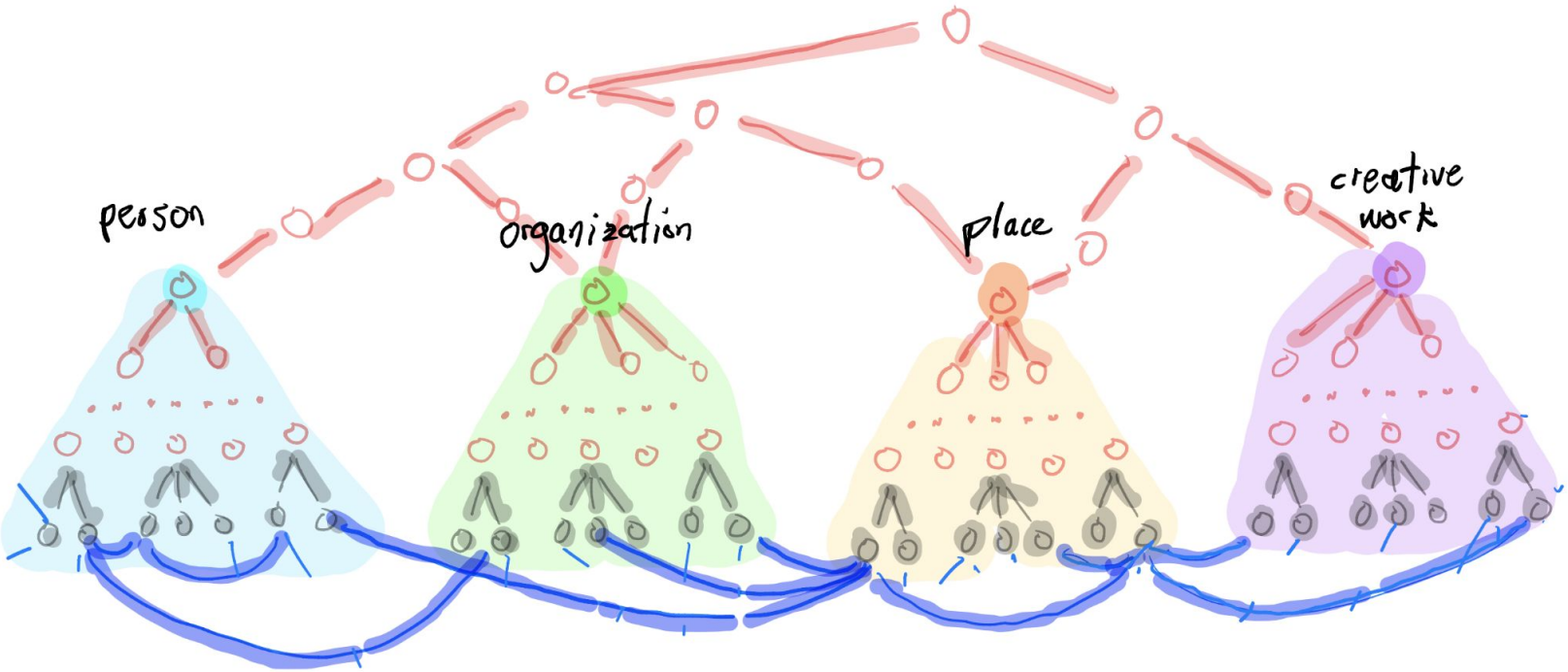
Starting with full Wikidata

Extract subgraph connecting items below person, organization, place and creative work

include "subclass of" edges to root



Wikidata person/organization/place/creative work subgraph

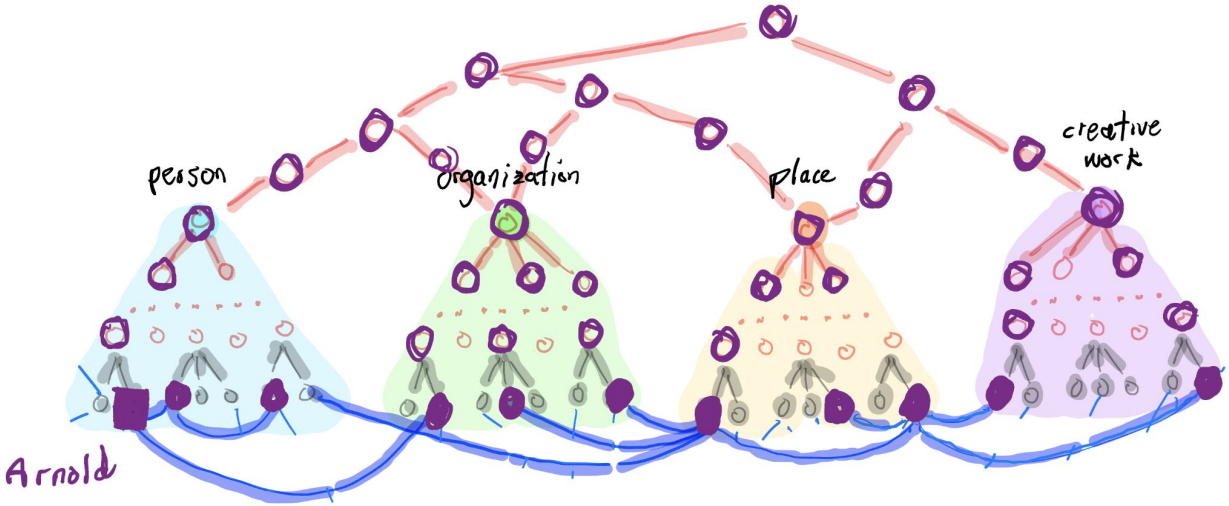


4 hours on my laptop

The Arnold Schwarzenegger graph

- Start from Arnold
- Go forward 3 hops
- Traverse subclass to top
- Get qualifiers of all edges
- Get literal values of all items

what	count
edges (all)	2,614,950
edges (qualifiers)	443,899
items	58,522
properties	3,831
classes	14,490



https://github.com/usc-isi-i2/kgtk/blob/master/tutorial/build-kg/build-tutorial-graph.ipynb

KGTK Browser

kgtk.isi.edu/iswc/browser/Q2685

KGTK Browser: ISWC Tutorial

Search...

Arnold Schwarzenegger

(Q2685)

Arnie | Arnold Alois Schwarzenegger | Arnold Strong | Governator | Governor Schwarzenegger | Schwarzenegger | The Styrian Oak | The Terminator

Austrian-American actor, businessman, bodybuilder and politician

Properties

instance of	human	
sex or gender	male	
birth name	Arnold Alois Schwarzenegger [German]	
name in native language	Arnold Alois Schwarzenegger [German]	
nickname	Arnie [mul]	
	Governator [English]	
	start time	2003
	The Austrian Oak [English]	
	The Styrian Oak [English]	
	literal translation	Roble Austriaco [Spanish]
pseudonym	Arnold Strong i ``governator´´	
date of birth	July 30, 1947	
work period (start)	1969	
child	Christina Schwarzenegger	
	Christopher Schwarzenegger	
	Joseph Baena	
	Katherine Schwarzenegger	
	Patrick Schwarzenegger	
sibling	Meinhard Schwarzenegger	
spouse	Maria Shriver	
	start time	April 26, 1986
	end time	July 01, 2011
relative	Patrick M. Knapp Schwarzenegger	

Gallery

Identifiers

abART person ID	79216
Acharts.co artist ID	arnold_schwarzenegger
Alexander Turnbull Library ID	118183
Allcinema person ID	11017
AllMovie person ID	p110501
AlloCiné person ID	1067
Amazon author ID	B000AP7VZW
Amazon Music artist ID	B0013691AA
American Film Institute person ID	223815
Babelio author ID	237798
Ballotpedia ID	Arnold_Schwarzenegger
Behind The Voice Actors person ID	Arnold-Schwarzenegger
BFI Filmography person ID	187025
BFI Films, TV and people ID	4ce2b9fbd2853
Biblioteca Nacional de España ID	XX974740
Bibliothèque nationale de France ID	13941991t
BIBSYS ID	90851780
Box Office Mojo person ID	arnoldschwarzenegger
Brockhaus Enzyklopädie online ID	schwarzenegger-arnold
C-SPAN person ID	arnoldschwarzenegger
CANTIC ID	a10891948
CCAB ID	000268391
Cinema.de ID	1567829
CineMagia person ID	459
cinematografo.it name or company ID	110845
Cineuropa person ID	261714
CiNii author ID (books)	DA1033823X

https://kgtk.isi.edu/iswc/browser/Q2685

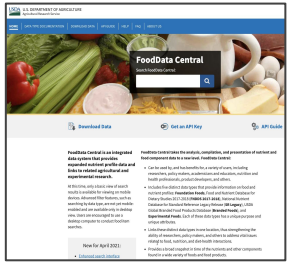
KGTK runs faster on my laptop than SPARQL on a server

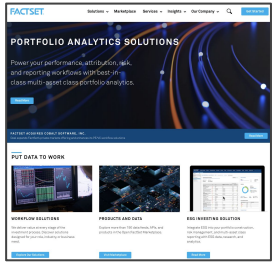
Query	Kypher 16GB laptop	Kypher 32GB laptop	SPARQL 256GB local server	SPARQL public
First names	24.37	8.28	31.05	time out
Class instances	104.97	88.97	>24 hours	time out
Film instances	0.03	0.04	1.91	time out
Author network	61.55	66.39	>24 hours	time out
Cancer network	3.18	2.62	40.19	time out
ULAN identifiers	0.56	0.20	1.08	*
DBpedia spouses	3.92	3.43	n/a	n/a

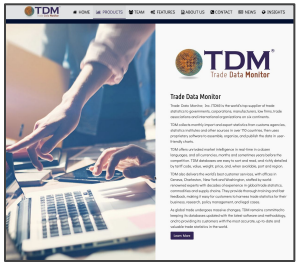
memory is RAM, all times in minutes except noted otherwise, (*) error, query too large

KGs we built using KGTK

KG for trade and supply chain analysis

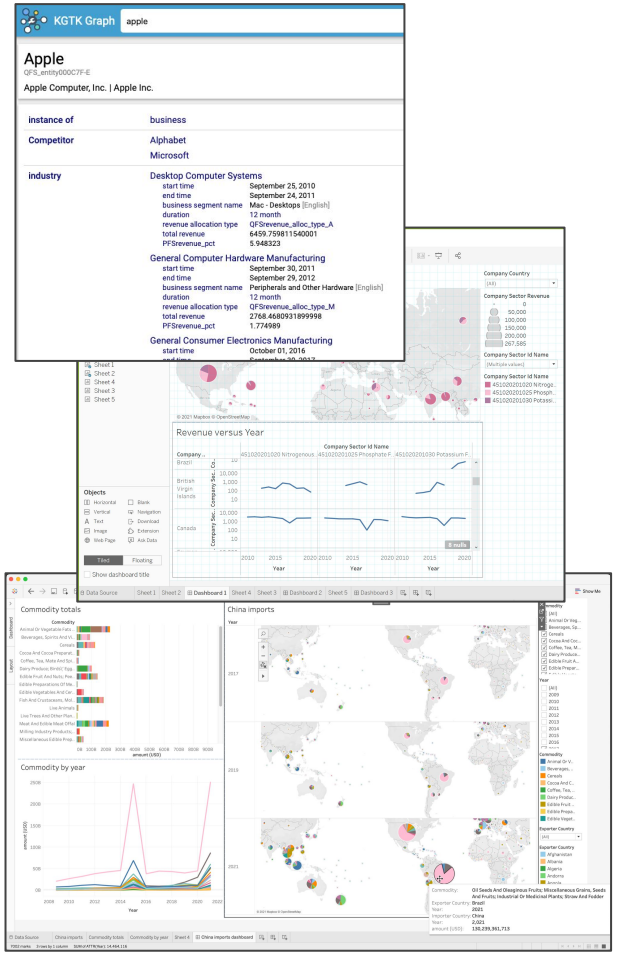






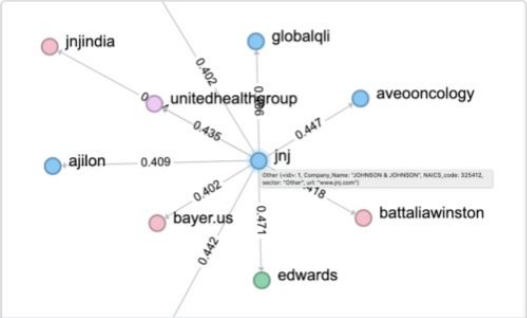
WIKIDATA

KGTK



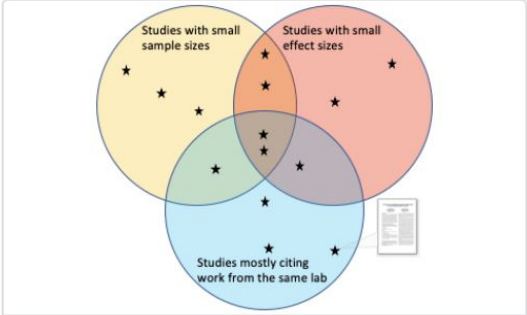
2X Wikidata

Other projects using KGTK to build KGs



Knowledge Graphs for Business

Creating a public resource containing knowledge about businesses, their



Scoring Scientific Research

Developing automated techniques for evaluating scientific claims and

https://usc-isi-i2.github.io/bokn/

https://usc-isi-i2.github.io/macro-score/

Tutorial Organization

Tutorial Program

Time (PDT - Los Angeles)	Content	Speaker
9:15 - 10:00	Intro to KGTK and Kypher	Hans
10:15 - 10:45	Profiling KGs	Pedro
10:45 - 11:15	Embeddings and similarity	Pedro
11:30 - 12:00	Extending KGs (CSV)	Pedro
12:00 - 12:30	Extending KGs (LOD)	Filip
12:30 - 13:00	Network analysis	Pedro
13:15 - 13:45	Analyzing Wikidata	Daniel
13:45 - 14:15	Commonsense KGs	Filip
14:15 - 15:00	Discussion	All

What You Will Learn

How to do analyses and transformations on KGs that you always wanted to do but didn't have the tools to do it

How to do sophisticated queries, do network analysis, work with embeddings, integrate CSV or LOD data

Interesting use cases that may inspire you to solve KG problems in your own work

That Wikidata is not scary-large and you can work with it on your laptop

A toolkit that is fun to use and integrates with Pandas and other tools that you like to use

Hands On

Runs on Google colab
Nothing to install

or just enjoy the show and try KGTK later

Introduction to KGTK

Profiling KGs

Embeddings

Extending with CSV data

Extending with LOD

Network analysis

Validating Wikidata constraints

Analysis of 300 Wikidata dumps

Building a commonsense reasoning KG

---

https://github.com/usc-isi-i2/kgtk-notebooks

Contacting us during the tutorial

slack

iswc-conf.slack.com
#kgtk-tutorial

zoom

chat, raise your hand or speak up

colab

if your colab doesn't work

- not the end of the world
- watch on zoom
- we'll try to get you going during a break

reminder:

- make a copy to your drive
- reset runtime after pip install kgtk==1.0.1