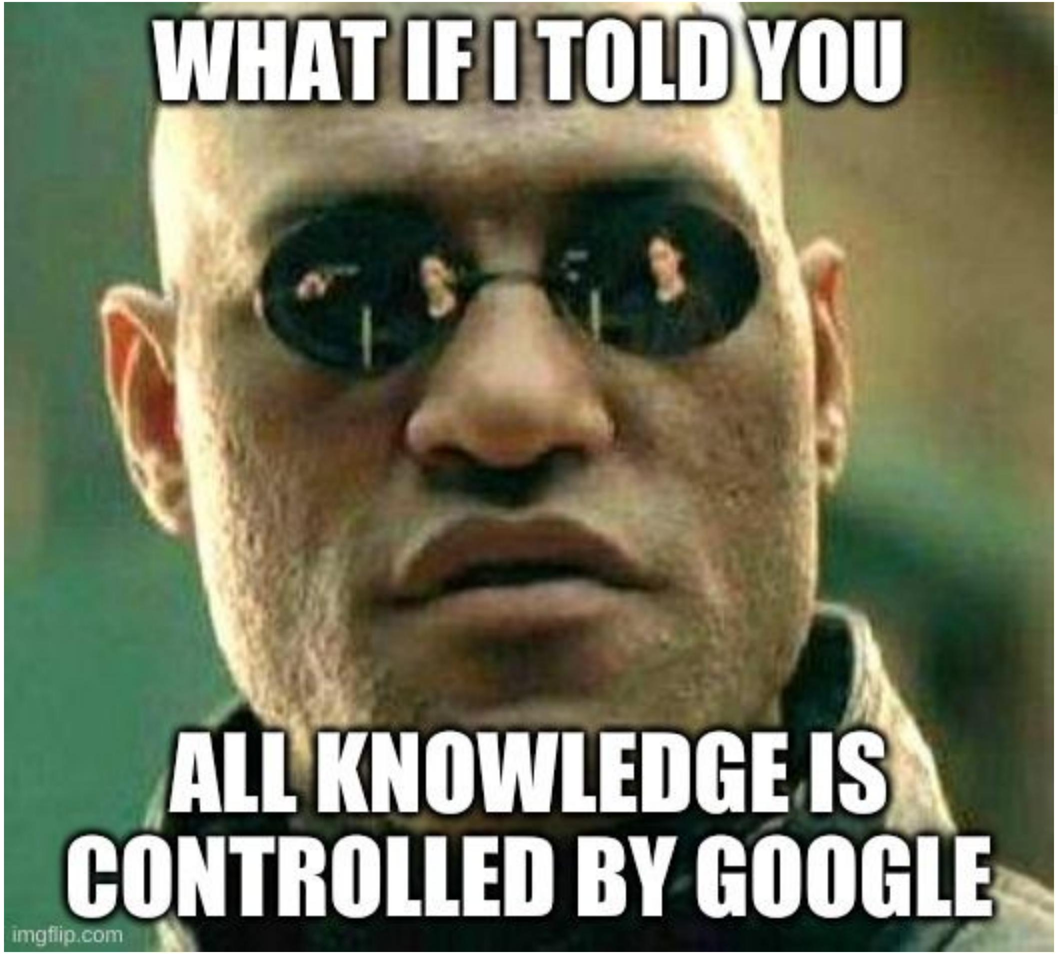
Knowledge Graphs Created through Basic Machine Learning

Clair J. Sullivan, PhD
Data Science Advocate

https://medium.com/@cj2001
@CJLovesData1

neo4j


WHAT IF I TOLD YOU
ALL KNOWLEDGE IS
CONTROLLED BY GOOGLE
imgflip.com

All materials for this demonstration are available on the workshop GitHub repo:

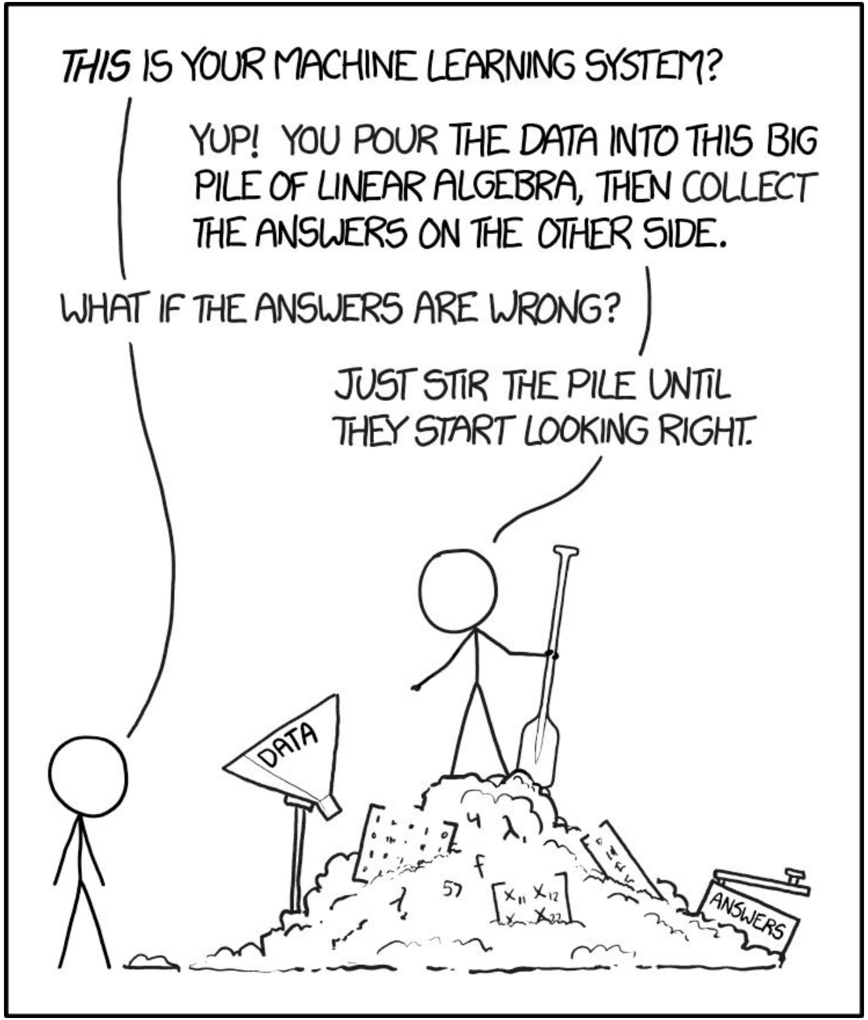
https://github.com/cj2001/nodes2021_kg_workshop

(I will put this link up several times!)

To run today’s code:

1. Jupyter or Google Colab
   - We will have some dependencies to manage in either
   - If you are bringing your own Jupyter, you probably want to create a virtual environment for this workshop
2. Neo4j Sandbox
   - https://dev.neo4j.com/sandbox

We can either populate the database manually, or I will show how to download a pre-populated one...


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

By the end of this workshop you will be able to...

1. Get some text and extract relevant information
2. Create a knowledge graph
3. Vectorize knowledge graph (create graph embeddings)
4. Apply data science and machine learning to knowledge graph

neo4j

By the end of this workshop you will be able to...

1. Get some text and extract relevant information
2. Create a knowledge graph
3. Vectorize knowledge graph (create graph embeddings)
4. Apply data science and machine learning to knowledge graph

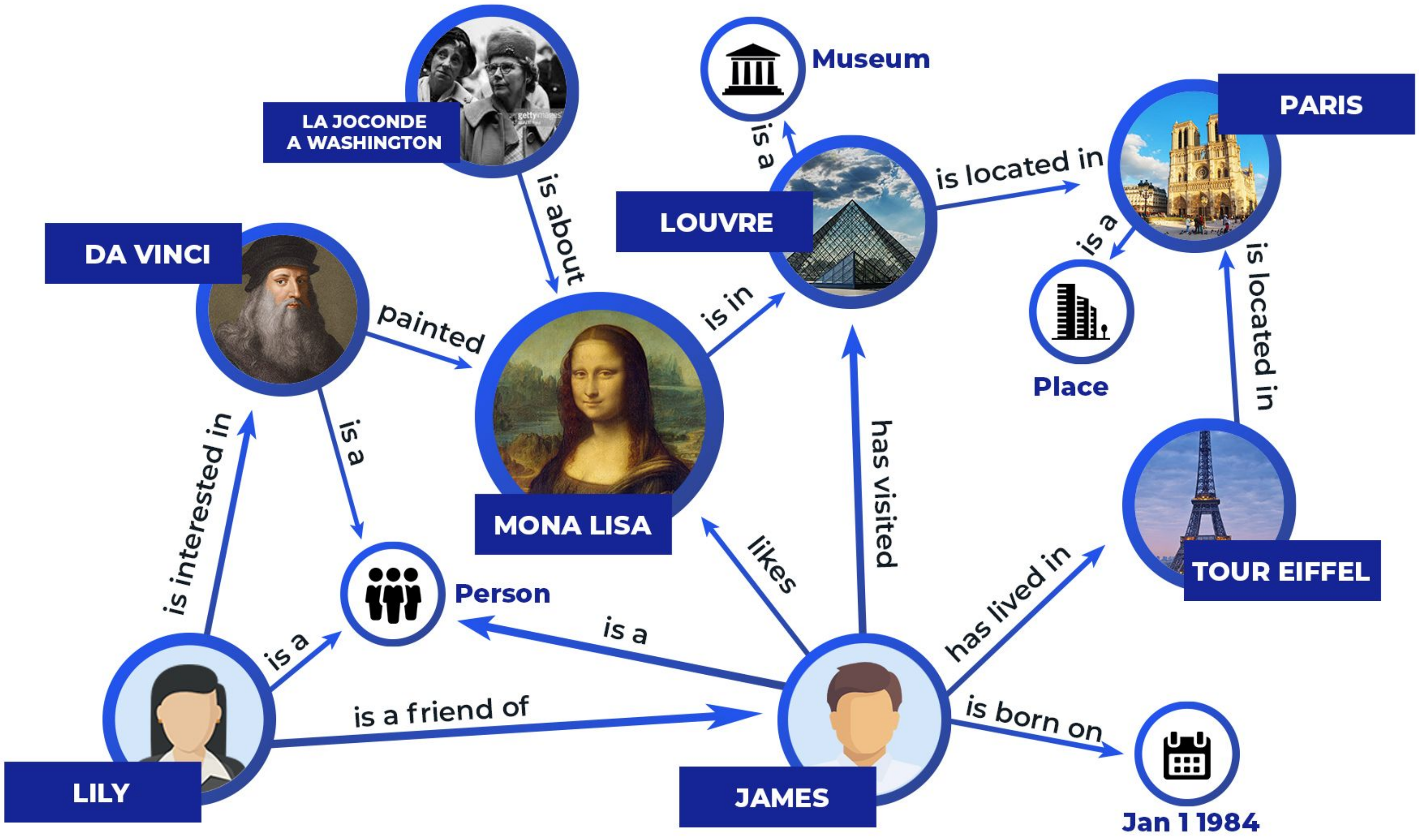
Natural Language Processing (NLP)

Graph Data Science Library + Basic ML

Two Key Concepts

1. There is no proverbial “silver bullet” with Natural Language Processing (NLP)

2. The quality of what you get out of a knowledge graph depends on the quality of what you put into it

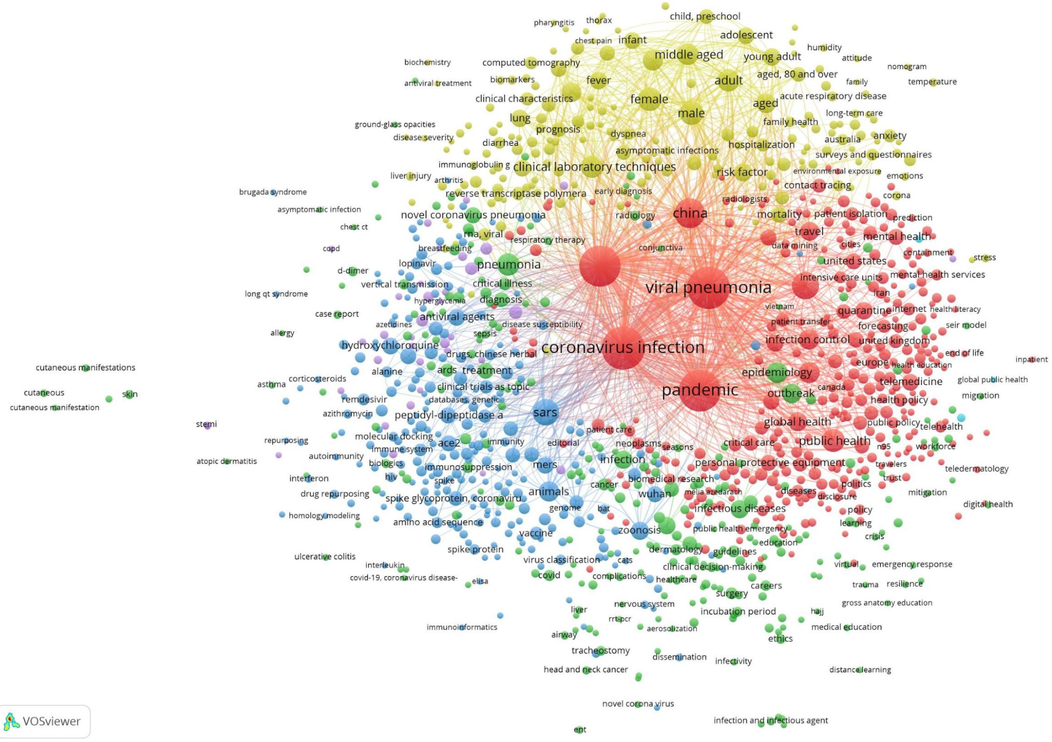


https://yashuseth.blog/2019/10/08/introduction-question-answering-knowledge-graphs-kgqa/

Introduction to knowledge graphs

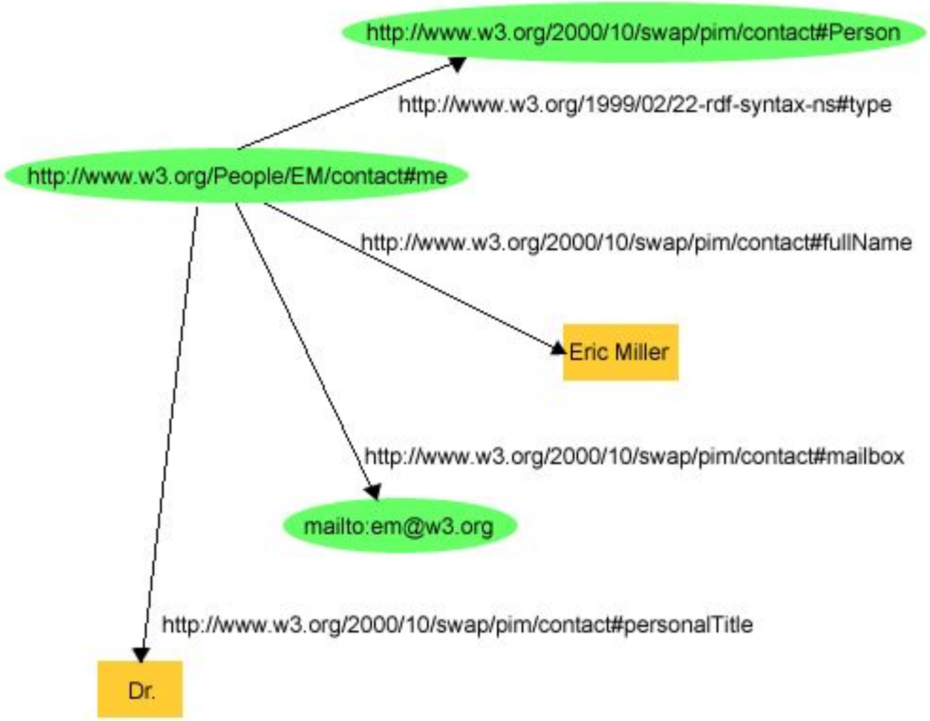
- “Things not strings”
- What knowledge graphs are useful for
  - Search
  - Question answering
  - Recommendation engine
- Can be generated a lot of different ways
  - Co-occurrence
  - Resource Description Framework (RDF)
  - Subject-Verb-Object (SVO)

Word co-occurrence



https://covid19biblio.com/2020/04/28/keyword-co-occurrence-network-graph-for-the-overall-research-field-on-covid-19-up-to-april-27th-2020/

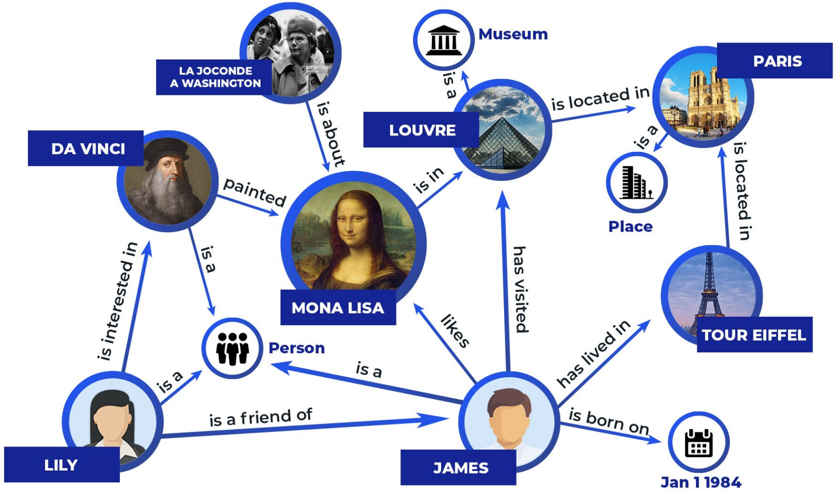
RDF triples



https://en.wikipedia.org/wiki/Resource_Description_Framework#Examples

SVO triples


Subject
:VERB
Object


LA JOCONDE A WASHINGTON
Museum
PARIS
is a
is located in
is about
LOUVRE
DA VINCI
is a
painted
is in
Place
is located in
is interested in
is a
has visited
MONA LISA
TOUR EIFFEL
likes
Person
has lived in
is a
is a
is a friend of
is born on
LILY
JAMES
Jan 1 1984

NLP considerations for knowledge graph creation

- Named Entity Recognition (NER)
- SVO / SPO triples
  - ...but verbs can be difficult to reliably detect via NLP!
- Very language dependent
- Very topic-area dependent

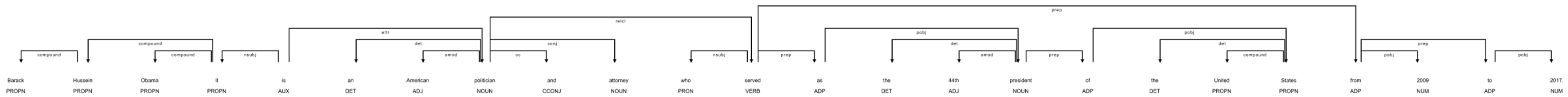
Barack Hussein Obama II PERSON ((listen) bə-RAHK hoo-SAYN oh-BAH-mə; born August 4, 1961 DATE) is an American NORP politician and attorney who served as the 44th ORDINAL president of the United States GPE from 2009 DATE to 2017 DATE . A member of the Democratic Party ORG , Obama PERSON was the first ORDINAL African-American NORP president of the United States GPE . He previously served as a U.S. GPE senator from Illinois GPE from 2005 to 2008 and as an Illinois GPE state senator from 1997 DATE to 2004.

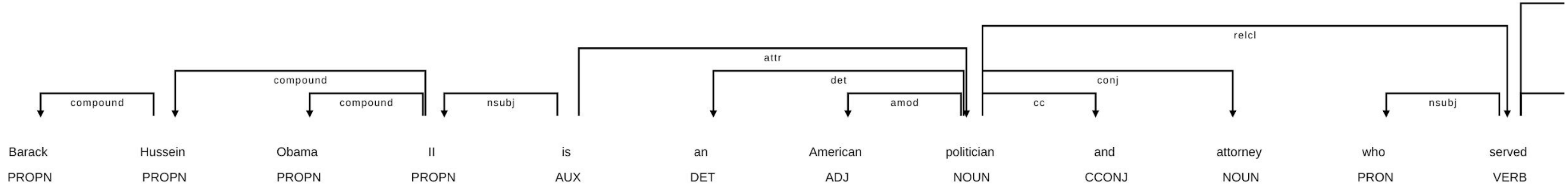
Obama PERSON was born in Honolulu GPE , Hawaii GPE . After graduating from Columbia University ORG in 1983 DATE , he worked as a community organizer in Chicago GPE . In 1988 DATE , he enrolled in Harvard Law School ORG , where he was the first ORDINAL black person to be president of the Harvard Law Review ORG . After graduating, he became a civil rights attorney and an academic, teaching constitutional law at the University of Chicago Law School ORG from 1992 DATE to 2004. Turning to elective politics, he represented the 13th ORDINAL district from 1997 DATE until 2004 DATE in the Illinois Senate, when he ran for the U.S. Senate ORG . Obama PERSON received national attention in 2004 DATE with his March Senate ORG primary win, his well-received July DATE Democratic National Convention keynote address, and his landslide November DATE election to the Senate ORG . In 2008 DATE , he was nominated by the Democratic Party ORG for president a year DATE after beginning his campaign, and after a close primary campaign against Hillary Clinton PERSON . Obama PERSON was elected over Republican NORP Senator John McCain PERSON in the general election and was inaugurated alongside his running mate, Joe Biden PERSON , on January 20, 2009 DATE . Nine months later DATE , he was named the 2009 DATE Nobel Peace Prize WORK_OF_ART laureate.

Obama PERSON signed many landmark bills into law during his first two years DATE in office. The main reforms that were passed include the Affordable Care Act LAW (commonly referred to as ACA ORG or " Obamacare WORK_OF_ART "), although without a public health insurance option, the Dodd–Frank Wall Street Reform and Consumer Protection Act, and the Don't Ask, Don't Tell Repeal Act of 2010 DATE . The American Recovery and Reinvestment Act ORG of 2009 DATE and Tax Relief, Unemployment Insurance Reauthorization, and Job Creation Act of 2010 DATE served as economic stimuli amidst the Great Recession EVENT . After a lengthy debate over the national debt limit, he signed the Budget Control ORG and the American Taxpayer Relief Acts ORG . In foreign policy, he increased U.S. GPE troop levels in Afghanistan GPE , reduced nuclear weapons with the United States– GPE Russia New START treaty, and ended military involvement in the Iraq War EVENT . He ordered military involvement in Libya GPE for the implementation of the UN Security Council ORG Resolution 1973 DATE , contributing to the overthrow of Muammar Gaddafi PERSON . He also ordered the military operations that resulted in the deaths of Osama bin Laden PERSON and suspected American NORP Al-Qaeda ORG operative Anwar al-Awlaki PERSON .

After winning re-election by defeating Republican NORP opponent Mitt Romney PERSON , Obama PERSON was sworn in for a second ORDINAL term in 2013 DATE . During this term, he promoted inclusion for LGBT Americans NORP . His administration filed briefs that urged the Supreme Court ORG to strike down same-sex marriage bans as unconstitutional (United States GPE v. Windsor PERSON and Obergefell ORG v. Hodges PERSON); same-sex marriage was legalized nationwide in 2015 DATE after the Court ORG ruled so in Obergefell ORG . He advocated for gun control in response to the Sandy Hook Elementary School ORG shooting, indicating support for a ban on assault weapons, and issued wide-ranging executive actions concerning global warming and immigration. In foreign policy, he ordered military intervention in Iraq GPE in response to gains made by ISIL ORG after the 2011 DATE withdrawal from Iraq GPE , continued the process of ending U.S. GPE combat operations in Afghanistan GPE in 2016 DATE , promoted discussions that led to the 2015 DATE Paris Agreement EVENT on global climate change, initiated sanctions against Russia GPE following the invasion in Ukraine GPE and again after interference in the 2016 DATE U.S. GPE elections, brokered the JCPOA ORG nuclear deal with Iran GPE , and normalized U.S. GPE relations with Cuba GPE . Obama PERSON nominated three CARDINAL justices to the Supreme Court ORG : Sonia Sotomayor PERSON and Elena Kagan PERSON were confirmed as justices, while Merrick Garland PERSON faced partisan obstruction from the Republican NORP -led Senate ORG led by Mitch McConnell PERSON , which never held hearings or a vote on the nomination. Obama PERSON left office in January 2017 DATE and continues to reside in Washington GPE , D.C.During Obama's PERSON term in office, the United States' GPE reputation abroad, as well as the American NORP economy, significantly improved. Obama PERSON 's presidency has generally been regarded favorably, and evaluations of his presidency among historians, political scientists, and the general public frequently place him among the upper tier of American NORP presidents.

Barack Hussein Obama II is an American politician and attorney who served as the 44th president of the United States from 2009 to 2017.





Barack Hussein Obama II is an American politician and attorney who served as the 44th president of the United States from 2009 to 2017.

Text	Lemma	Tag	POS	DEP	is_stop
Barack	Barack	NNP	PROPN	compound	FALSE
Hussein	Hussein	NNP	PROPN	compound	FALSE
Obama	Obama	NNP	PROPN	compound	FALSE
II	II	NNP	PROPN	nsubj	FALSE
is	be	VBZ	AUX	ROOT	TRUE
an	an	DT	DET	det	TRUE
American	american	JJ	ADJ	amod	FALSE
politician	politician	NN	NOUN	attr	FALSE
and	and	CC	CCONJ	cc	TRUE
attorney	attorney	NN	NOUN	conj	FALSE
who	who	WP	PRON	nsubj	TRUE
served	serve	VBD	VERB	relcl	FALSE
as	as	IN	ADP	prep	TRUE
the	the	DT	DET	det	TRUE
44th	44th	JJ	ADJ	amod	FALSE
president	president	NN	NOUN	pobj	FALSE
of	of	IN	ADP	prep	TRUE
the	the	DT	DET	det	TRUE
United	United	NNP	PROPN	compound	FALSE
States	States	NNP	PROPN	pobj	FALSE
from	from	IN	ADP	prep	TRUE
2009	2009	CD	NUM	pobj	FALSE
to	to	IN	ADP	prep	TRUE
2017	2017	CD	NUM	pobj	FALSE
.	.	.	PUNCT	punct	FALSE

https://github.com/explosion/spaCy/blob/master/spacy/glossary.py

Barack Hussein Obama II is an American politician and attorney who served as the 44th president of the United States from 2009 to 2017.

Text	Lemma	Tag	POS	DEP	is_stop
Barack	Barack	NNP	PROPN	compound	FALSE
Hussein	Hussein	NNP	PROPN	compound	FALSE
Obama	Obama	NNP	PROPN	compound	FALSE
II	II	NNP	PROPN	nsubj	FALSE
is	be	VBZ	AUX	**ROOT**	TRUE
an	an	DT	DET	det	TRUE
American	american	JJ	ADJ	amod	FALSE
politician	politician	NN	NOUN	attr	FALSE
and	and	CC	CCONJ	cc	TRUE
attorney	attorney	NN	NOUN	conj	FALSE
who	who	WP	PRON	nsubj	TRUE
served	serve	VBD	VERB	relcl	FALSE
as	as	IN	ADP	prep	TRUE
the	the	DT	DET	det	TRUE
44th	44th	JJ	ADJ	amod	FALSE
president	president	NN	NOUN	pobj	FALSE
of	of	IN	ADP	prep	TRUE
the	the	DT	DET	det	TRUE
United	United	NNP	PROPN	compound	FALSE
States	States	NNP	PROPN	pobj	FALSE
from	from	IN	ADP	prep	TRUE
2009	2009	CD	NUM	pobj	FALSE
to	to	IN	ADP	prep	TRUE
2017	2017	CD	NUM	pobj	FALSE
.	.	.	PUNCT	punct	FALSE

https://github.com/explosion/spaCy/blob/master/spacy/glossary.py

Barack Hussein Obama II is an American politician and attorney who served as the 44th president of the United States from 2009 to 2017.

Text	Lemma	Tag	POS	DEP	is_stop
Barack	Barack	NNP	PROPN	compound	FALSE
Hussein	Hussein	NNP	PROPN	compound	FALSE
Obama	Obama	NNP	PROPN	compound	FALSE
II	II	NNP	PROPN	nsubj	FALSE
is	be	VBZ	AUX	ROOT	TRUE
an	an	DT	DET	det	TRUE
American	american	JJ	ADJ	amod	FALSE
politician	politician	NN	NOUN	attr	FALSE
and	and	CC	CCONJ	cc	TRUE
attorney	attorney	NN	NOUN	conj	FALSE
who	who	WP	PRON	nsubj	TRUE
served	serve	VBD	VERB	relcl	FALSE
as	as	IN	ADP	prep	TRUE
the	the	DT	DET	det	TRUE
44th	44th	JJ	ADJ	amod	FALSE
president	president	NN	NOUN	pobj	FALSE
of	of	IN	ADP	prep	TRUE
the	the	DT	DET	det	TRUE
United	United	NNP	PROPN	compound	FALSE
States	States	NNP	PROPN	pobj	FALSE
from	from	IN	ADP	prep	TRUE
2009	2009	CD	NUM	pobj	FALSE
to	to	IN	ADP	prep	TRUE
2017	2017	CD	NUM	pobj	FALSE
.	.	.	PUNCT	punct	FALSE

https://github.com/explosion/spaCy/blob/master/spacy/glossary.py

An introduction to the tools we will use today

- spacy
- Wikipedia Python package
- Google Knowledge Graph
- Pywikibot
- Neo4j
  - Awesome Procedures on Cypher (APOC)
  - Graph Data Science (GDS) Library
  - Cypher

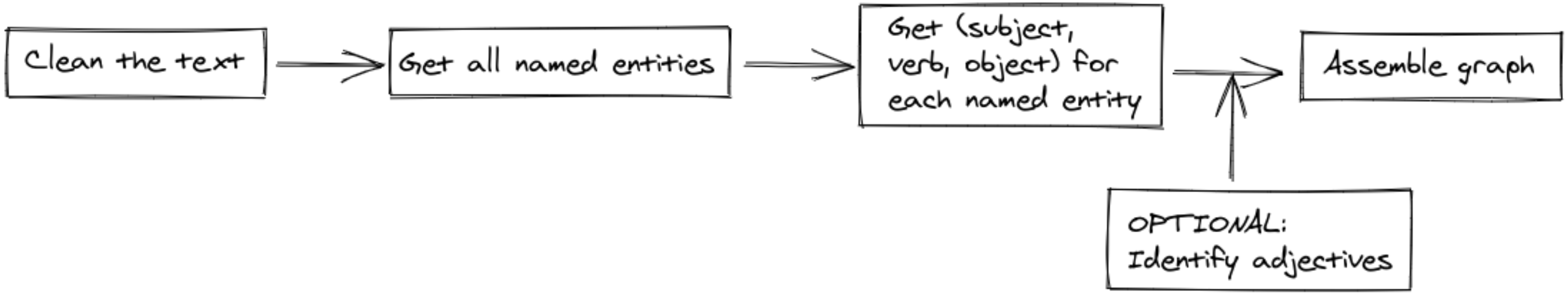
neo4j

Clone the GitHub repository at (OPTIONAL)

https://github.com/cj2001/nodes2021_kg_workshop

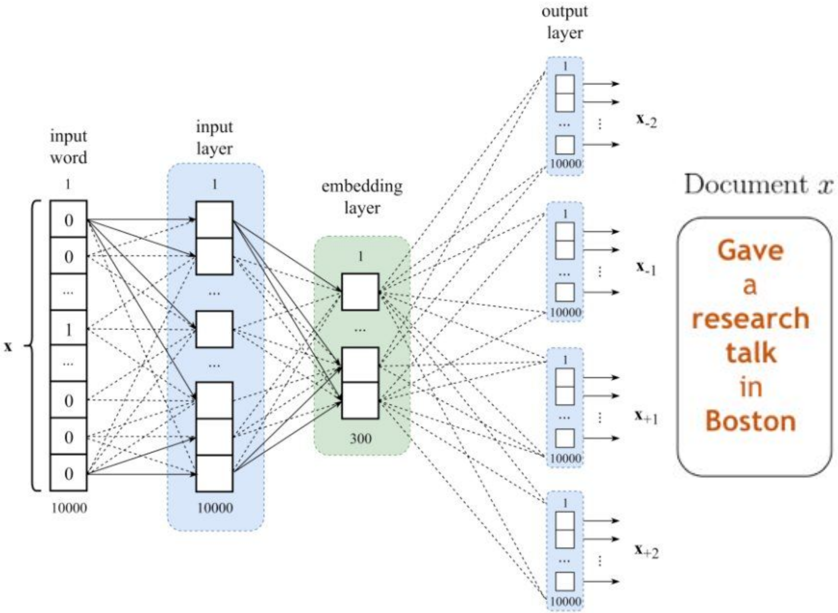
Method 1: The NLP Only Approach

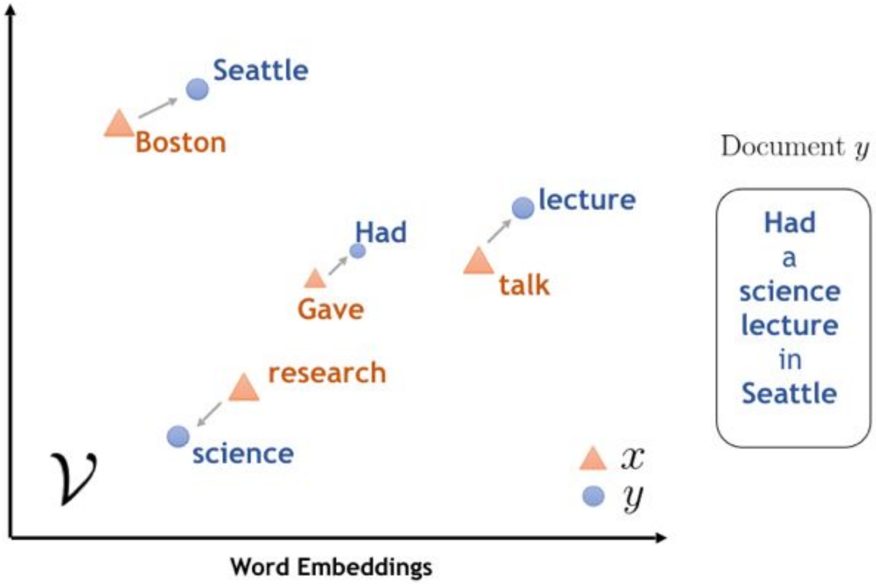
Some ways we could get this done: NLP only approach



Advantage: limitless verbs
Drawback: entity disambiguation

word2vec

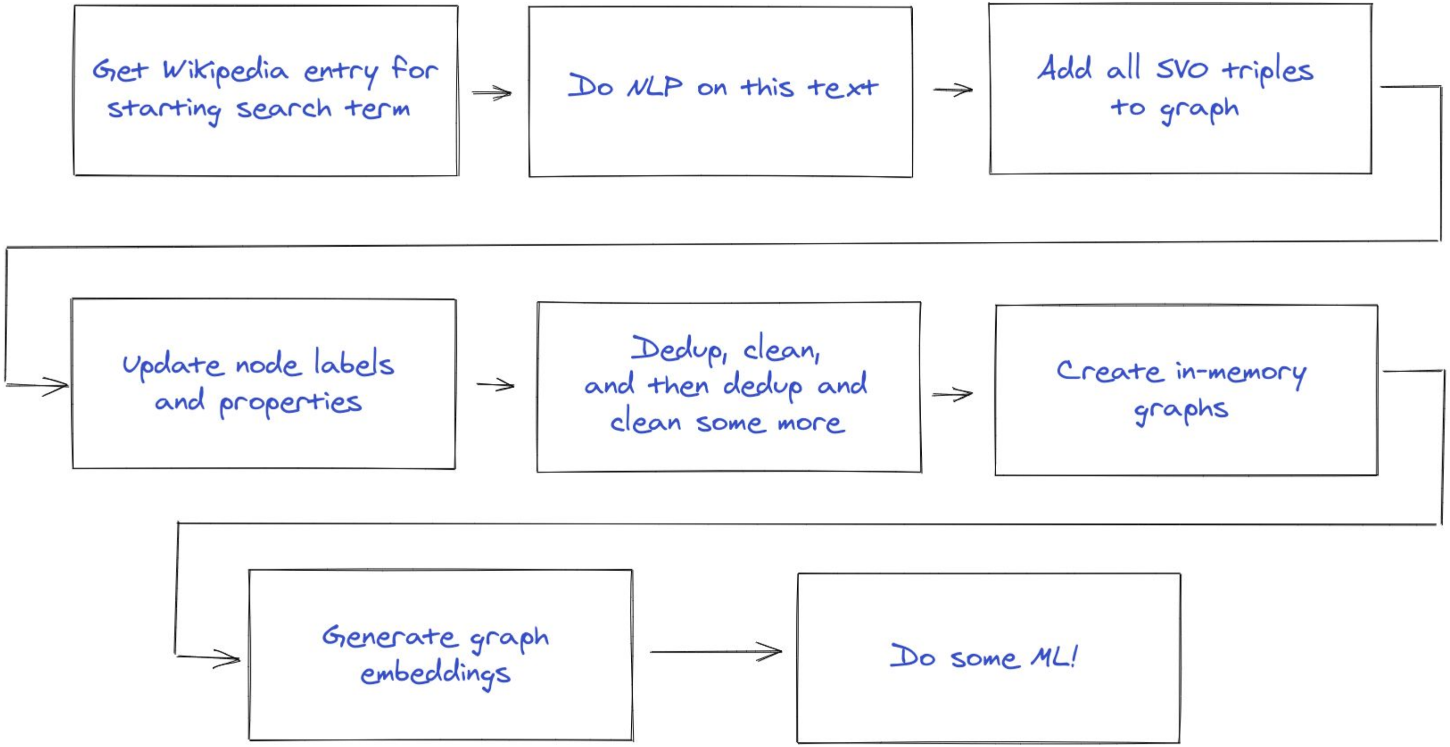




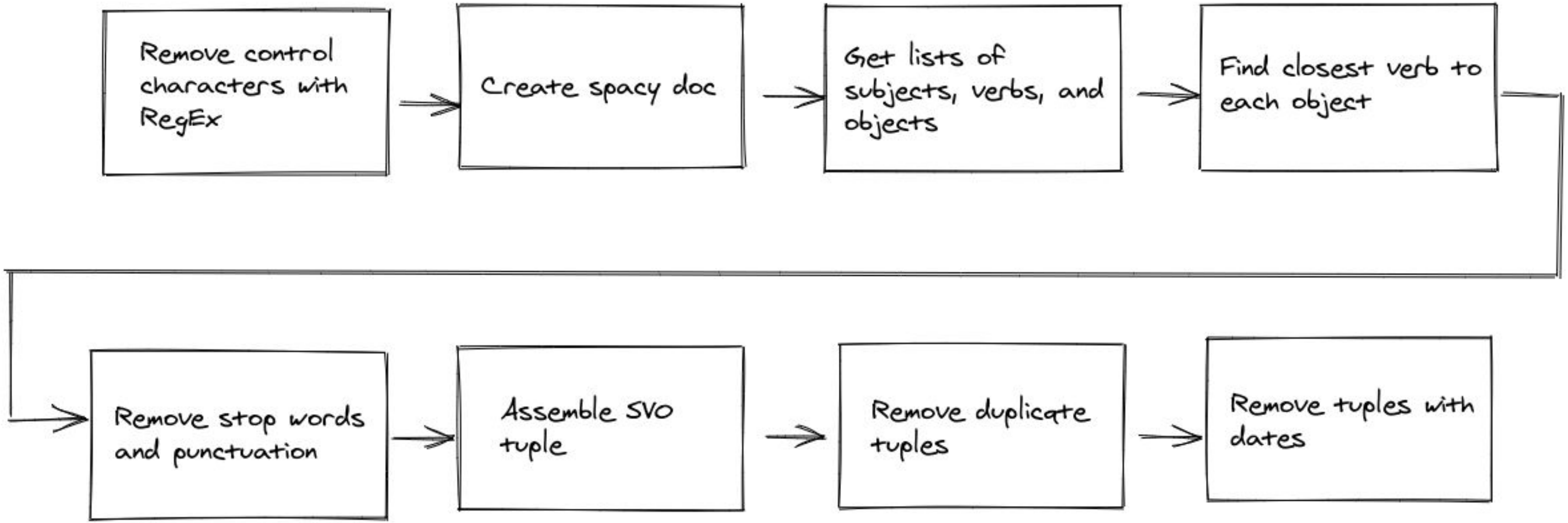
https://www.kdnuggets.com/2019/01/burkov-self-supervised-learning-word-embeddings.html

https://medium.com/swlh/word2vec-in-practice-for-natural-language-processing-a179b3286a21

Overview of workflow



NLP workflow



neo4j

Create a Google Knowledge API key

https://developers.google.com/knowledge-graph/how-tos/authorizing

Home › Search Central › Knowledge Graph Search API

Rate and review

Authorize Requests

When your application requests public data, the request doesn't need to be authorized, but does need to be accompanied by an identifier, such as an API key.

Your application needs to identify itself every time it sends a request to the Google Knowledge Graph Search API, by including an API key with each request.

Acquiring and using an API key

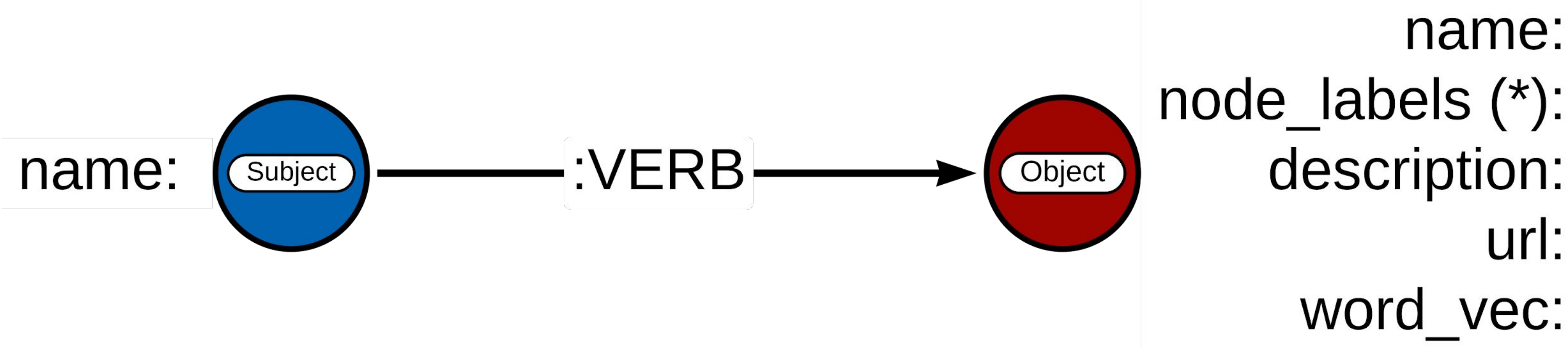
To acquire an API key:

1. Open the Credentials page in the API Console.
2. This API supports two types of credentials. Create whichever credentials are appropriate for your project:

Enhance the existing data with Google Knowledge Graph

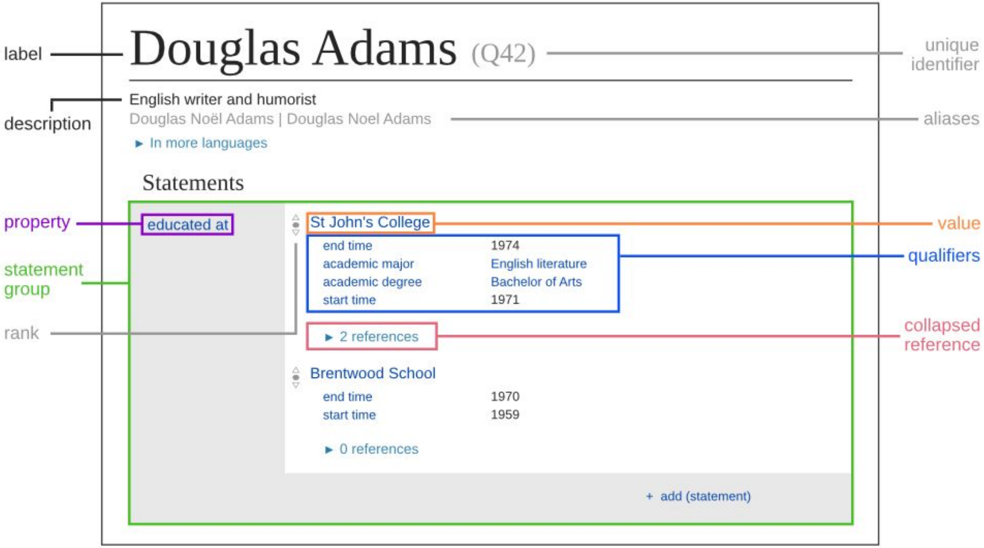

```
{...
  "@type": "ItemList",
  "itemListElement": [
    {
      "@type": "EntitySearchResult",
      "result": {
        "@id": "kg:/m/0dl567",
        "name": "Taylor Swift",
        "@type": [
          "Thing",
          "Person"
        ],
...
        "detailedDescription": {
          "articleBody": "Taylor Alison Swift is an American singer-songwriter and
actress. Raised in Wyomissing, Pennsylvania, she moved to Nashville, Tennessee, at the
age of 14 to pursue a career in country music. ",
          "url": "http://en.wikipedia.org/wiki/Taylor_Swift",
...
}
```

Detailed knowledge graph data model



Method 2: The NLP Lite Approach

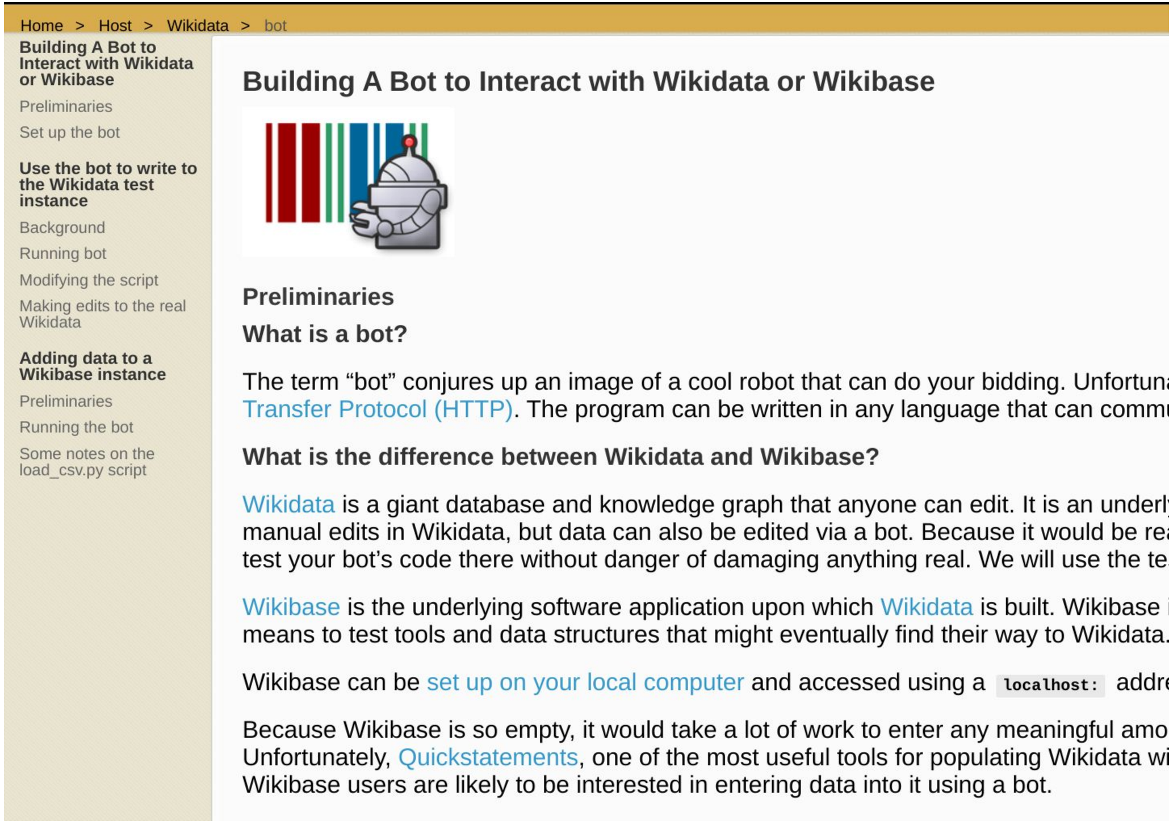
Some ways we could get this done: NLP “lite”



Advantage: entity disambiguation
Drawback: must specify which verbs you are interested in

Create a PyWikiBot token

https://heardlibrary.github.io/digital-scholarship/host/wikidata/bot/



Home > Host > Wikidata > bot

Building A Bot to Interact with Wikidata or Wikibase

Preliminaries

Set up the bot

Use the bot to write to the Wikidata test instance

Background

Running bot

Modifying the script

Making edits to the real Wikidata

Adding data to a Wikibase instance

Preliminaries

Running the bot

Some notes on the load_csv.py script

Building A Bot to Interact with Wikidata or Wikibase

Preliminaries

What is a bot?

The term "bot" conjures up an image of a cool robot that can do your bidding. Unfortuna
Transfer Protocol (HTTP). The program can be written in any language that can commu

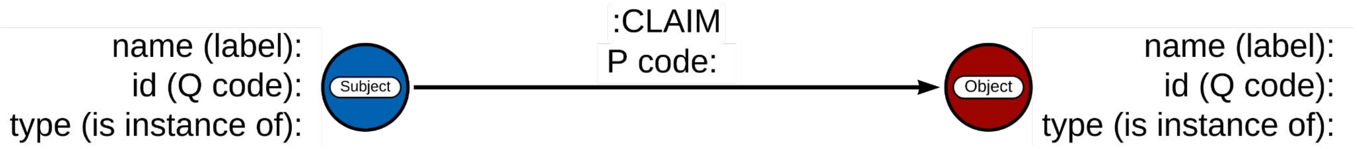
What is the difference between Wikidata and Wikibase?

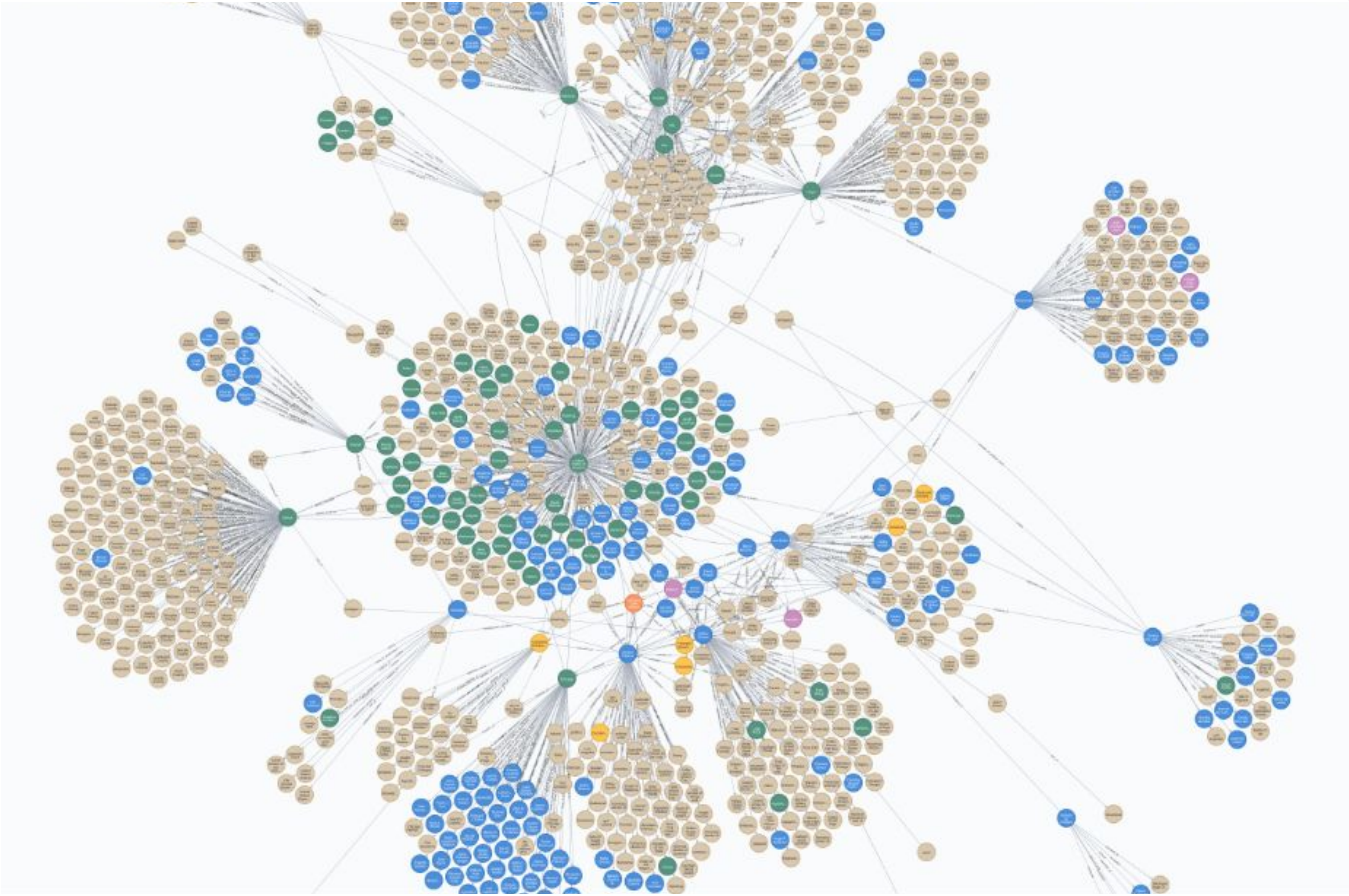
Wikidata is a giant database and knowledge graph that anyone can edit. It is an underl
manual edits in Wikidata, but data can also be edited via a bot. Because it would be re
test your bot's code there without danger of damaging anything real. We will use the te

Wikibase is the underlying software application upon which Wikidata is built. Wikibase i
means to test tools and data structures that might eventually find their way to Wikidata.

Wikibase can be set up on your local computer and accessed using a `localhost:` addr

Because Wikibase is so empty, it would take a lot of work to enter any meaningful amo
Unfortunately, Quickstatements, one of the most useful tools for populating Wikidata wi
Wikibase users are likely to be interested in entering data into it using a bot.


name (label):
id (Q code):
type (is instance of):
Subject
:CLAIM
P code:
Object
name (label):
id (Q code):
type (is instance of):



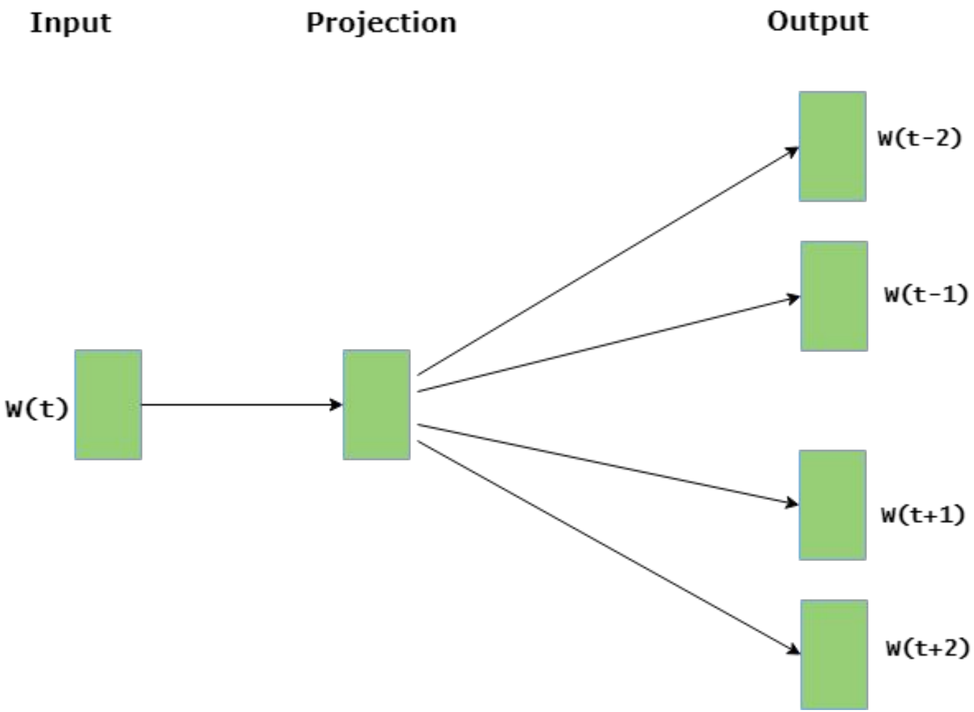
Machine Learning on Graphs

Why machine learning with (knowledge) graphs?

- Traditional ML uses a relational database-type model
  - All data points are are independent of each other
    - Example: churn prediction based on user behavior
- Graphs (and graph databases) treat relationships as a “first class citizen”
  - Models can include homophily
    - Example: churn prediction includes the churn of neighbors within the graph
  - Models can also include the same data as the traditional, independent data point models

Example: making a better recommendation engine

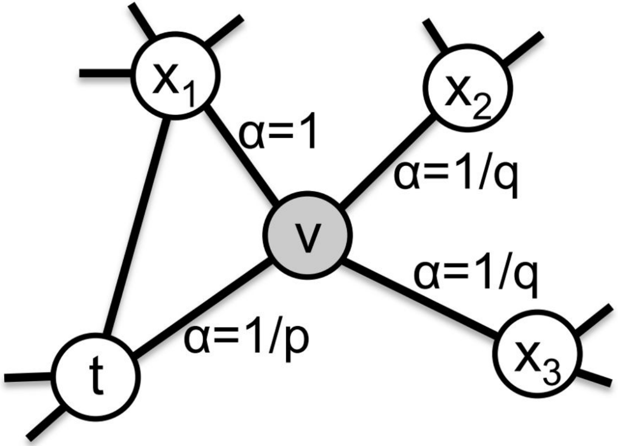
word2vec

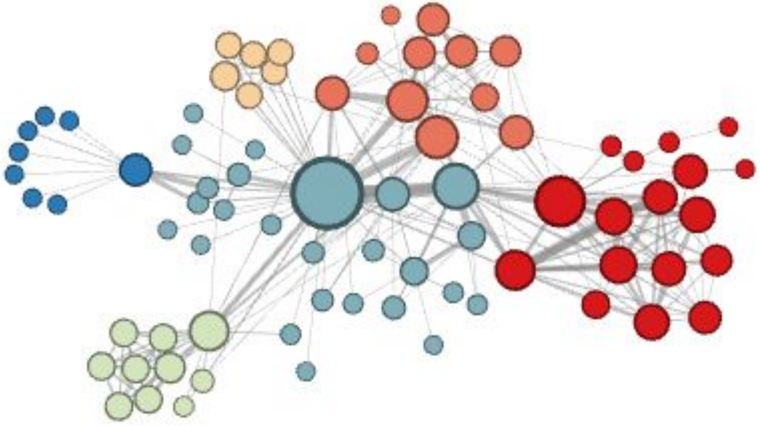


Note: word2vec typically creates one vector per word. The spacy implementation of vectorization takes a document (sentence) and averages the word vectors across the sentence.

https://www.geeksforgeeks.org/python-word-embedding-using-word2vec/

node2vec





https://snap.stanford.edu/node2vec/

node2vec

	name	embedding
1	"Barack Obama"	[-0.5534299612045288, 0.319044291973114, -0.06302239000797272, 0.8757740259170532, -0.5034562945365906, 0.23735041916370392, 0.15117834508419037, 0.8566229939460754, 0.36209946870803833, -0.6797583103179932]
2	"United States of America"	[0.1451471447944641, 0.4807624816894531, 0.4366985261440277, 0.42269086837768555, 0.20110560953617096, -0.5779915452003479, -0.10465864837169647, -0.04139380902051926, 0.3290615975856781, -0.495746374130249]
3	"Democratic Party"	[0.19489407539367676, 0.5242791771888733, -0.7722358703613281, 0.15843135118484497, -0.5232059955596924, 0.5803580284118652, 0.17970407009124756, -0.6052649617195129, -0.7680787444114685, 0.5753467679023743]
4	"Illinois"	[0.8053464293479919, 0.2300983965396881, 0.7079187035560608, -0.03613480180501938, 0.4691833555698395, -0.38967591524124146, 0.0824178084731102, -0.27919191122055054, 0.23097757995128632, 0.7676025032997131]
5	"Honolulu"	[0.5645546317100525, -0.8217434287071228, 0.5297825336456299, 0.31918787956237793, -0.4469479024410248, 0.8200243711471558, -0.4185393154621124, -0.3348398506641388, -0.5732030272483826, -0.39435911178588867]
6	"Hawaii"	[-0.5108500719070435, 0.12454281747341156, 0.2569912075996399, -0.6748168468475342, 0.24597491323947906, 0.5893328189849854, 0.26128533482551575, -0.7081990838050842, -0.5841189622879028, 0.13823509216308594]

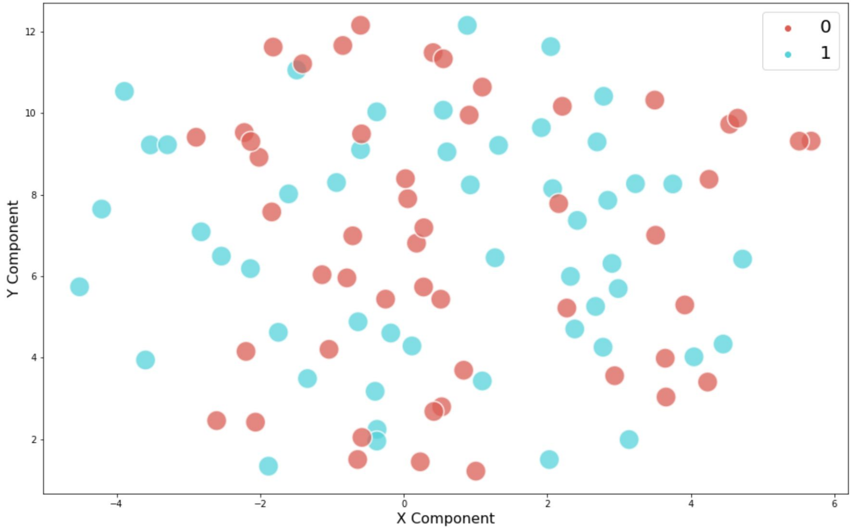
Embedding dimension: 10

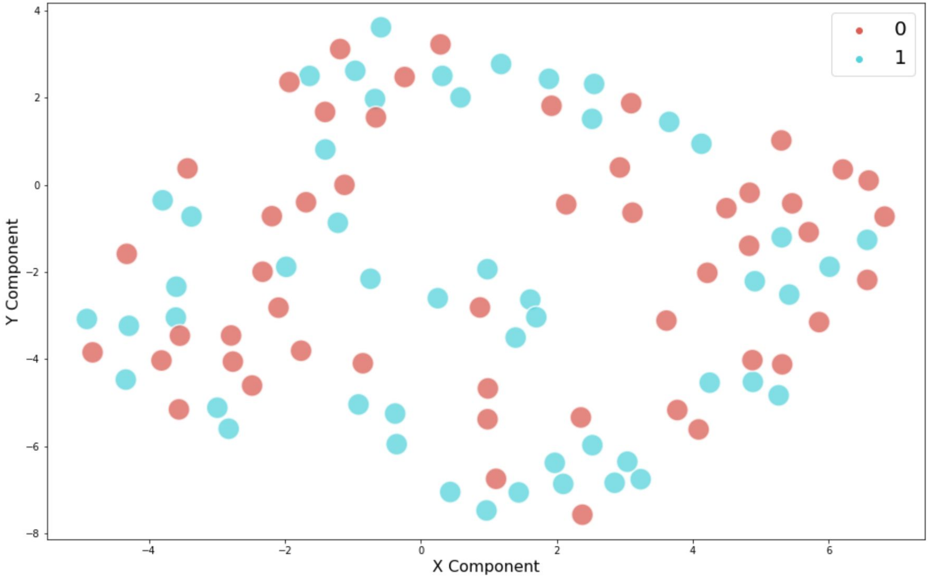
Node similarity via embeddings

	u.name	n.name	n.type_ls	similarity
1	"mixed-sex education"	"Columbia University"	["private university"]	0.13848771879851898
2	"capital"	"Washington, D.C."	["capital"]	0.1296056718640691
3	"landlocked country"	"Afghanistan"	["sovereign state"]	0.11288100303291404
4	"big city"	"Chicago"	["city of the United States"]	0.08881633377710593
5	"social state"	"Ukraine"	["sovereign state"]	0.08480072967702501

Embedding dimension: 300

Visualizing embeddings with t-SNE





Clone the GitHub repository at (OPTIONAL)

https://github.com/cj2001/nodes2021_kg_workshop

Where to go from here?

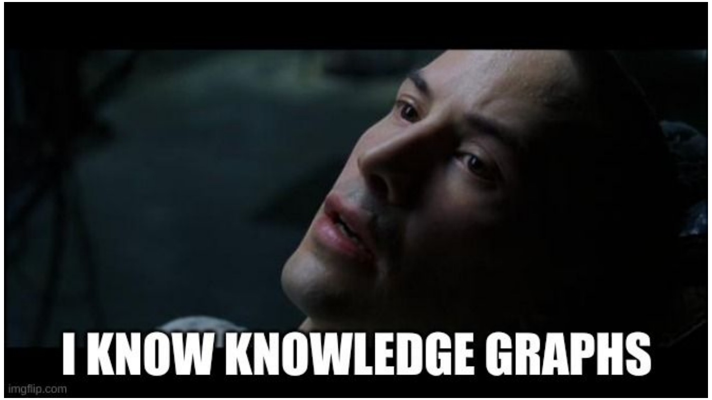
Two Key Concepts

1. There is no proverbial “silver bullet” with Natural Language Processing (NLP)

2. The quality of what you get out of a knowledge graph depends on the quality of what you put into it

What could we do from here?

- Add nodes to the graph!
- Various embedding optimization techniques
- Add data for creating embeddings
    - Ex: Word vectors from text descriptions
- Different embedding/modeling techniques
    - GraphSAGE
    - GNN



Problems we could solve

- Community/cluster detection
- Node classification, link prediction
- Graph-to-graph classification
- Unstructured text, NLP
- Question answering systems

Thank you!

https://medium.com/@cj2001
@CJLovesData1