cs520

A survey of open source tools for building knowledge graphs

Naren Chittar

Disclaimer: My own views. No affiliation to any software. Fan of several.

June 2021

Topics

- Architecting software today
- Joining lists
- Information extraction
- Graph databases
- Graph compute engines

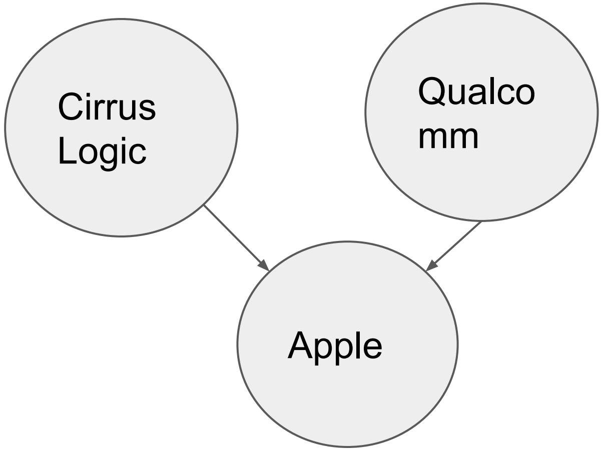
Architecting software today

- Minimal code from scratch, reuse code
- Decompose high level problem and map to open source components
- Too many choices. Ranking problem
  - Github, Apache
  - Stars, forks, contributors, Update dates, companies logos
- Write code to configure, glue, scale

Joining lists: a common task

Company Name	Address	Revenue
Apple Inc	One Apple Park Way, Cupertino CA 95014	$100B

Company Name	Address	Suppliers
Apple Computers	1 Apple Pkwy, Cupertino CA	Qualcomm, Cirrus Logic



Joining lists : Dedupe

- https://github.com/dedupeio/dedupe
- Join multiple lists
- Remove duplicates
- Data type
  - Text, Short Text, Date Time,...
  - Price, Address, Name, Phone, Lat-Long,...
- Blocking
- Unsupervised learning (cosine distance. threshold)
- Supervised learning (L2 regularized logistic regression classifier)
- Active learning

Person Name Matching

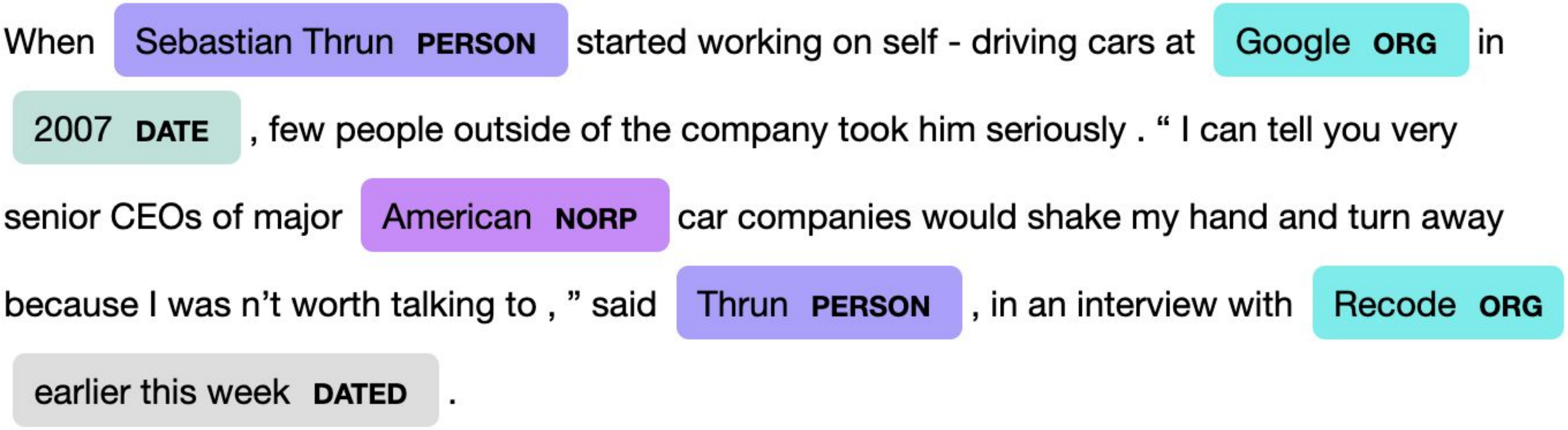
- Bill Gates, WIlliam Gates
  - https://github.com/carltonnorthern/nickname-and-diminutive-names-lookup
  -
- (Jennifer, Jenifer), (Kaitlin, Kaitlyn)
  - Soundex
  - Metaphone
  - Double Metaphone

Comparison of record linkage packages

https://github.com/J535D165/data-matching-software

Software	API	GUI	Linking	Deduplication	Supervised Learning	Unsupervised Lrng	Active Lrng
Atylmo	Pyspark	❌	✅	✅	❌	❌	❌
Depupe	Python	❌	✅	✅	✅	❌	✅
fastLink	R	❌	✅	?	❌	✅	❌
FEBRL	Python	✅	✅	✅	❌	❌	❌
FRIL	Java	✅	✅	❌	?	✅	❌
Fuzzy Matcher	Python	❌	✅	❌	❌	✅	❌
JedAI	Java	✅	✅	?	✅	?	?
PRIL	SQL	❌	✅	?	?	?	?
RecordLinkage	R	❌	✅	✅	✅	✅	❌
RELAIS	❌	✅	✅	?	?	✅	❌
ReMaDDer	❌	✅	✅	✅	❌	✅	❌
The Link King	❌	✅	✅	✅	?	✅	❌

Spacy: Information Extraction

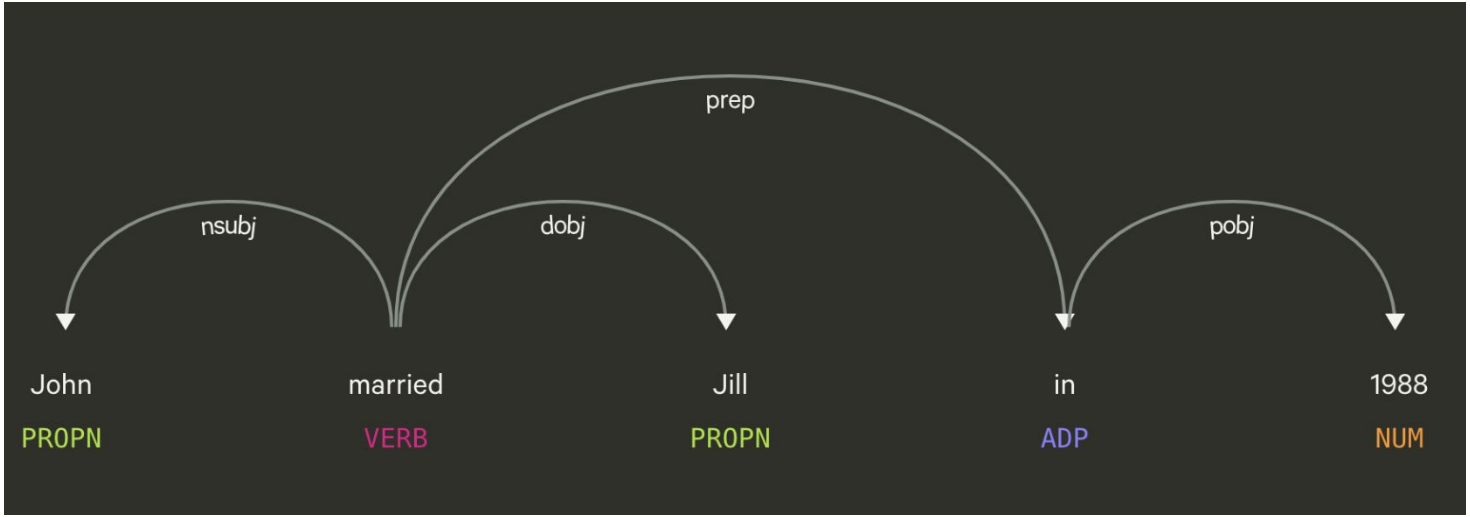


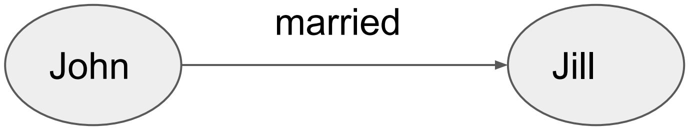
When Sebastian Thrun **PERSON** started working on self - driving cars at Google **ORG** in 2007 **DATE** , few people outside of the company took him seriously . “ I can tell you very senior CEOs of major American **NORP** car companies would shake my hand and turn away because I was n’t worth talking to , ” said Thrun **PERSON** , in an interview with Recode **ORG** earlier this week **DATED** .

- Pre built models for POS, NER
- Noun chunks
- NER, Custom NER
- Entity Linking framework
- https://pypi.org/project/spacy-entity-linker/

Spacy: Dependency Parsing

- Contains information to build graphs
- Subject-Predicate-Object Triple (multiple implementations)





SpikeX

- SpaCy Pipes for Knowledge Extraction : A collection of pipes ready to be plugged in a spaCy pipeline
- **WikiPageX** links Wikipedia pages to chunks in text
- **ClusterX** picks noun chunks in a text and clusters them based on a revisiting of the Ball Mapper algorithm, Radial Ball Mapper
- **AbbrX** detects abbreviations and acronyms, linking them to their long form. It is based on scispacy's one with improvements
- **LabelX** takes labelings of pattern matching expressions and catches them in a text, solving overlappings, abbreviations and acronyms
- **PhraseX** creates a Doc's underscore extension based on a custom attribute name and phrase patterns. Examples are **NounPhraseX** and **VerbPhraseX**, which extract noun phrases and verb phrases, respectively
- **SentX** detects sentences in a text, based on Splitta with refinements

Graph Databases

- In memory/disk based
- OLTP rather than OLAP
- Distributed
- ACID
- Query language
  - No standardization yet link SQL
  - Gremlin, SPARQL, Cypher, custom
- Managed
- Pricing
- Support and ecosystem

Neo4J Community

- Property graph
- Scales horizontally
- Data access controls
- Declarative query language : Cypher
- Drivers for multiple programming languages for etl as well as query
- Options for managed, failover, backups in professional version
- https://web.stanford.edu/class/cs520/2020/abstracts/rathle.html

Amazon Neptune (closed source)

- Managed
- Property Graph and RDF. Apache TinkerPop Gremlin and SPARQL
- ACID
- Continuous backup to Amazon S3 and point-in-time recovery
- Replication across Availability Zones
- Security: HTTPS, Encryption at rest
- Auditing
- Three higher level applications using graphs
  - Knowledge Graphs
  - Identity Graphs
  - Fraud Detection

Graph Compute Engines

- Run analytics and ML on graphs offline (OLAP rather than OLTP)
- Usually different software than graph dbs
- Some like Neo4j offer both

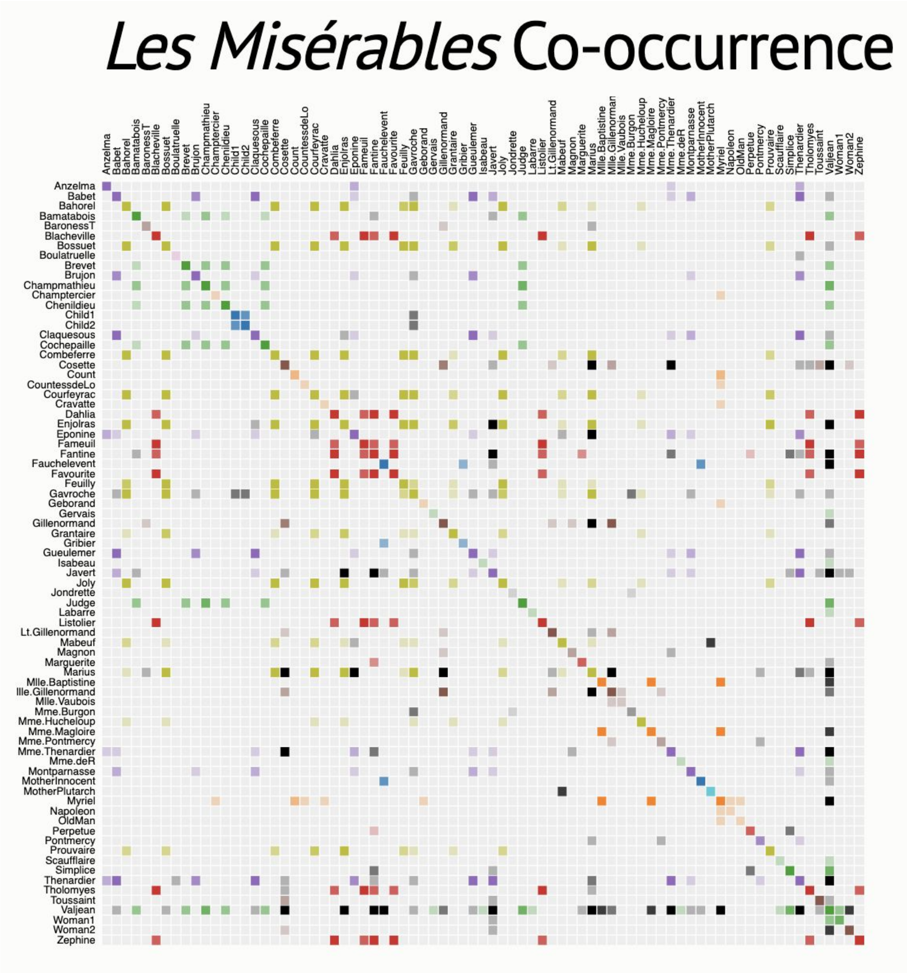
networkx

- Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks
- 500 contributors, 9k stars, 2k forks
- In memory
- Python interface
- Flexible: all NetworkX graph classes allow (hashable) Python objects as nodes and any Python object can be assigned as an edge attribute
- Rich in algorithms (click. Too many to show)

```
>>> import networkx as nx
>>> G = nx.Graph()
>>> G.add_edge('A', 'B', weight=4)
>>> G.add_edge('B', 'D', weight=2)
>>> G.add_edge('A', 'C', weight=3)
>>> G.add_edge('C', 'D', weight=4)
>>> nx.shortest_path(G, 'A', 'D',
weight='weight')
['A', 'B', 'D']
```

Networkx community detection .

[link]



Apache Spark: GraphX

- Parallel graph computation with *RDD* (Resilient Distributed Dataset).
- 100s of commodity nodes. Robust to node failures
- *Graph* extends *RDD*
- Property operators
- Structural operators
- Join operators
- Neighborhood operators
- Discrete algorithms
  - Partitioning
  - ConnectedComponents
  - TriangleCounting
- ML algorithms
  - PageRank

Q&A

Acknowledgements for help with slides to Wanying Ding, Manoj Cherukumalli, Aarohi Chittar