Machine Learning With Graphs: Going Beyond Tabular Data

Clair J. Sullivan, PhD
Data Science Advocate
Twitter: @CJLovesData1
Medium: https://medium.com/@cj2001

What we are going to do today

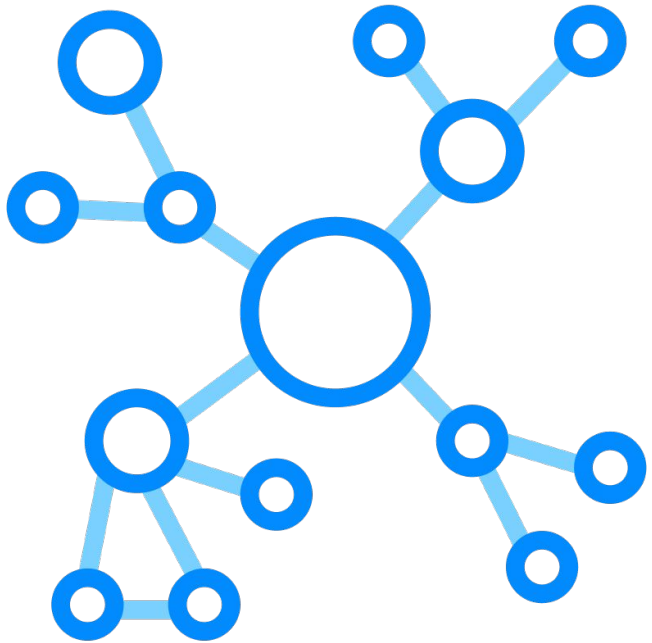
- Why graphs? Why not just SQL?
- How do I know if I have a “graph-y” problem?
- A very brief introduction to graph theory
- Graph machine learning (ML)
- Wrap up

Two Key Concepts

1. It can be good to break the assumption that individual data points are independent

2. Modeling relationships can result in models that are less noisy, more accurate

What is a graph?



Common examples

- Social media
- Internet routing
- Maps, wayfinding
- Recommender systems
- Search
- Knowledge graphs, question answering

Columnar data for churn prediction

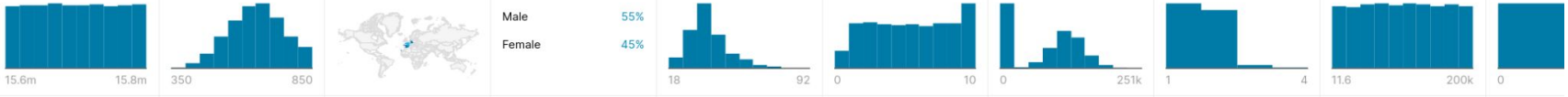
Churn_Modelling.csv (684.86 kB)

Detail | Compact | Column

10 of 14 columns

About this file

Based upon data of employees of a bank we calculate whether a employee stands a chance to stay in the company or not.



CustomerId	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	EstimatedSalary	Exited
The unique customer id	Their credit score	Which Country they belong to	Their Gender	Age	The time of bond with company	The amount left with them	The products they own.	Their estimated salary	Whether the or leave
15634602	619	France	Female	42	2	0	1	101348.88	1
15647311	608	Spain	Female	41	1	83807.86	1	112542.58	0
15619304	502	France	Female	42	8	159660.8	3	113931.57	1
15701354	699	France	Female	39	1	0	2	93826.63	0
15737888	850	Spain	Female	43	2	125510.82	1	79084.1	0
15574012	645	Spain	Male	44	8	113755.78	2	149756.71	1
15592531	822	France	Male	50	7	0	2	10062.8	0
15656148	376	Germany	Female	29	4	115046.74	4	119346.88	1
15792365	501	France	Male	44	4	142051.07	2	74940.5	0
15592389	684	France	Male	27	2	134603.88	1	71725.73	0
15767821	528	France	Male	31	6	102016.72	2	80181.12	0
15737173	497	Spain	Male	24	3	0	2	76390.01	0

A graph model of churn prediction

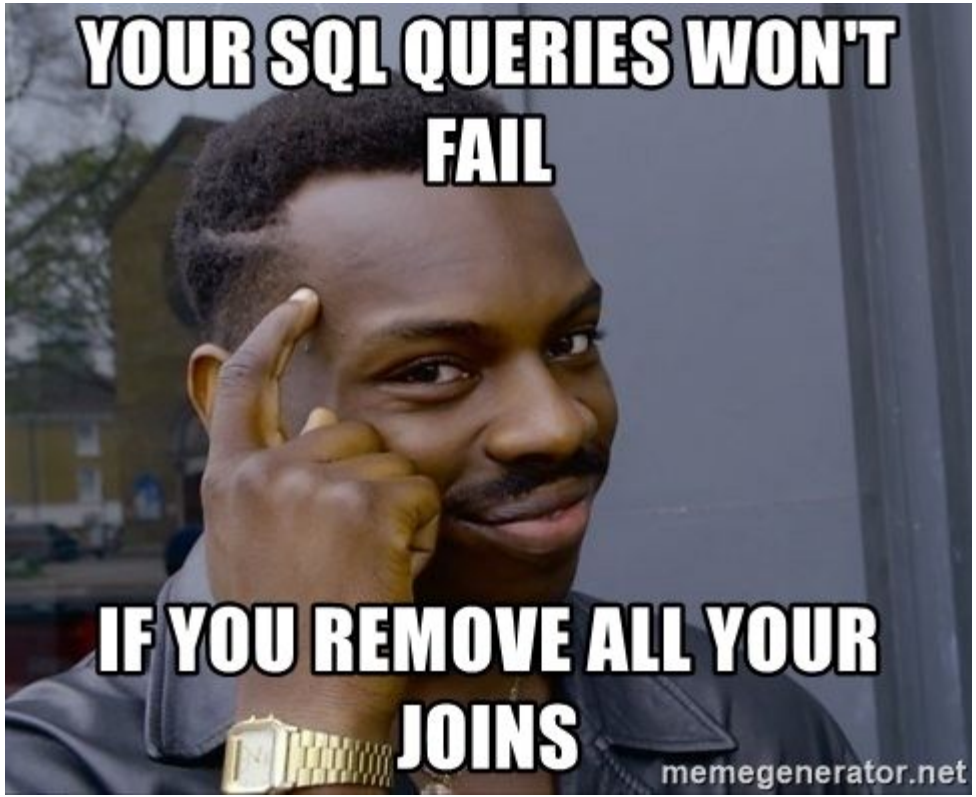
A recommendation engine problem

How do I know if I have a “graph-y” problem?

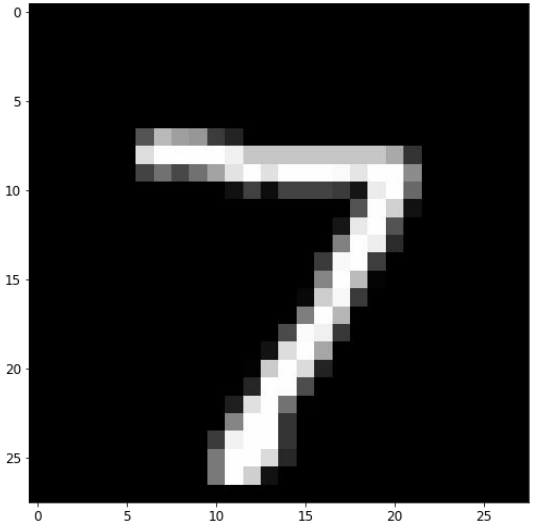
How do you know if you have a graph-y problem?

Rule of thumb:

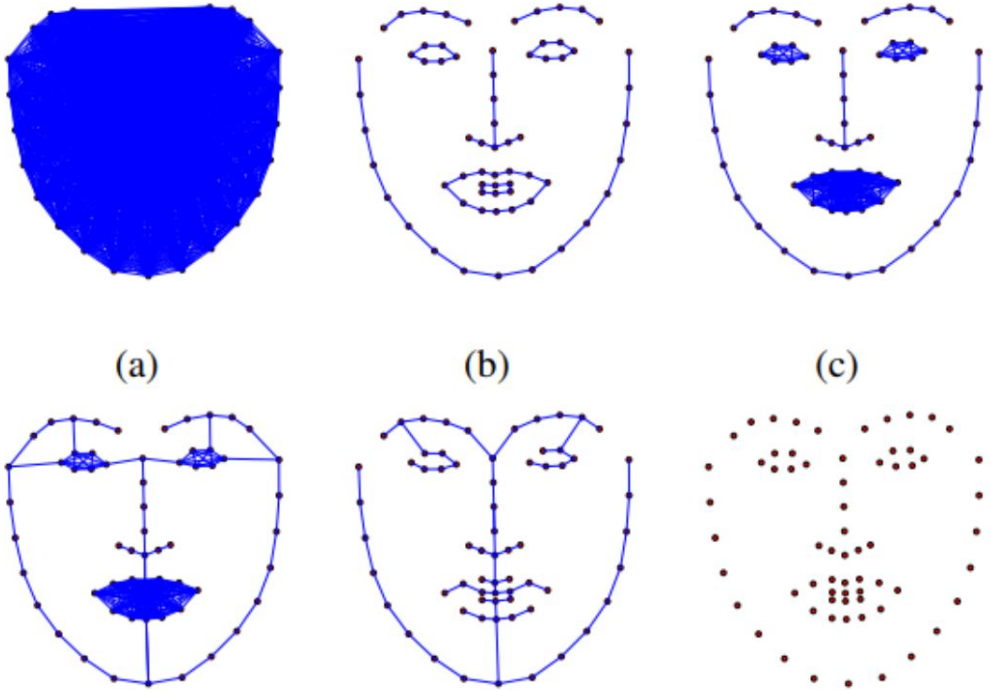
If you have to do more than a couple SQL JOINs then suspect you have a graph-y problem



Will it graph: MNIST

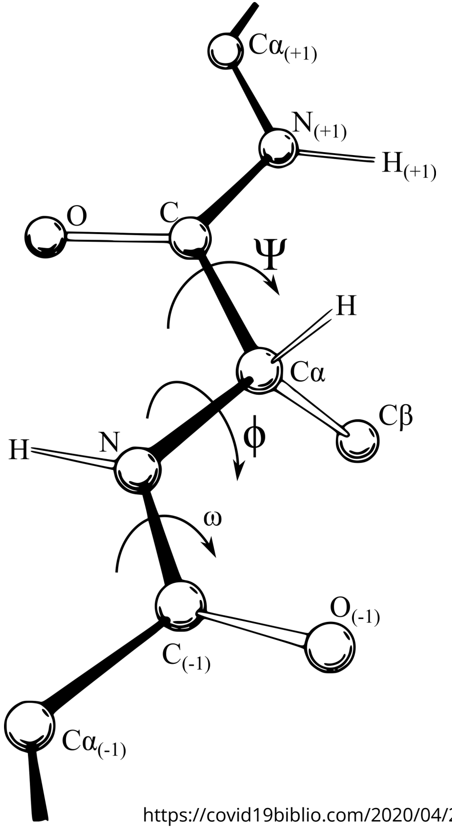


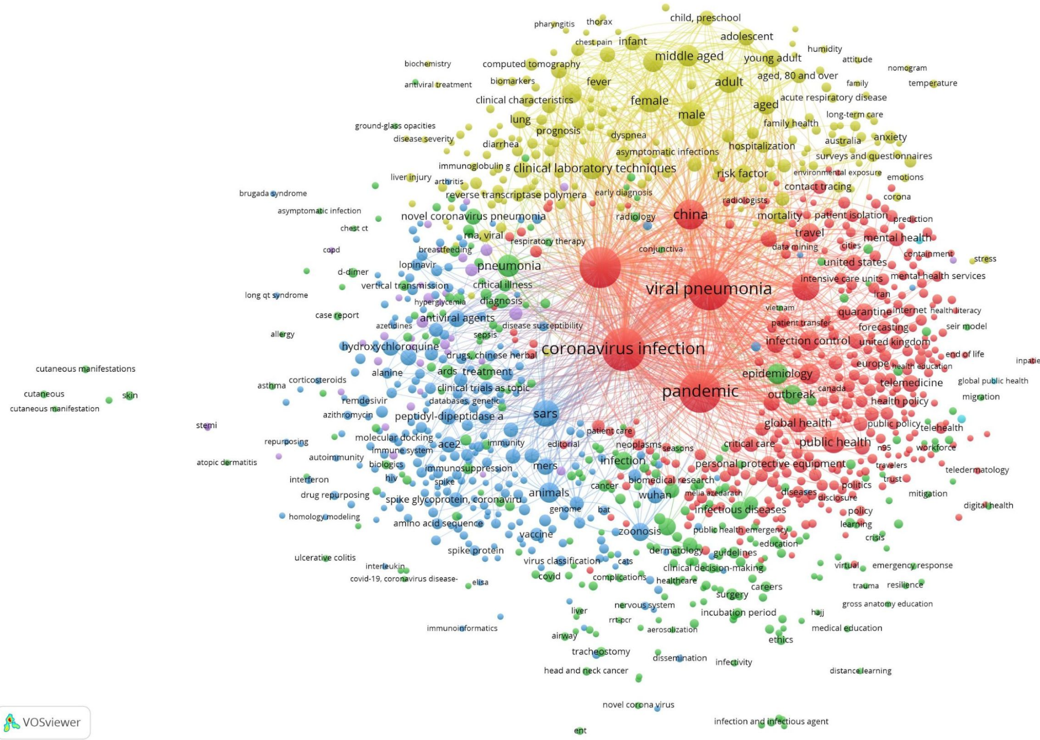
Will it graph: facial recognition



https://medium.com/@BorisAKnyazev/tutorial-on-graph-neural-networks-for-computer-vision-and-beyond-part-1-3d9fada3b80d

Will it graph: drug discovery

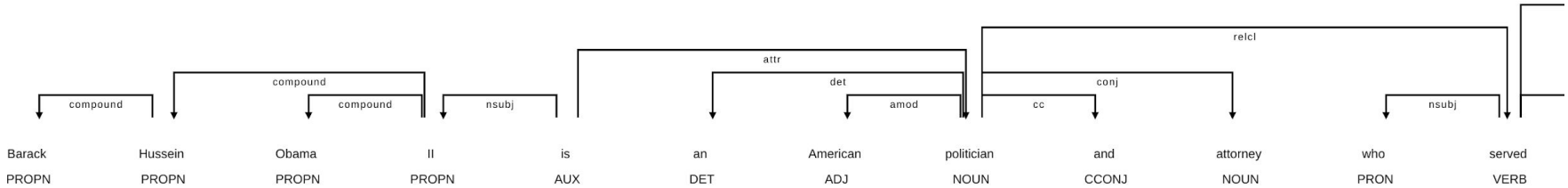




https://covid19biblio.com/2020/04/28/keyword-co-occurrence-network-graph-for-the-overall-research-field-on-covid-19-up-to-april-27th-2020/

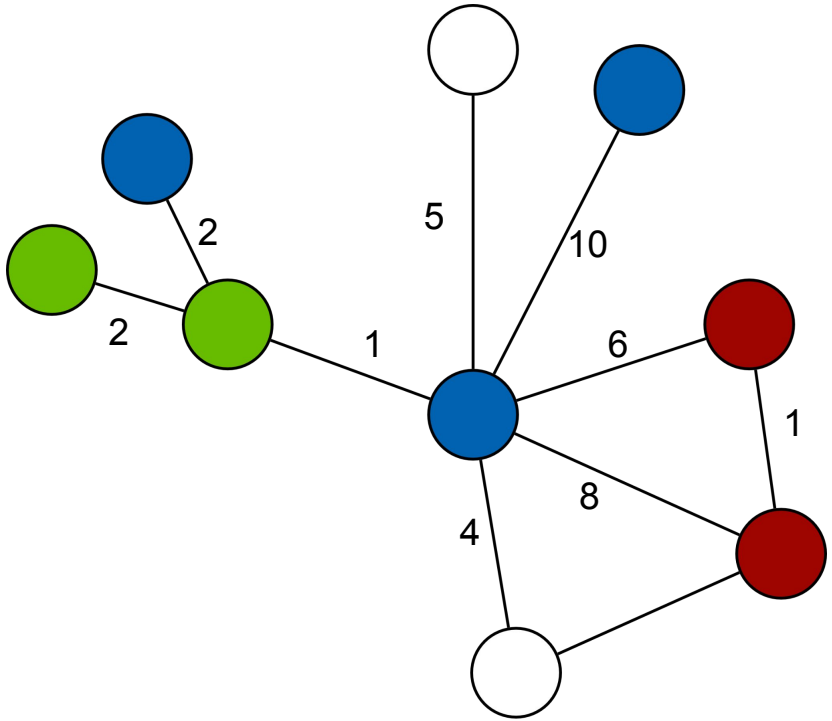
neo4j

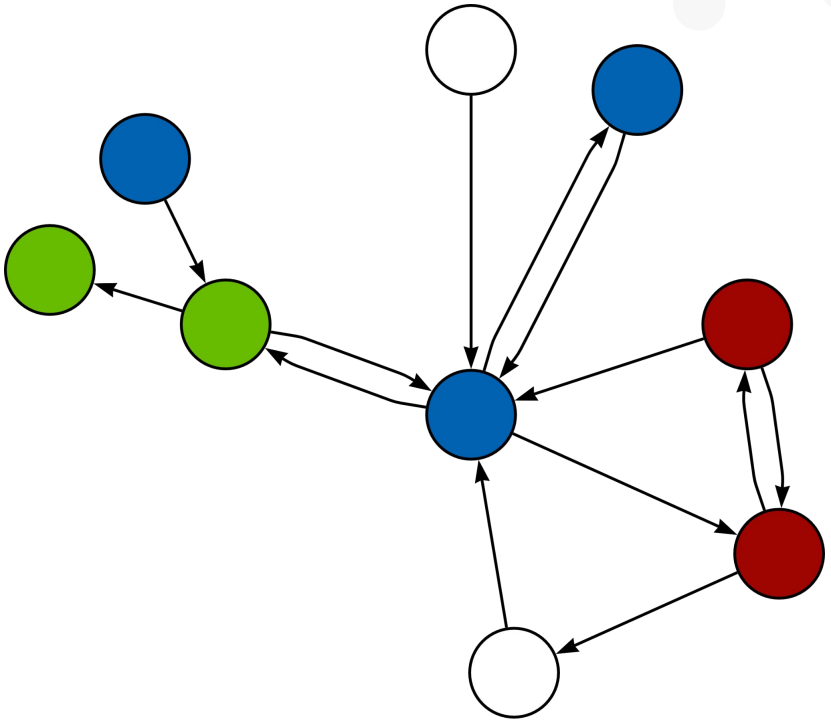
Will it graph: natural language processing



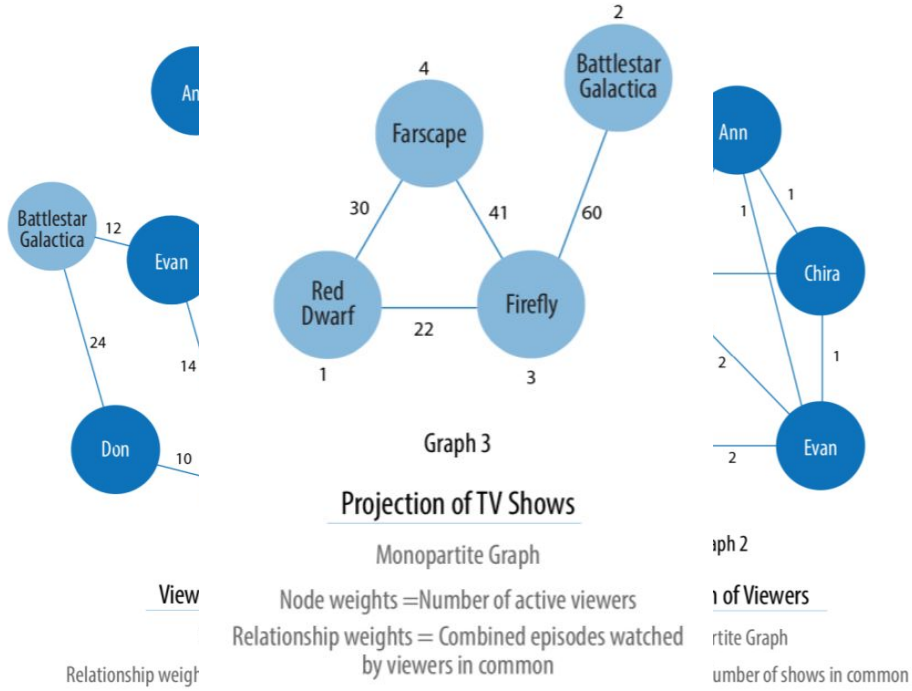
Some basic graph theory

Directed vs. Undirected vs. Weighted



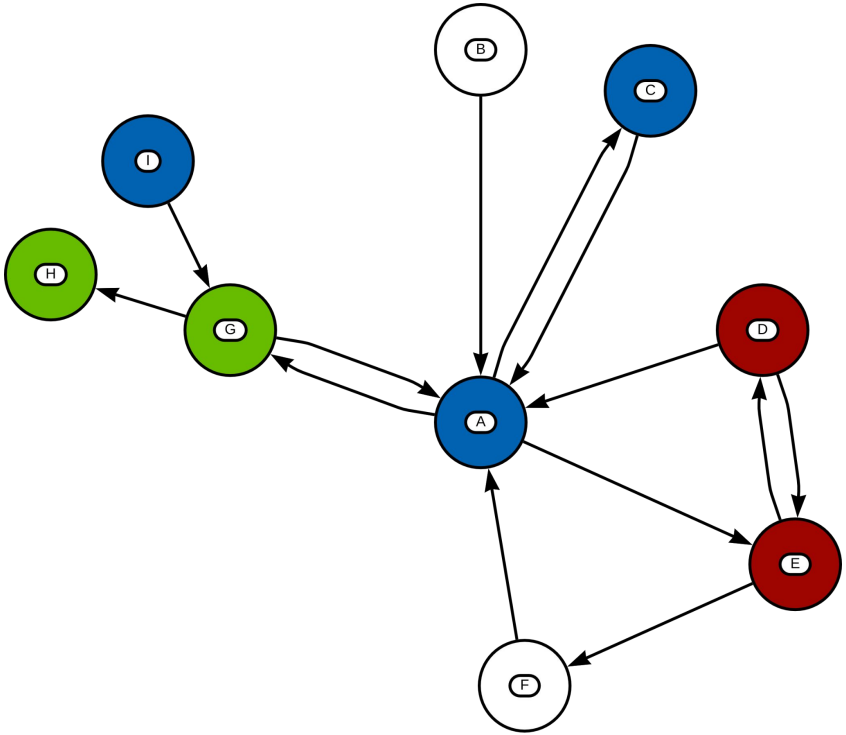


Monopartite vs. Bipartite



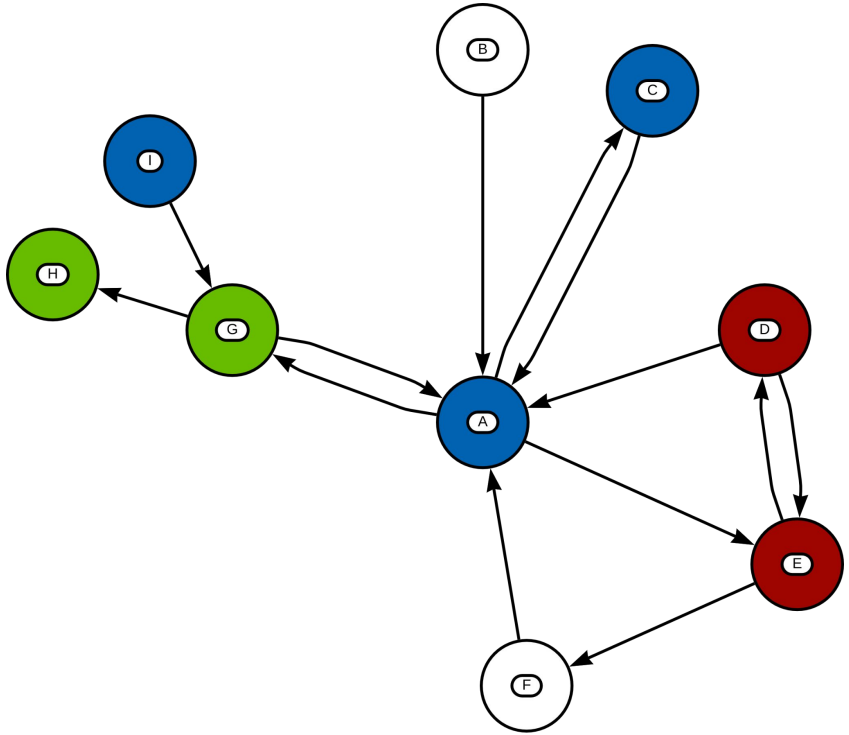
Graph Algorithms: Practical Examples in Apache Spark and Neo4j, M. Needham and A.E. Hodler (2019)

Adjacency matrix



	A	B	C	D	E	F	G	H	I
A	0	0	1	0	1	0	1	0	0
B	1	0	0	0	0	0	0	0	0
C	1	0	0	0	0	0	0	0	0
D	1	0	0	0	1	0	0	0	0
E	0	0	0	1	0	1	0	0	0
F	1	0	0	0	0	0	0	0	0
G	1	0	0	0	0	0	0	1	0
H	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	1	0	0

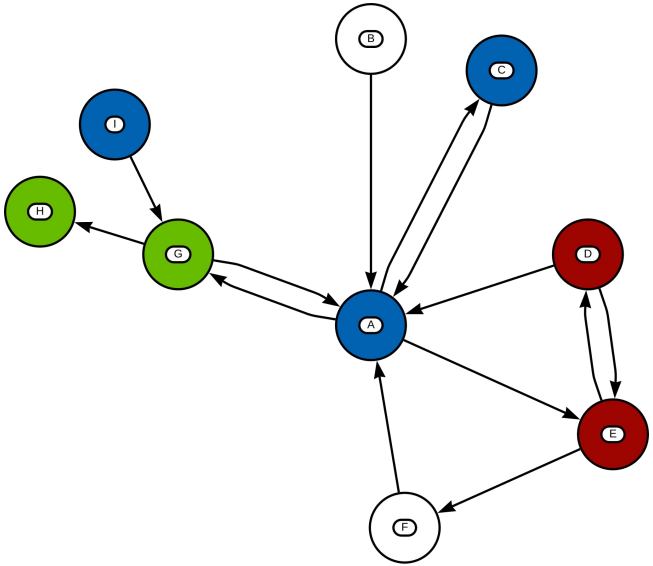
Degree

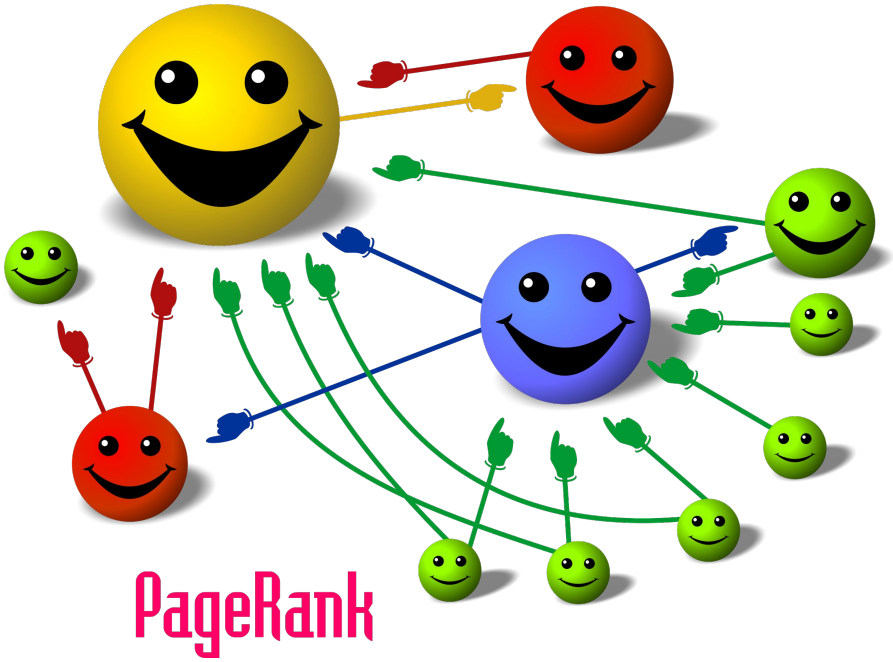


	A	B	C	D	E	F	G	H	I
A	8	0	0	0	0	0	0	0	0
B	0	1	0	0	0	0	0	0	0
C	0	0	2	0	0	0	0	0	0
D	0	0	0	3	0	0	0	0	0
E	0	0	0	0	4	0	0	0	0
F	0	0	0	0	0	2	0	0	0
G	0	0	0	0	0	0	4	0	0
H	0	0	0	0	0	0	0	1	0
I	0	0	0	0	0	0	0	0	1

Node importance (centrality algorithms)

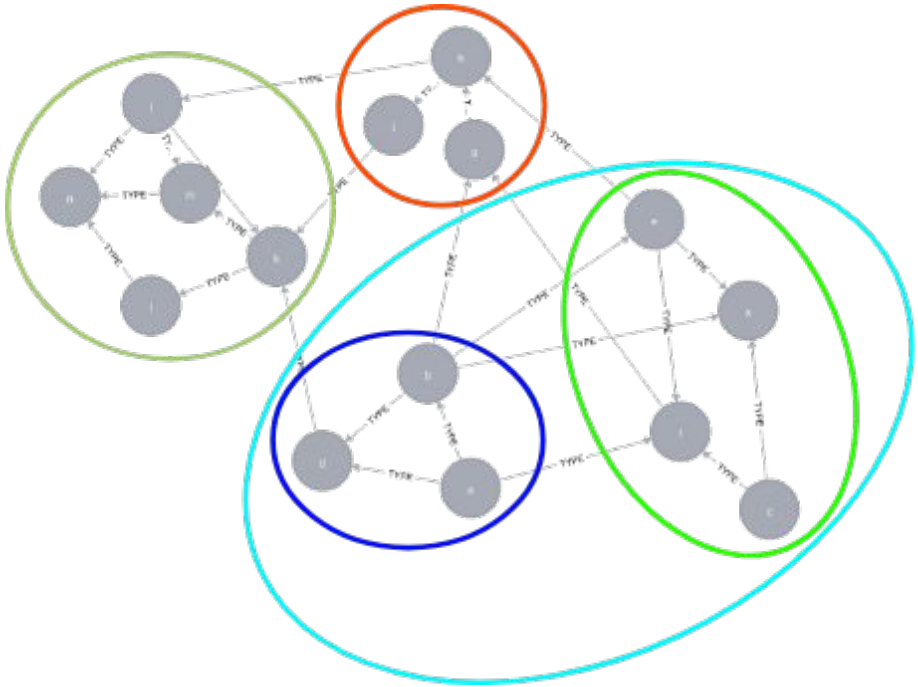
- Degree Centrality
- Betweenness centrality
- PageRank and friends





Community detection

- Connected components (union find)
- Label propagation
- Speaker listener label propagation
- Louvain modularity



How to assemble a graph

Typically you will need...

- Node list(s)
  - Unique identifiers help
  - Node properties, if available
- Edge list(s)
  - Edge names help
  - Weights, if available
- A graph model
- Answer the in-memory versus database question
  - Most of what I am going to show today can be done in either NetworkX or Neo4j
  - Choice depends on infrastructure and scalability needs

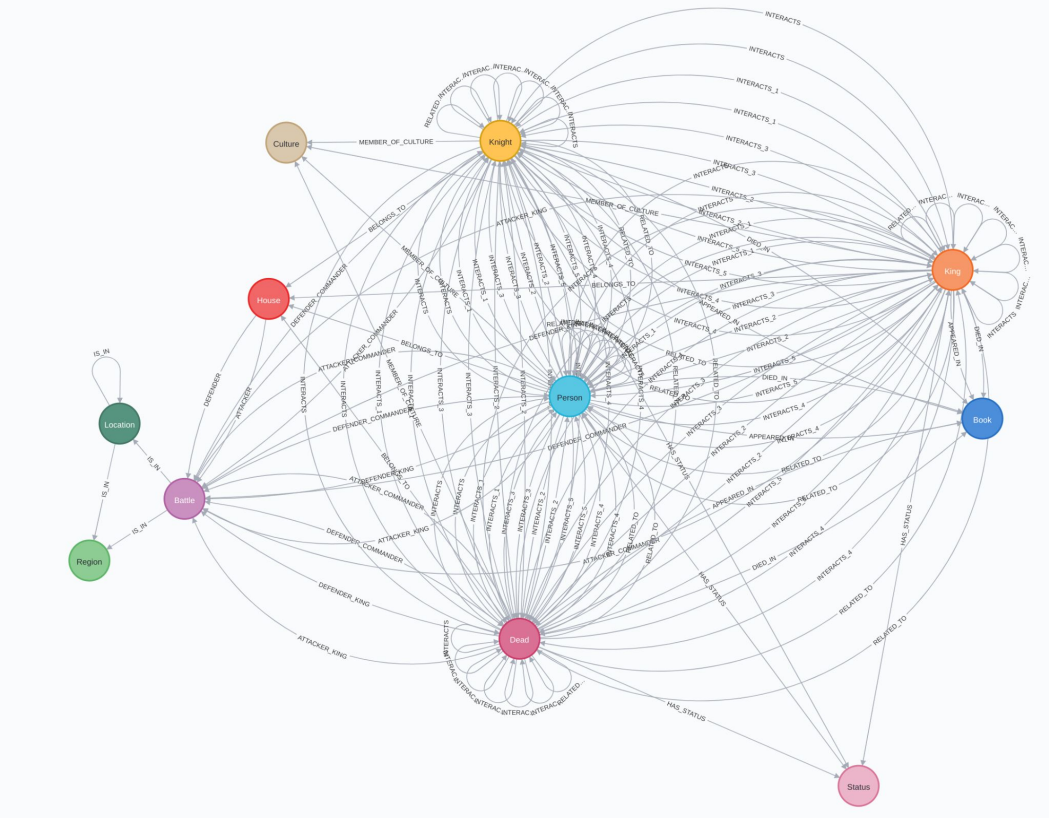
Node and edge lists

S.No	actual	pred	alive	plod	name	title	male	culture
1	0	0	0.054	0.946	Viserys II Targaryen		1	
2	1	0	0.387	0.613	Walder Frey	Lord of the Crossing	1	Rivermen
3	1	0	0.493	0.507	Addison Hill	Ser	1	
4	0	0	0.076	0.924	Aemma Arryn	Queen	0	
5	1	1	0.617	0.383	Sylva Santagar	Greenstone	0	Dornish
6	1	0	0.021	0.979	Tommen Baratheon		1	
7	0	0	0.014	0.986	Valarr Targaryen	Hand of the King	1	Valyrian
8	0	0	0.036	0.964	Viserys I Targaryen		1	
9	0	1	0.724	0.276	Wilbert	Ser	1	
10	1	0	0.391	0.609	Wilbert Osgrey	Ser	1	

name	year	battle_number	attacker_king	defender_king	attac
Battle of the Golden Tooth	298	1	Joffrey Baratheon/Tommen Baratheon	Robb Stark	Lanni
Battle at the Mummer's Ford	298	2	Joffrey Baratheon/Tommen Baratheon	Robb Stark	Lanni
Battle of Riverrun	298	3	Joffrey Baratheon/Tommen Baratheon	Robb Stark	Lanni
Battle of the Green Fork	298	4	Robb Stark	Joffrey Baratheon/Tommen Baratheon	Stark
Battle of the Whispering Wood	298	5	Robb Stark	Joffrey Baratheon/Tommen Baratheon	Stark
Battle of the Camps	298	6	Robb Stark	Joffrey Baratheon/Tommen Baratheon	Stark
Sack of Darry	298	7	Joffrey Baratheon/Tommen Baratheon	Robb Stark	Lanni
Battle of Moat Cailin	299	8	Balon Greyjoy/Euron Greyjoy	Robb Stark	Greyj
Battle of Deepwood Motte	299	9	Balon Greyjoy/Euron Greyjoy	Robb Stark	Greyj

https://github.com/tomasonjo/neo4j-game-of-thrones

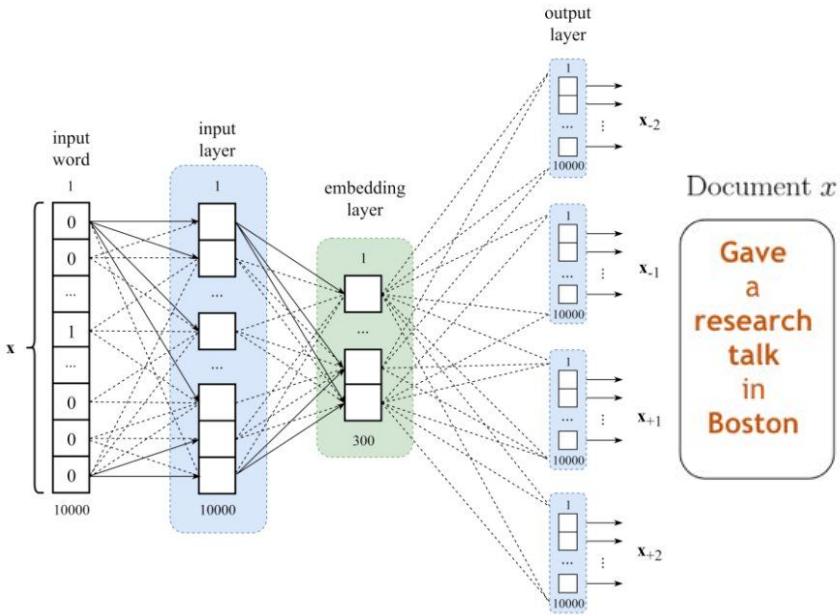
The importance of graph modeling

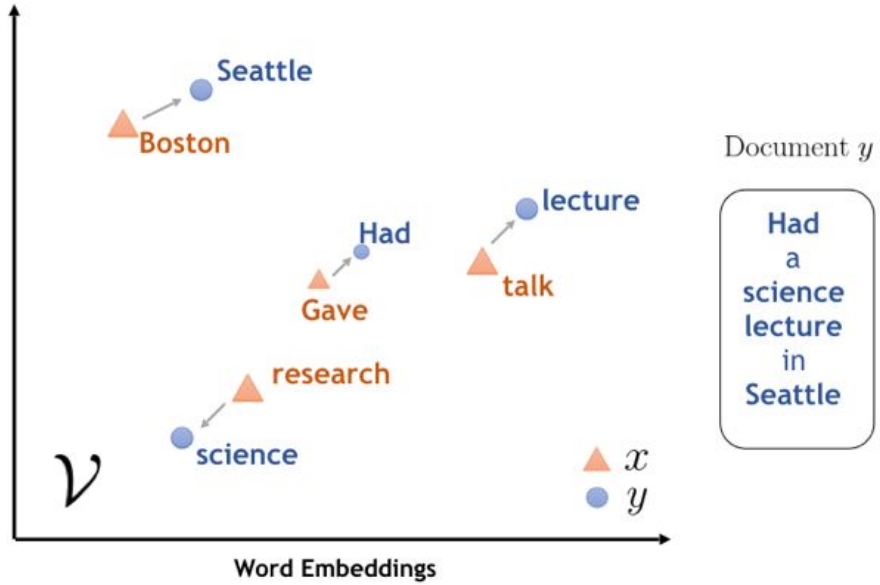


https://github.com/tomasonjo/neo4j-game-of-thrones

Graph machine learning

An example of traditional ML: word vectors (NLP)



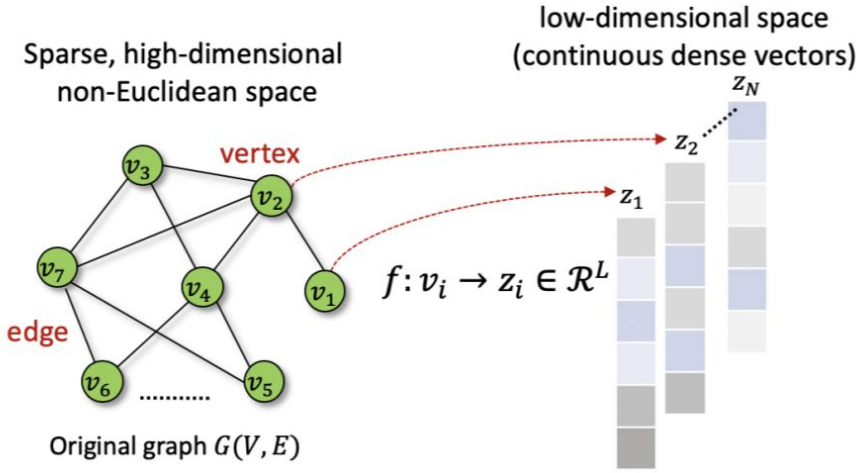


https://www.kdnuggets.com/2019/01/burkov-self-supervised-learning-word-embeddings.html

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

Graph embeddings

- Transductive
- Inductive
- Matrix factorization
- Methods based on random walks
  - FastRP
  - node2vec
- Methods based on neural networks

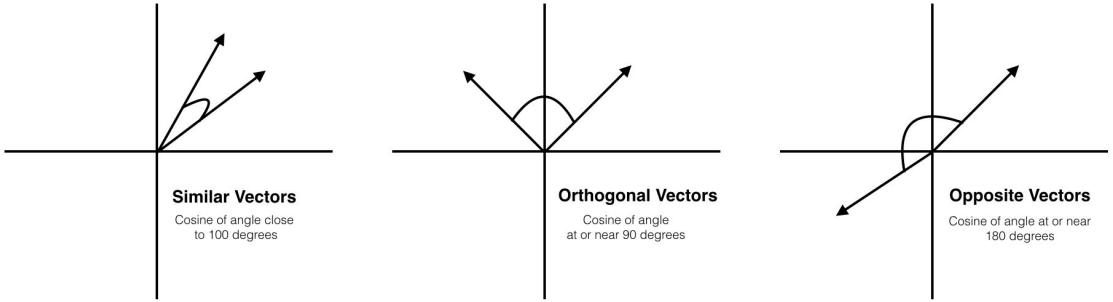


M. Xu (2020) *arXiv:2012.08019v1*

Node similarity

- Jaccard
- Cosine
- Euclidean distance
- K-Nearest Neighbors (KNN)

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

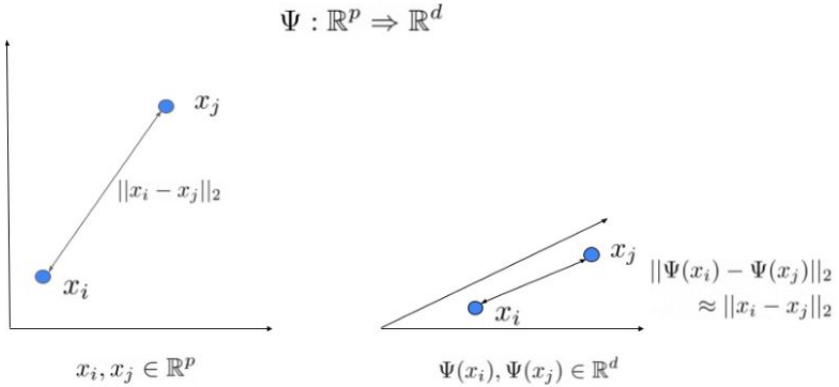


https://learning.oreilly.com/library/view/mastering-machine-learning/9781785283451/ba8bef27-953e-42a4-8180-cea152af8118.xhtml

All of the same ML models can be run using graph embeddings!

- Classification (binary, multi-class, multi-label)
- Regression
- Clustering
- Dimensionality reduction
- Similarity
- Plus more that are unique to graphs!
  - Link prediction
  - (Sub)graph-level structural similarity

FastRP



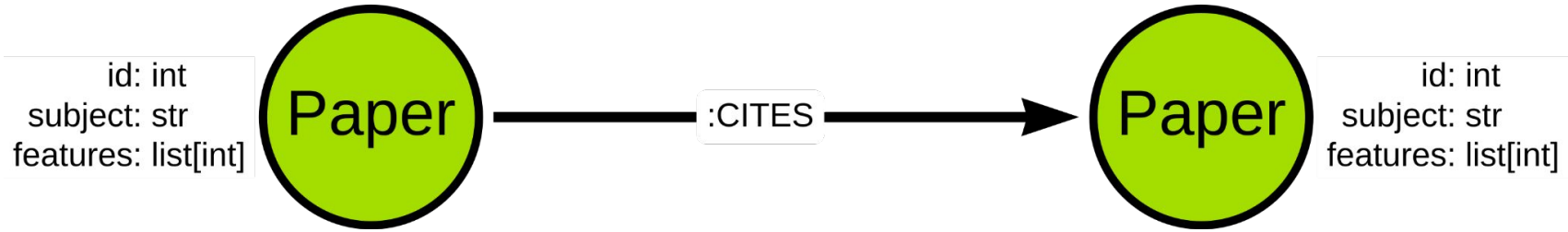
$$N = (\alpha_1 \tilde{A} + \alpha_2 \tilde{A} + \alpha_3 \tilde{A} + \alpha 4 \tilde{A}) \cdot R$$

$$\tilde{A} = \tilde{A}(D, \beta)$$

https://dev.neo4j.com/fastrp_background

CORA Dataset

- 2708 scientific publications in data science
- 7 classes
- 5429 citation relationships
- Abstracts one-hot encoded to a vocabulary of 1433 words

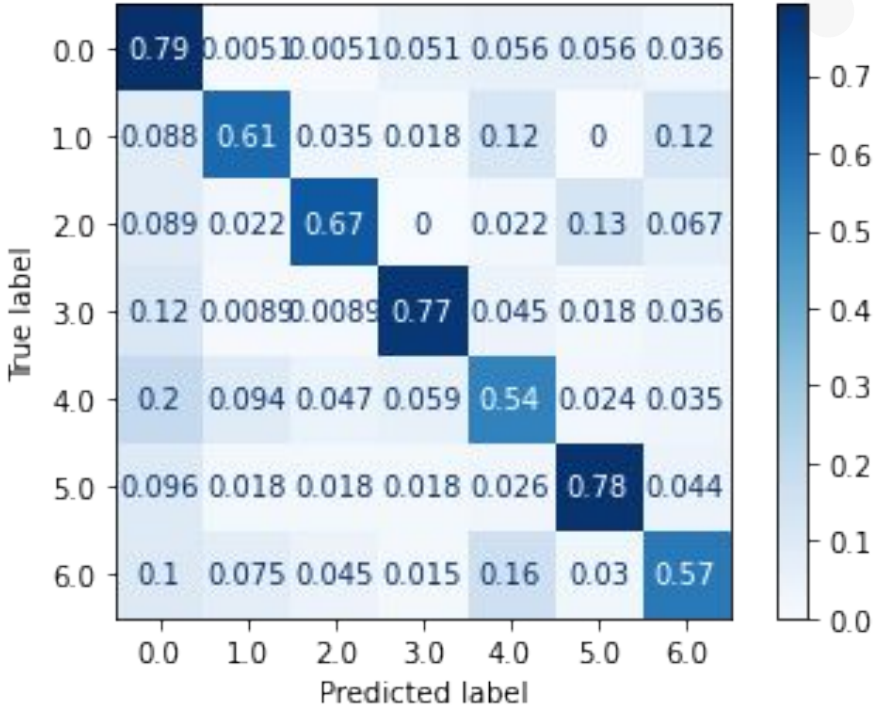


Two “models”

- Goal: compare embeddings
  - Keep ML model identical
  - Keep ML model simple
- Caveat
  - No access to the word vectors, vocabularies for tuning
- Both “models” given the same task
  - Multi-class classification
  - Imbalanced dataset

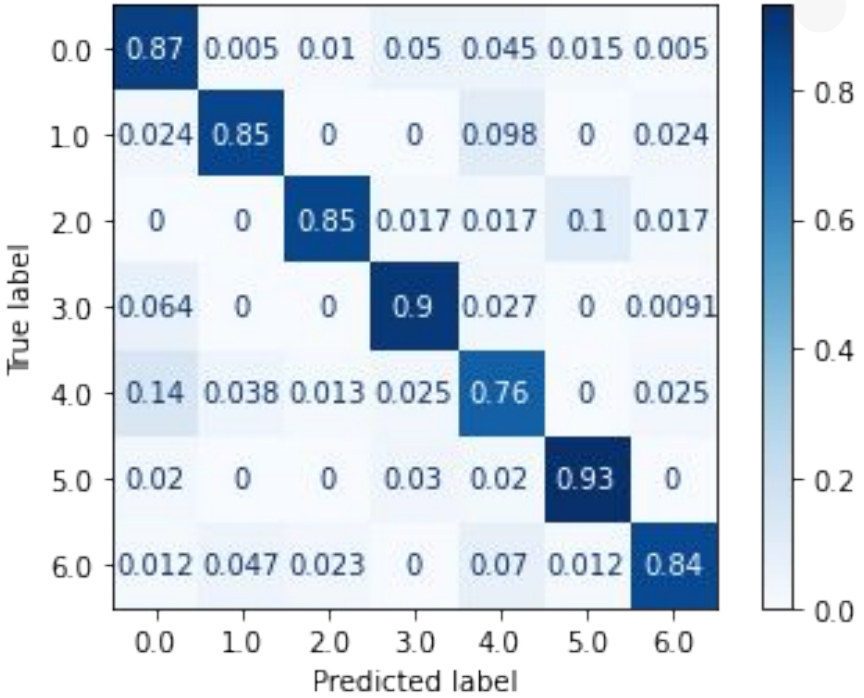
Results using word vectors

Mean accuracy: 0.726

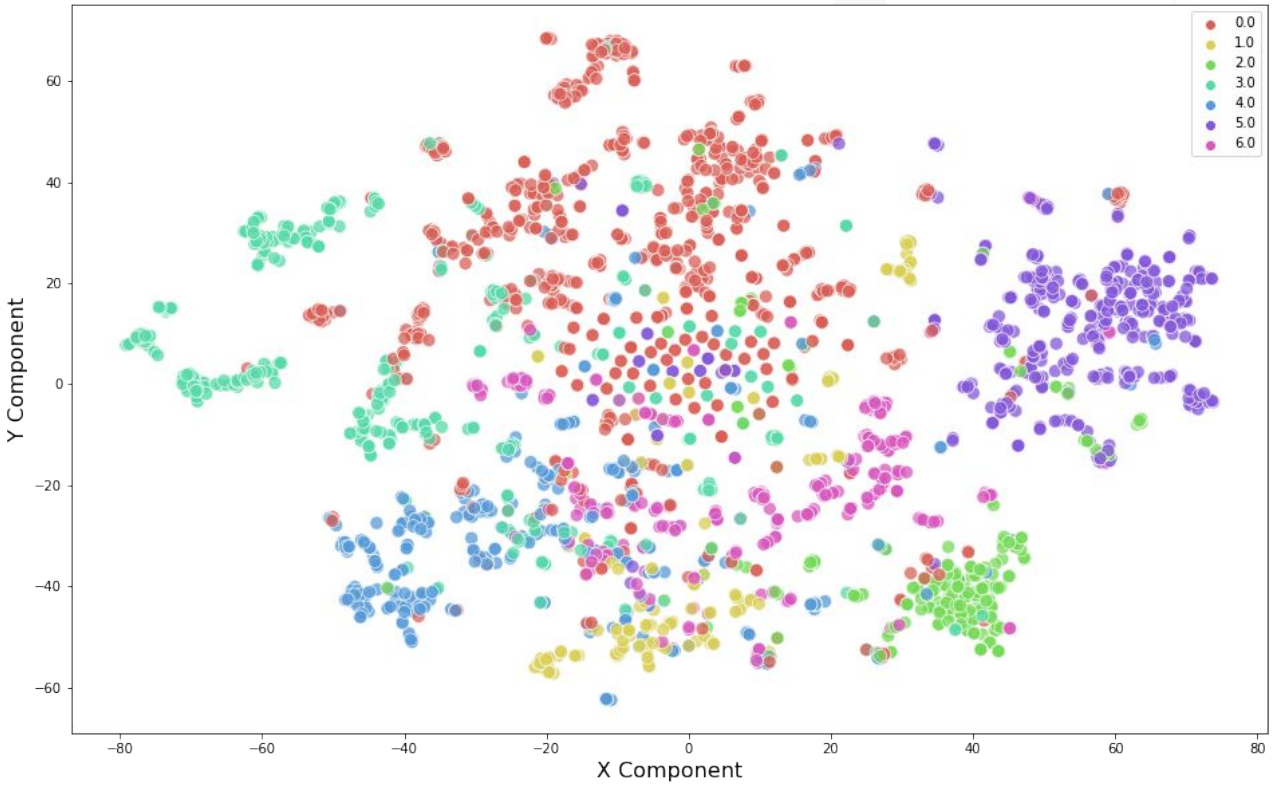


FastRP embeddings with default hyperparameters

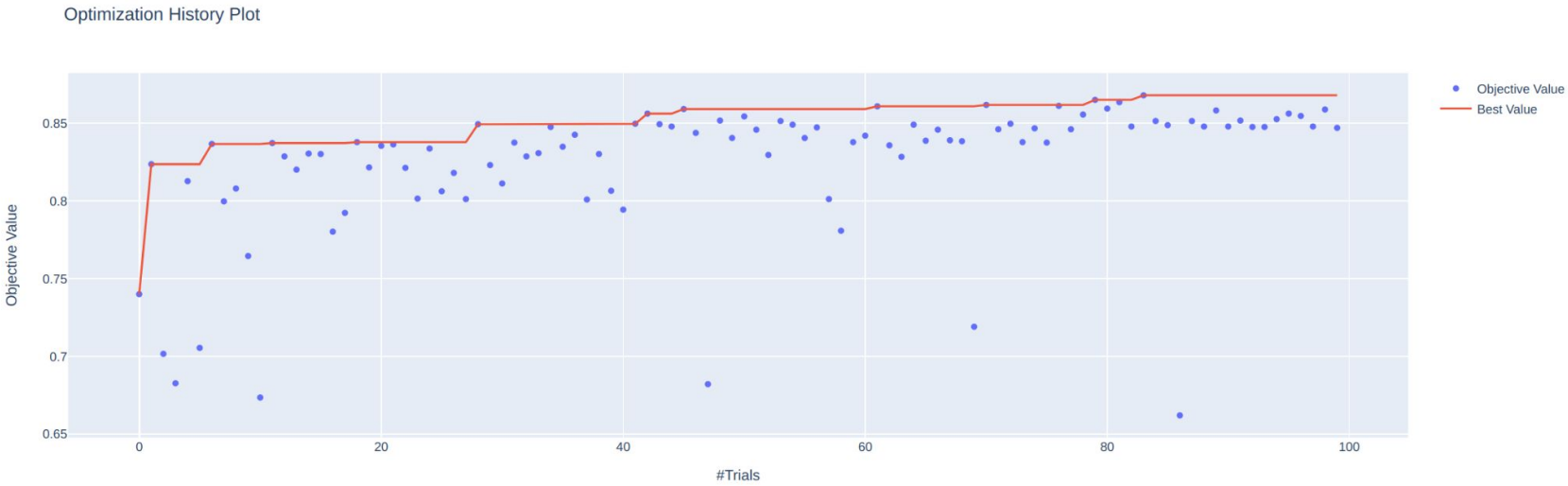
Mean accuracy: 0.850



Neural_Networks: 0.0
Rule_Learning: 1.0
Reinforcement_Learning: 2.0
Probabilistic_Methods: 3.0
Theory: 4.0
Genetic_Algorithms: 5.0
Case_Based: 6.0

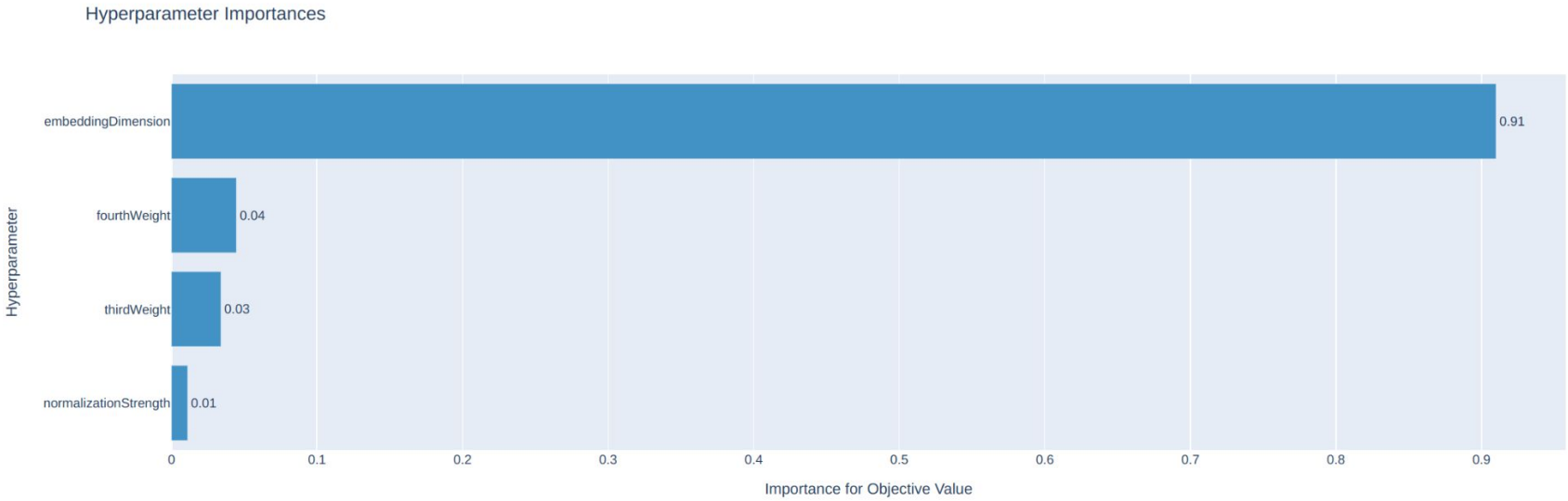


Results of tuning hyperparameters



Mean accuracy: 0.868

Results of tuning hyperparameters



Two Key Concepts

1. It can be good to break the assumption that individual data points are independent

2. Modeling relationships can result in models that are less noisy, more accurate

https://github.com/cj2001/odsc_west_2021

Thank you!

@CJLovesData1