CSE Capstone Project

Entity Linking (aka Entity Disambiguation) using Kensho-Derived Wikimedia Dataset

Weiru Chen, Dean Hathout, Tyler Yoo, David Zheng

IACS INSTITUTE FOR APPLIED COMPUTATIONAL SCIENCE AT HARVARD UNIVERSITY

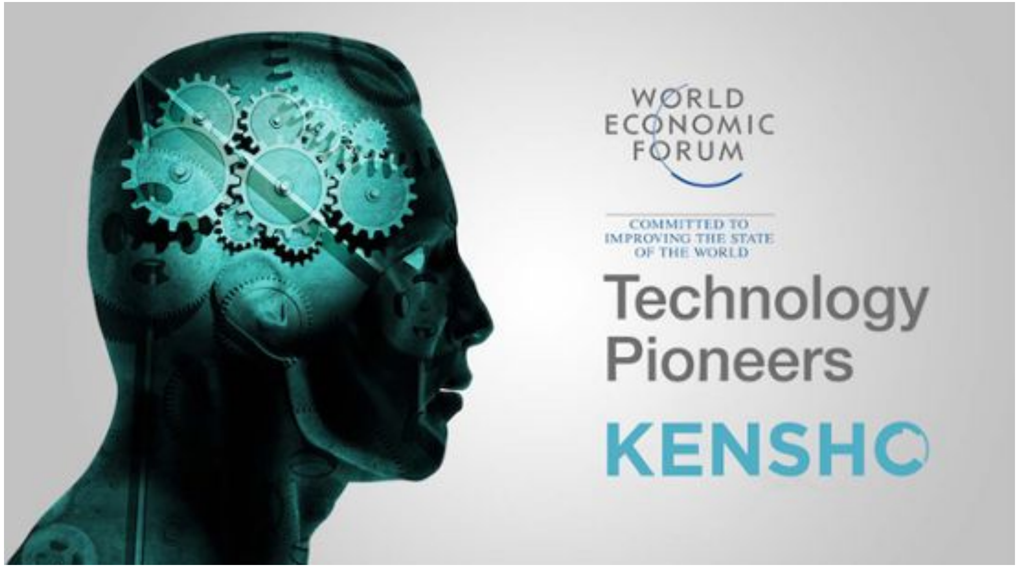
KENSHO

Final Presentation

Background & Motivation

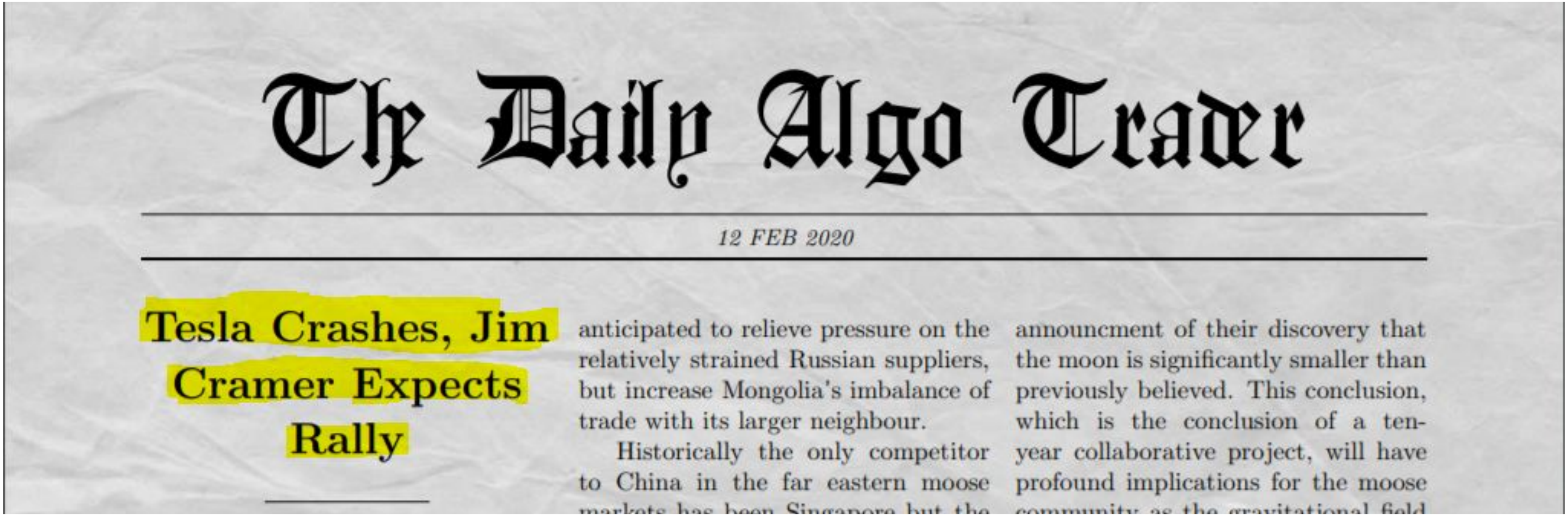
Partner Background

S&P Global



Kensho Technologies is a leading technology company focused on ML and AI solutions

“Tesla” - Car or Company?



The Daily Algo Trader

12 FEB 2020

Tesla Crashes, Jim Cramer Expects Rally

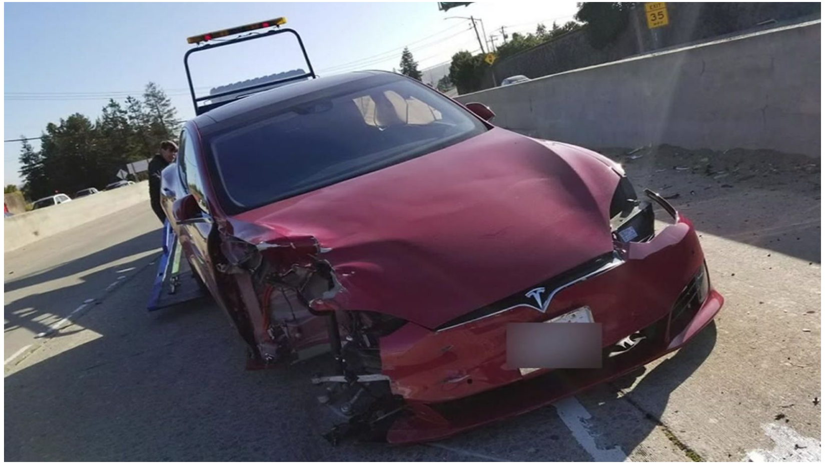
anticipated to relieve pressure on the relatively strained Russian suppliers, but increase Mongolia's imbalance of trade with its larger neighbour.

Historically the only competitor to China in the far eastern moose markets has been Singapore but the

announcment of their discovery that the moon is significantly smaller than previously believed. This conclusion, which is the conclusion of a ten-year collaborative project, will have profound implications for the moose community as the gravitational field

Motivation: Does the word “Tesla” here refer to the car, the company, or even the person?

“Tesla” - Car or Company?



VS.

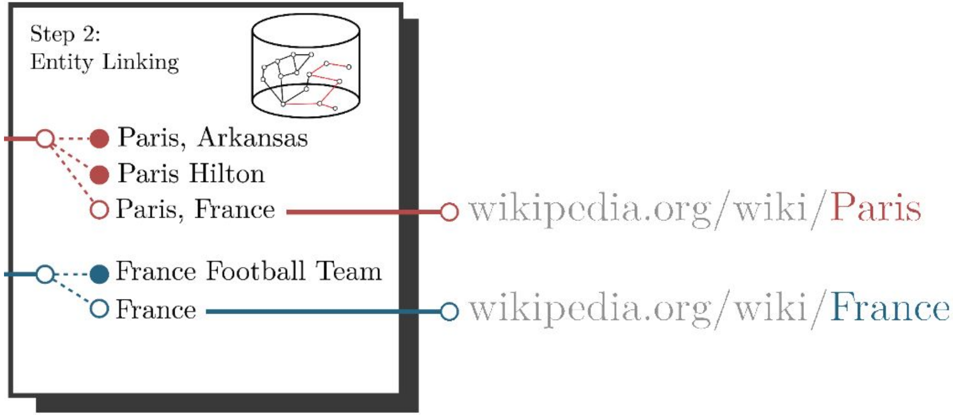
?
We Said Sell
We Said Buy
1890.9
1746.0
1601.2
1456.4
1311.6
1166.8
1022.0
877.23
732.42
587.61
442.80
06-1996
06-1998
05-2010
05-2012
05-2014

Motivation: Determining the exact reference has important implications

Background - Entity Linking Definition

In natural language processing, entity linking, is the task of assigning a unique identity to entities (such as famous individuals, locations, or companies) mentioned in text.

For example, given the sentence "Paris is the capital of France", the idea is to determine that "Paris" refers to the city of Paris and not to Paris Hilton or any other entity that could be referred to as "Paris".



We link entities to our knowledge base by using context to disambiguate

Background - Entity Linking Applications

In countless applications (text analysis, recommender systems, search engines, etc), we must be able to separate relevant concepts in text from non-meaningful data.

E.g. when typing “the capital of France” in a search engine, entity is necessary to actually retrieve containing the word “Paris”

Entity linking has been used to improve the performance of information retrieval systems and to improve search performance on digital libraries.

Entity linking is also a key input for semantic search, which seeks to improve search accuracy by understanding the searcher's intent and the contextual meaning of terms as they appear in the searchable dataspace.

Why is entity linking important?

Problem Statement & Statement of Work

Problem Statement

Our goal is to generate a model that can accurately and efficiently execute named entity linking.

We wish to try a novel approach that uses the predictions of the other entities' types in a context window to refine our predictions.

"Paris is the capital of France"

Can we improve entity linking accuracies by considering model predictions for other word in context window?

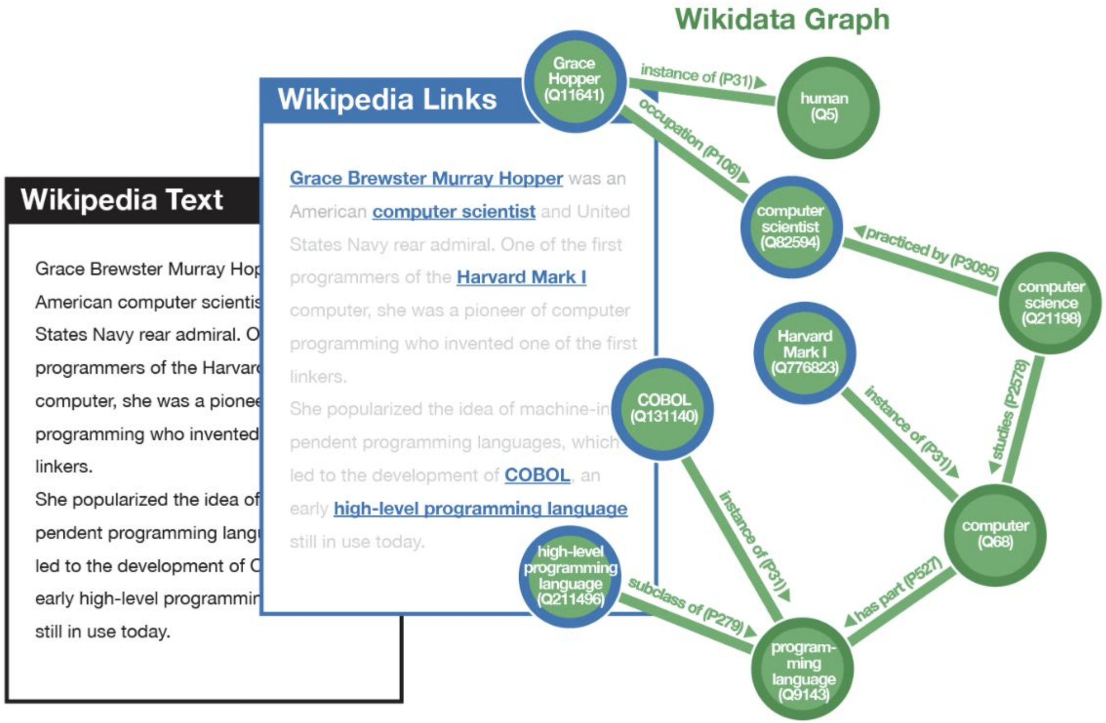
Statement of Work

What are we focusing on?	What are we not focusing on?	Nice-to-haves/Future Work
Named Entity Disambiguation via Context Window Type Predictions	Named Entity Recognition	State-of-the-art baseline (e.g. OpenAI)
KDWD	Text-based entity linking	Speed improvements

We will focus on building a robust graph-based entity linking model using the Kensho knowledge graph

KDWD EDA

Kensho-Derived Wikimedia Dataset (KDWD)



Wikipedia Text

Grace Brewster Murray Hop
American computer scientis
States Navy rear admiral. O
programmers of the Harvar
computer, she was a pionee
programming who invented
linkers.
She popularized the idea of
pendent programming lang
led to the development of C
early high-level programmin
still in use today.

Wikipedia Links

Grace Brewster Murray Hopper was an American **computer scientist** and United States Navy rear admiral. One of the first programmers of the **Harvard Mark I** computer, she was a pioneer of computer programming who invented one of the first linkers.

She popularized the idea of machine-independent programming languages, which led to the development of **COBOL**, an early **high-level programming language** still in use today.

Kensho-Derived Wikimedia Dataset (KDWD)

2 Core Problems

1. **How do we create the training/test dataset?**
    a. Where do we get the data?
    b. How do we label it?

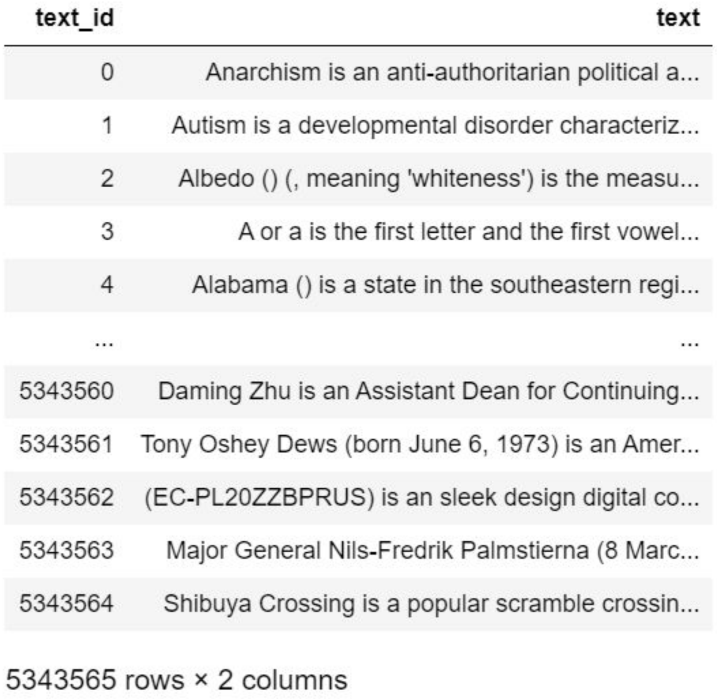
2. **Parsing 58 million items is slow…**
    a. How do we speed it up?
    b. What is the ideal data structure for our purposes?

(KDWD) Create Dataset - Text Extraction

JSON File

{'page_id': 12,
'sections': [{'name': 'Introduction',
'text': "Anarchism is an anti-authoritarian political and social philosophy that rejects hierarchies deemed unjust and Strains of anarchism have often been divided into the categories of social and individualist anarchism, or similar dual classifications.",
'link_lengths': [18, 9, 17, 11, 12, 13, 11, 19, 17, 5, 8, 18, 9, 16, 9, 12, 11, 9, 23, 30, 28, 13, 12, 6, 23],
'link_offsets': [16, 35, 49, 80, 143, 157, 201, 269, 405, 500, 586, 602, 638, 652, 715, 726, 740, 753, 767, 886, 964, 1052, 1078, 1160, 1171],
'target_page_ids': [867979, 23040, 586276, 13998, 40949353, 191161, 89313, 4228181, 26271818, 23604120, 18247344, 23490, 1228884, 5483237, 17865, 46399294, 49938, 1799997, 76393, 1063286, 46399274, 15181, 5708, 39353100, 14936]}]}

16.8Gb



text_id	text
0	Anarchism is an anti-authoritarian political a...
1	Autism is a developmental disorder characteriz...
2	Albedo () (, meaning 'whiteness') is the measu...
3	A or a is the first letter and the first vowel...
4	Alabama () is a state in the southeastern regi...
...	...
5343560	Daming Zhu is an Assistant Dean for Continuing...
5343561	Tony Oshey Dews (born June 6, 1973) is an Amer...
5343562	(EC-PL20ZZBPRUS) is an sleek design digital co...
5343563	Major General Nils-Fredrik Palmstierna (8 Marc...
5343564	Shibuya Crossing is a popular scramble crossin...

5343565 rows × 2 columns

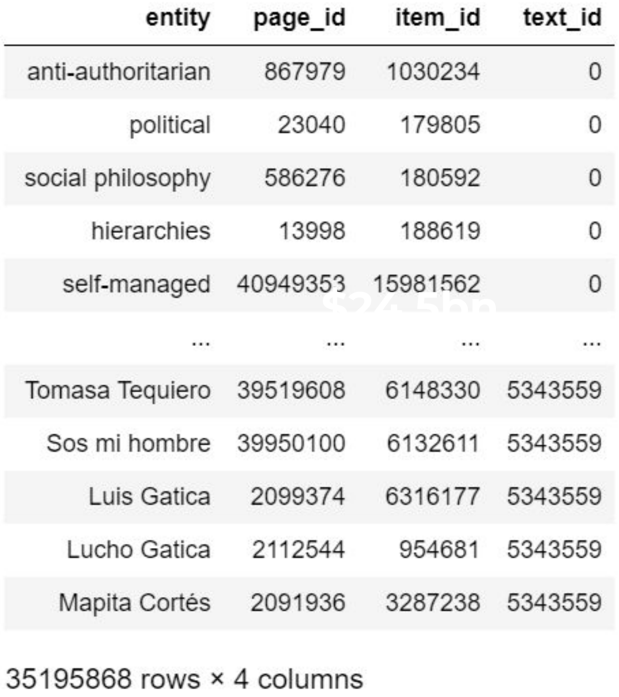
2.7Gb

(KDWD) Create Dataset - Entity Extraction

JSON File

{'page_id': 12,
'sections': [{'name': 'Introduction',
'text': "Anarchism is an anti-authoritarian political and social philosophy that rejects hierarchies deemed unjust and Strains of anarchism have often been divided into the categories of social and individualist anarchism, or similar dual classifications.",
'link_lengths': [18, 9, 17, 11, 12, 13, 11, 19, 17, 5, 8, 18, 9, 16, 9, 12, 11, 9, 23, 30, 28, 13, 12, 6, 23],
'link_offsets': [16, 35, 49, 80, 143, 157, 201, 269, 405, 500, 586, 602, 638, 652, 715, 726, 740, 753, 767, 886, 964, 1052, 1078, 1160, 1171],
'target_page_ids': [867979, 23040, 586276, 13998, 40949353, 191161, 89313, 4228181, 26271818, 23604120, 18247344, 23490, 1228884, 5483237, 17865, 46399294, 49938, 1799997, 76393, 1063286, 46399274, 15181, 5708, 39353100, 14936]}]}

16.8Gb



entity	page_id	item_id	text_id
anti-authoritarian	867979	1030234	0
political	23040	179805	0
social philosophy	586276	180592	0
hierarchies	13998	188619	0
self-managed	40949353	15981562	0
...	...	...	...
Tomasa Tequiero	39519608	6148330	5343559
Sos mi hombre	39950100	6132611	5343559
Luis Gatica	2099374	6316177	5343559
Lucho Gatica	2112544	954681	5343559
Mapita Cortés	2091936	3287238	5343559

35195868 rows × 4 columns

1.22Gb

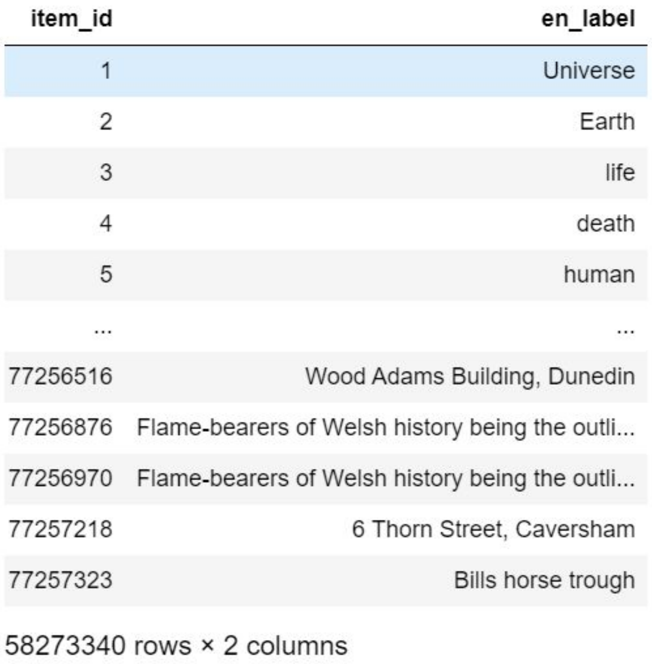
Kensho-Derived Wikimedia Dataset (KDWD)

2 Core Problems

1. **How do we create the training/test dataset?**
   a. Where do we get the data?
   b. How do we label it?

2. **Parsing 58 million items is slow…**
   a. How do we speed it up?
   b. What is the ideal data structure for our purposes?

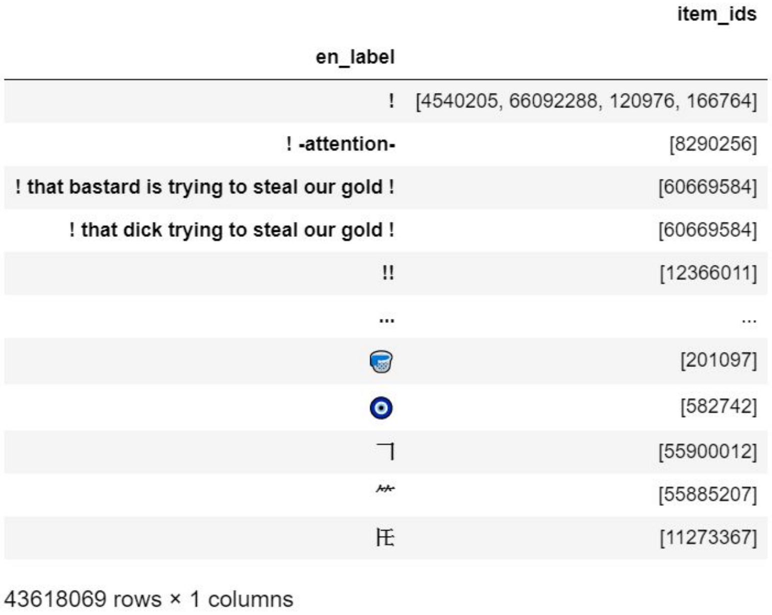
(KDWD) Item Database



item_id	en_label
1	Universe
2	Earth
3	life
4	death
5	human
...	...
77256516	Wood Adams Building, Dunedin
77256876	Flame-bearers of Welsh history being the outli...
77256970	Flame-bearers of Welsh history being the outli...
77257218	6 Thorn Street, Caversham
77257323	Bills horse trough

58273340 rows × 2 columns

4.04Gb

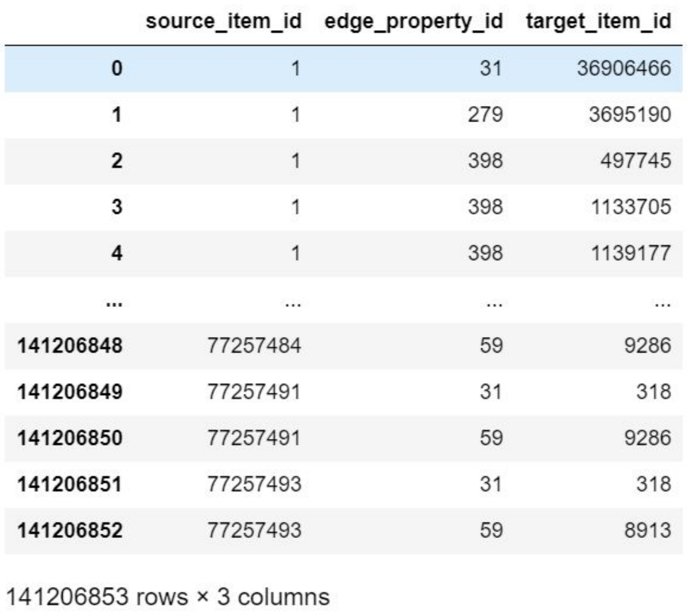


	item_ids
en_label	
!	[4540205, 66092288, 120976, 166764]
! -attention-	[8290256]
! that bastard is trying to steal our gold !	[60669584]
! that dick trying to steal our gold !	[60669584]
!!	[12366011]
...	...
🪣	[201097]
🧿	[582742]
㇕	[55900012]
⺮	[55885207]
厇	[11273367]

43618069 rows × 1 columns

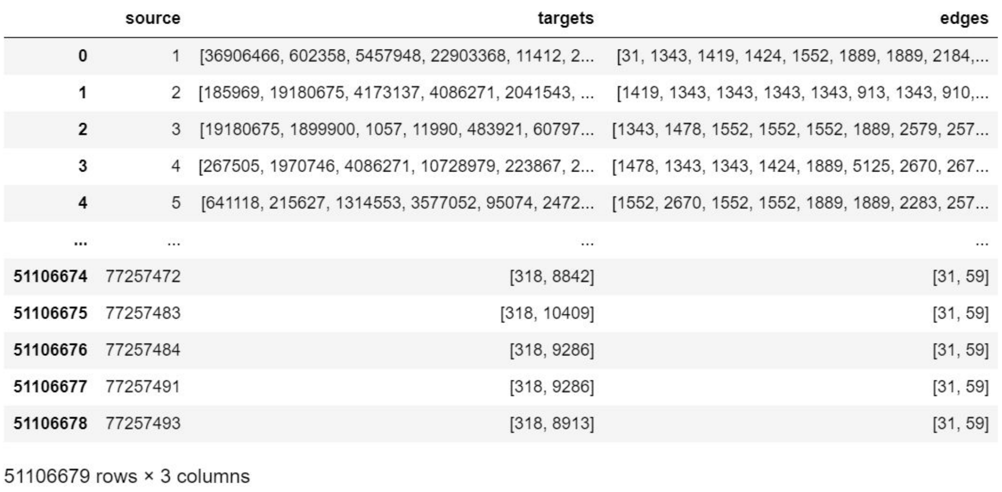
2.6Gb

(KDWD) Statement Graph



	source_item_id	edge_property_id	target_item_id
0	1	31	36906466
1	1	279	3695190
2	1	398	497745
3	1	398	1133705
4	1	398	1139177
...	...	...	...
141206848	77257484	59	9286
141206849	77257491	31	318
141206850	77257491	59	9286
141206851	77257493	31	318
141206852	77257493	59	8913

141206853 rows × 3 columns



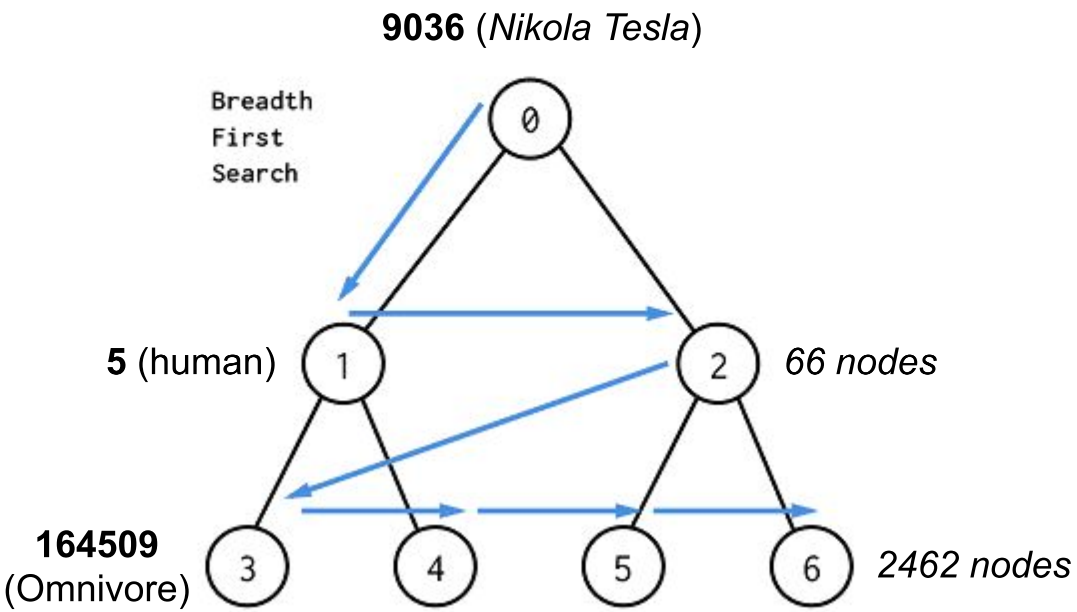
	source	targets	edges
0	1	[36906466, 602358, 5457948, 22903368, 11412, 2...	[31, 1343, 1419, 1424, 1552, 1889, 1889, 2184,...
1	2	[185969, 19180675, 4173137, 4086271, 2041543, ...	[1419, 1343, 1343, 1343, 1343, 913, 1343, 910,...
2	3	[19180675, 1899900, 1057, 11990, 483921, 60797...	[1343, 1478, 1552, 1552, 1552, 1889, 2579, 257...
3	4	[267505, 1970746, 4086271, 10728979, 223867, 2...	[1478, 1343, 1343, 1424, 1889, 5125, 2670, 267...
4	5	[641118, 215627, 1314553, 3577052, 95074, 2472...	[1552, 2670, 1552, 1552, 1889, 1889, 2283, 257...
...	...	...	...
51106674	77257472	[318, 8842]	[31, 59]
51106675	77257483	[318, 10409]	[31, 59]
51106676	77257484	[318, 9286]	[31, 59]
51106677	77257491	[318, 9286]	[31, 59]
51106678	77257493	[318, 8913]	[31, 59]

51106679 rows × 3 columns

2.45Gb

1.91Gb

(KDWD) Graph Traversal Performance



9036 (*Nikola Tesla*)

5 (human)

66 nodes

164509 (Omnivore)

2462 nodes

Level	# Nodes	Time
1	66	997 us
2	2462	3 ms
3	37805	21 ms
4	166908	183 ms
5	326619	1.03 s
6	371614	3.04 s
7	305859	5.97 s
8	211902	7.59 s
9	156552	6.46 s
10	99839	7.54 s

Statements are unidirectional (source -> target): 9036 Nikola Tesla -> (5) human -> 42 items
Graph breaks if reversed (target -> source): 9036 Nikola Tesla -> (5) human -> 6,701,310 items

Current Project Ideas

Baseline Modeling

- Algorithm: Given a mention w_i, its associated entity candidates $e_{i1}, e_{i2}, ..., e_{ik}$, and their corresponding popularity (page views) $p_{i1}, p_{i2}, ..., p_{ik}$, select e_{im} as the predicted entity for the mention w_i, where $m = argmax_j p_{ij}$.

- Metric: Randomly sampled k = 20,000 mentions and labeled each mention with the entity based on the above algorithm, then computed the overall accuracy rate.

- Accuracy rate of baseline model: 61.83%

Baseline Modeling

- Algorithm: Given a mention w_i, its associated entity candidates $e_{i1}, e_{i2}, \ldots, e_{ik}$, and their corresponding popularity (target anchor count) $p_{i1}, p_{i2}, \ldots, p_{ik}$, select e_{im} as the predicted entity for the mention w_i, where $m = argmax_j p_{ij}$.

- Metric: Randomly sampled k = 20,000 mentions and labeled each mention with the entity based on the above algorithm, then computed the overall accuracy rate.

- Accuracy rate of baseline model: 59.135%

Baseline Modeling

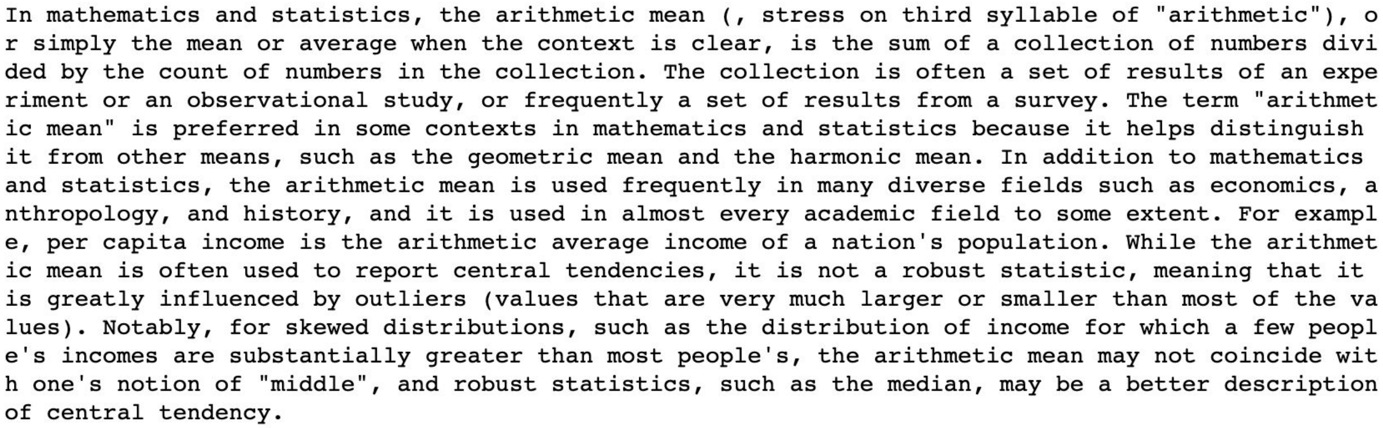
- Algorithm: Given a mention w_i, select the entity that has the minimum cosine distance with the neighboring named entities identified by the spaCy package from its entity candidates. The distance between two entities here is defined as the cosine distance between their pre-trained word embeddings from the google2vec model.

- Metric: Randomly sampled k = 20,000 mentions and labeled each mention with the entity based on the above algorithm, then computed the overall accuracy rate.

- Accuracy rate of baseline model: 66.74%

Examples

Example Text:



```
In mathematics and statistics, the arithmetic mean (, stress on third syllable of "arithmetic"), o
r simply the mean or average when the context is clear, is the sum of a collection of numbers divi
ded by the count of numbers in the collection. The collection is often a set of results of an expe
riment or an observational study, or frequently a set of results from a survey. The term "arithmet
ic mean" is preferred in some contexts in mathematics and statistics because it helps distinguish
it from other means, such as the geometric mean and the harmonic mean. In addition to mathematics
and statistics, the arithmetic mean is used frequently in many diverse fields such as economics, a
nthropology, and history, and it is used in almost every academic field to some extent. For exampl
e, per capita income is the arithmetic average income of a nation's population. While the arithmet
ic mean is often used to report central tendencies, it is not a robust statistic, meaning that it
is greatly influenced by outliers (values that are very much larger or smaller than most of the va
lues). Notably, for skewed distributions, such as the distribution of income for which a few peopl
e's incomes are substantially greater than most people's, the arithmetic mean may not coincide wit
h one's notion of "middle", and robust statistics, such as the median, may be a better description
of central tendency.
```

Mentions in the text:

mathematics, statistics, mean, experiment, observational study, survey, geometric mean, harmonic mean, economics, anthropology, history, per capita income, central tendencies, robust statistic, outlier, skewed distribution, distribution of income, median

Examples

Example Mention:
mean

Entity Candidates:
Arithmetic mean, Mean (song), Mean, Mean (album), Mean (magazine), MEAN (software bundle)

Named Entity Recognitions (spaCy):
third - ORDINAL

Selected Entity for mention ‘mean’:
Arithmetic mean

Results of Baseline Modeling

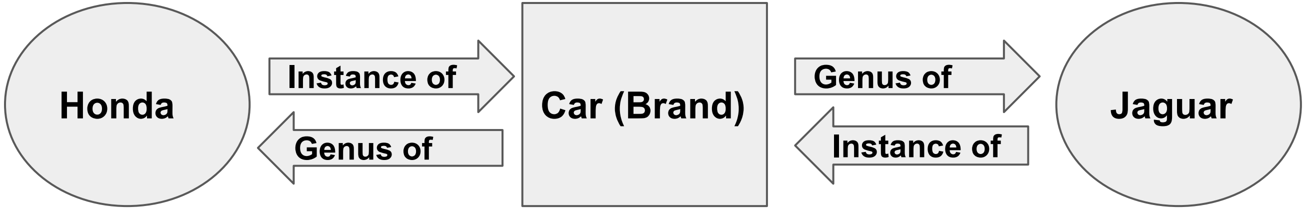
Model	Accuracy Rate
Model with highest page views as popularity	61.83%
Model with most links directed as popularity	59.135%
Model with word embeddings	66.74%

What Have We Learned Already? (Baseline Model)

Beyond the Baseline Model

Measuring Semantic Distance Via Shortest Path

Using the KDWD graph data:
Assumption: The number of steps needed represents the semantic distance



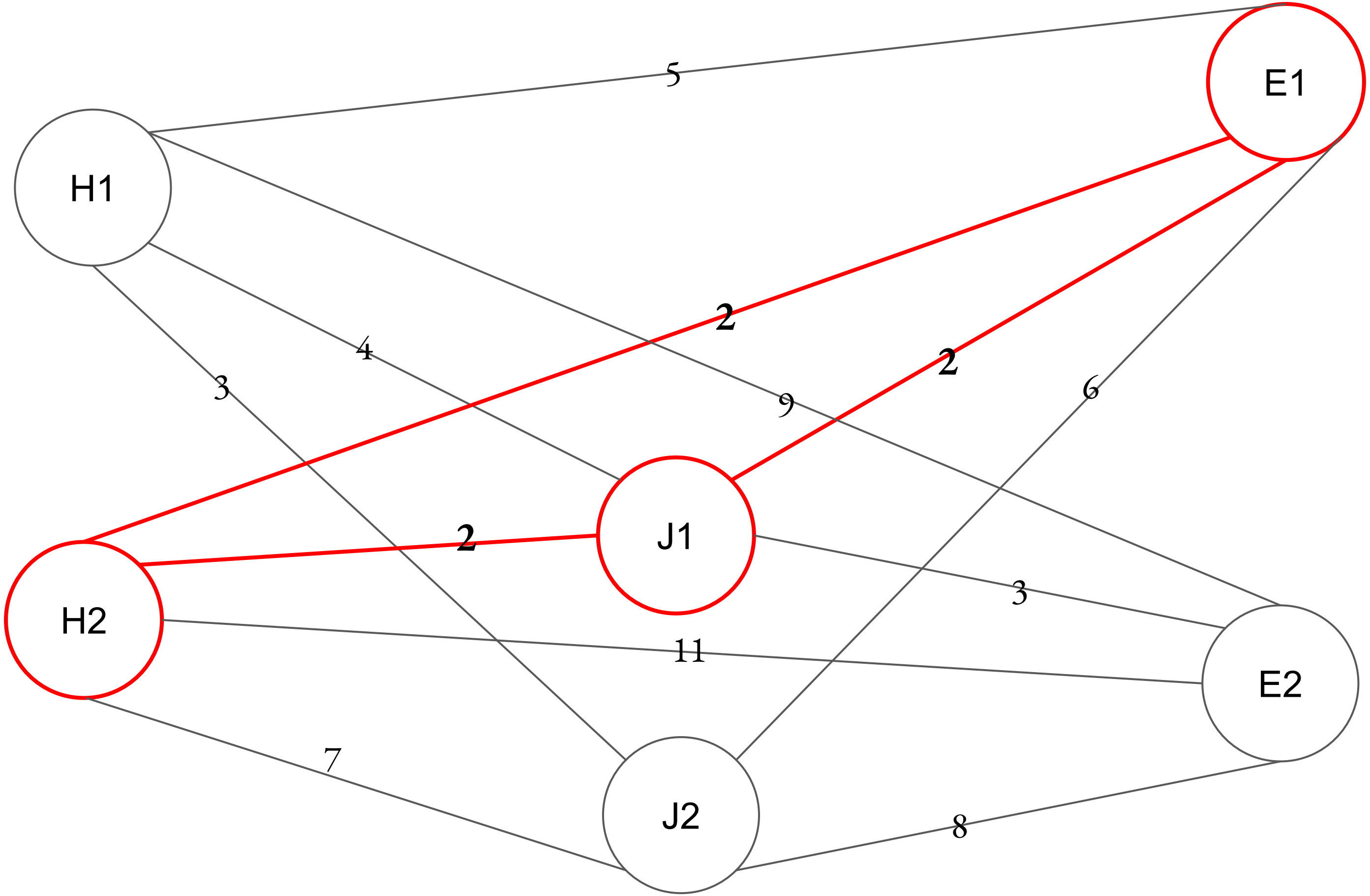
Disambiguation as a Combinatorial Optimization

Example: “Honda is competing against Jaguar in EV industry”

Honda: [H1: the entrepreneur, H2: the car brand]
Jaguar: [J1: the car brand, J2: the animal]
EV: [E1: electric vehicle, E2: expected value]

Combinations: (H1,J1,E1),(H1,J1,E2),(H1,J2,E1),(H1,J2,E2),(H2,J1,E1),(H2,J1,E2),(H2,J2,E1),(H2,J2,E2)

```
For each combination:
        Measure = sum(all pairwise semantic distances: e.g. D(H1,J1)+D(H1,C1)+D(C1,J1))
        Update least if measure < least
```


E1
H1
5
2
4
3
2
9
6
J1
2
H2
3
11
E2
7
J2
8

Achieving speed-up

1. **Data-processing to enable serial implementation**

   -Collapsed all the triplets into a single dictionary

   -Key: outbound node; Value: a list of inbound nodes

2. **Keeping a dictionary of all the previously seen distances**

   -As we loop through combinations, (a,b,c) and (a,b,d), the distance D(a,b) is seen 2x

   -The dictionary sped up the calculation by a lot.

Results

Overall Accuracy score of 85.5% over sample of 10,000 sentences

-We came pretty close to state of the art under a very different paradigm.

-Fine tuning of ad-hoc parameters could boost the performance

Future Steps

amazon
EMR

GraphFrames 0.7.0

APACHE Spark™

How did this contribute to our goal?

Thank You

Q&A