Kensho Capstone Project

Ruochen Zhao, Johannes Kolberg, Will Seaton, Hardik Gupta
October 5, 2020

Problem statement

Named-entity disambiguation (a core part of NLP pipelines) has typically treated the task of mapping each named entity in text to a node in the knowledge graph independently.

We want to improve the performance of named-entity disambiguation models by incorporating the concept of “congruence”, i.e., incorporating nearby mappings into identifying the named-entity.

Benefit to Kensho

S&P processes text data like public filings, earning call transcripts, PR announcements and high-quality news sources and links each of these to the entities discussed in them. They maintain their own internal knowledge graph called Capital IQ database in addition to KDWD.

They sell access to various front-ends for this and are building a new one that lets people see all entities mentioned in a document like an earnings call transcript.

Scope of work

- Incorporate “congruence” in these models to improve entity recognition
- Expand upon the work of previous teams that used bidirectional LSTM and feed forward neural network models for named-entity linking
- Deliver final model that improves final predictive scores, as measured by accuracy and AUC, some credible percentage above an existing baseline model
- Deliver an end-to-end pipeline such that the final model can have practical use for our partner Kensho

Team members

Ruochen Zhao
Shanghai, China

Hardik Gupta
Cambridge, MA

Johannes Kolberg
Oslo, Norway

Will Seaton
Atlanta, Georgia

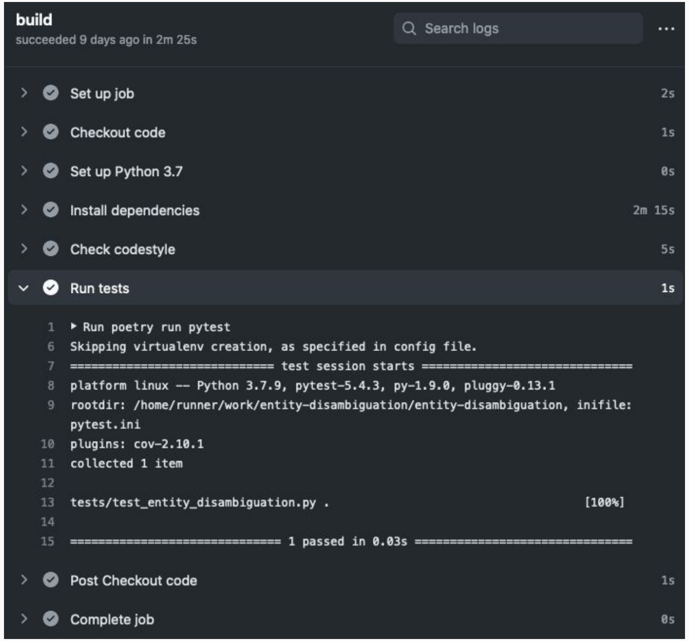
Team infrastructure

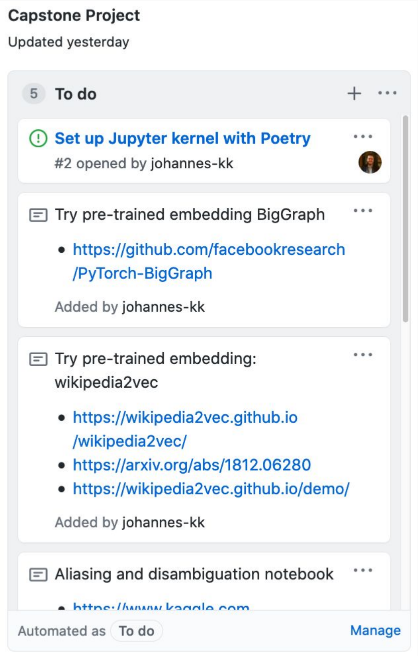
PYTHON PACKAGING AND DEPENDENCY MANAGEMENT MADE EASY

Poetry

Important Packages

- Spacy
- Gensim
- Wikipedia2vec
- NetworkX





Learning goals

- Implement an end-to-end NLP pipeline
- Understand Knowledge Graph data structure and associated analysis packages
- Approximate industry-high performance on entity disambiguation

Literature review

Literature on this sphere of NLP is vast and growing, with some of the most relevant papers for “congruence” being accepted to conferences *this* month. We’ve highlighted some of our key papers informing our approach.

Pair-Linking for Collective Entity Disambiguation: Two Could Be Better Than All. Minh C. Phan, Aixin Sun, Yi Tay, Jialong Han, and Chenliang Li. URL.

Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, Yuji Matsumoto. URL.

Improving Entity Linking through Semantic Reinforced Entity Embeddings. Feng Hou, Ruili Wang, Jun He, Yi Zhou. URL.

A Primer in BERTology: What we know about how BERT works. Anna Rogers, Olga Kovaleva, Anna Rumshisky. URL.

Robust Disambiguation of Named Entities in Text, Johannes Hoffart, Mohamed Amir Yosey, Ilaria Bordino, URL.

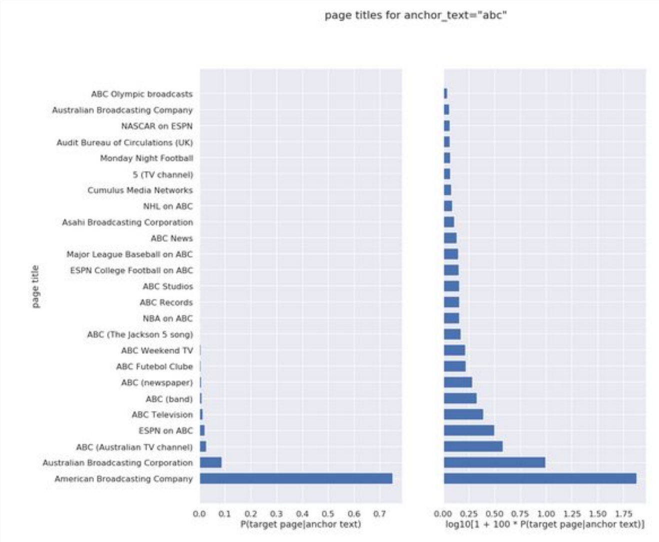
Project ideas

- Pre-trained models that incorporate knowledge implicitly through weights
  - E.g. sequence models (BERT)
  - E.g. embedding a (Google2Vec, Wikipedia2Vec)
- Congruence (aka community level entity linking)
  - Computation-based, e.g.KG traversal
  - Graph-based, e.g. minimum spanning pairwise tree
  - Entity (+ word?) embeddings
- Performance and generalisability
  - Serve predictions fast while maintaining acceptable performance
  - Transparent pipeline, preparation and processing to avoid a one-off POC
  - Mitigate tailoring to specific dataset to facilitate plug-comparability with other data

Project Approach: Establish a Baseline

We will deploy "**Anchor Link Statistics**", a method of counting the number of hyperlinks in each Wiki node and adopting that as the entity identified.

https://www.kaggle.com/kenshoresearch/kdwd-aliases-and-disambiguation



"abc" links to these pages N many times, so pick "American Broadcasting Company" because it is most populous

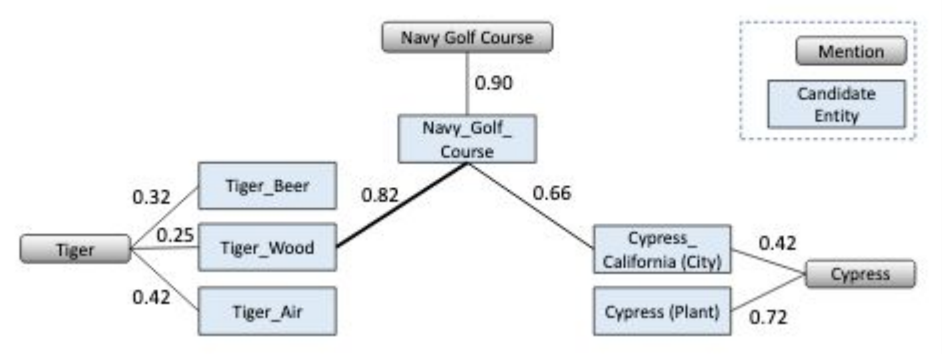
Project Approach: Word Vector Similarity

We will use Wikipedia2Vec and its pre-computed word embeddings to find similarity between the word in the input text and the title label of an associated Wikipedia “node”.

Project Approach: Page Vector Similarity

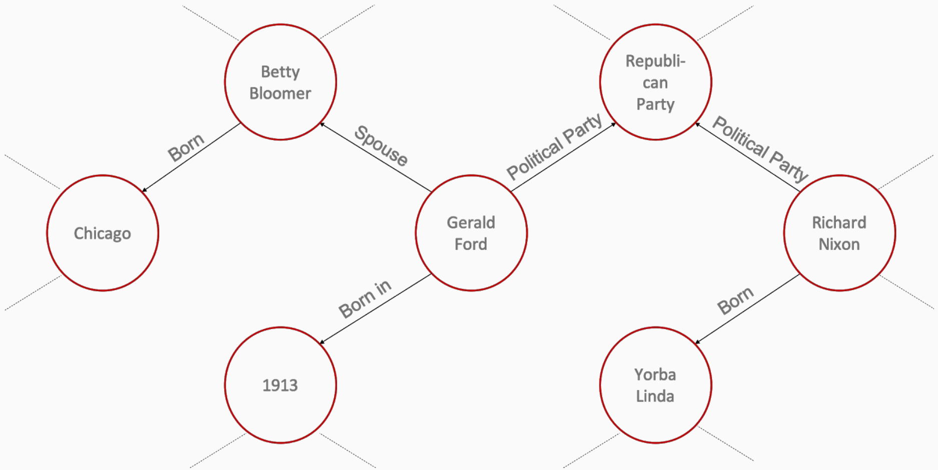
Using wiki2vec's page embeddings, we can adopt a two-step approach to entity disambiguation that incorporates scores for other entities in same text block.

How much should we limit comparative entities? Same sentence? Same paragraph? Two sentences on either side?



Project Approach: Graph Relations

Instead of comparing Wikipedia pages based on their text, we can specify strength of relationship based on node proximity on the knowledge graph, using a selected distance metric.



Exploratory data analysis

KDWD dataset

Highlights

- Derived version of the Wikidata knowledge graph with additional structure released by Kensho
- Total size of 24 GB
    - 5 million pages with raw text data
    - 51 million unique items identified
    - 7,000 different properties defined
    - 141 million entity-to-entity “statements”

Data structure

	item_id	en_label	en_description
0	1	Universe	totality of space and all contents
1	2	Earth	third planet from the Sun in the Solar System
2	3	life	matter capable of extracting energy from the e...
3	4	death	permanent cessation of vital functions
4	5	human	common name of Homo sapiens, unique extant spe...

Items

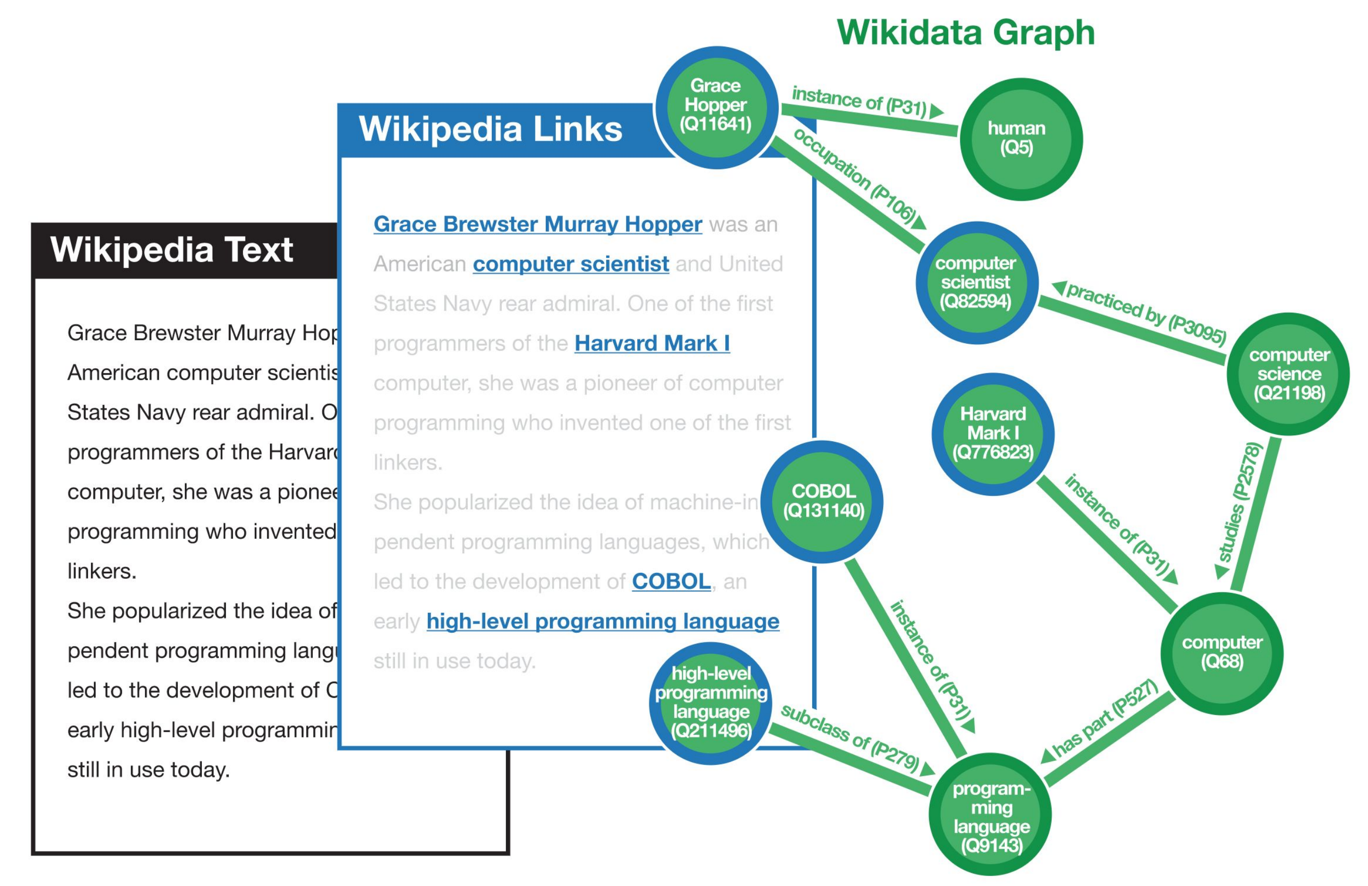
	property_id	en_label	en_description
0	6	head of government	head of the executive power of this town, city...
1	10	video	relevant video. For images, use the property P...
2	14	traffic sign	graphic symbol describing the item, used at th...
3	15	route map	image of route map at Wikimedia Commons
4	16	highway system	system (or specific country specific road type...

Properties

	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

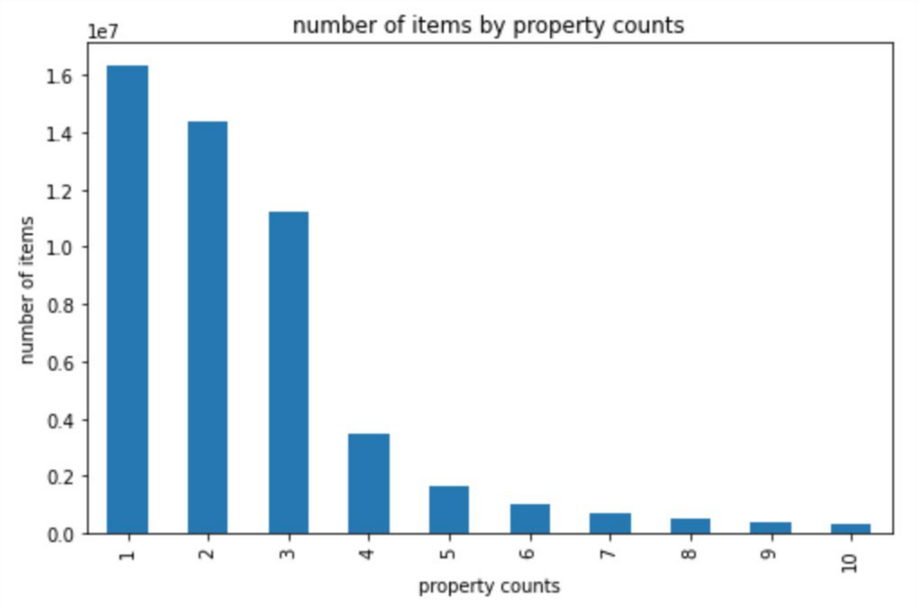
Statements

Data structure



Number of statements per item

- Average item has 2 properties
- ~90% of items have 5 or less properties
- ~2,300 items have 100 or more properties

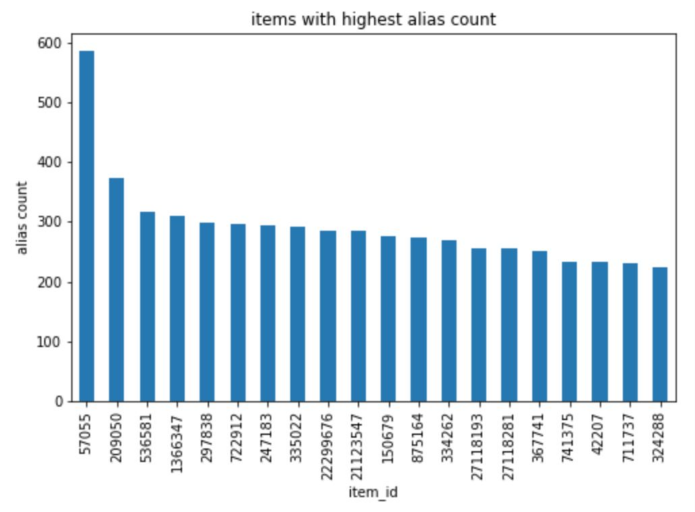


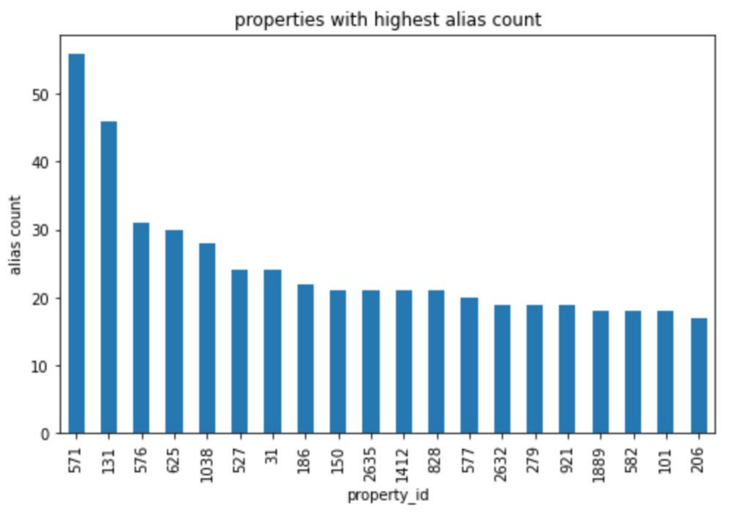
Two ontological properties

- **instance of** (P31): that class of which this subject is a particular example and member (subject typically an individual member with a proper name label)
  - “Washington DC” is an instance of “capital”, “big city”, “city in United States”
- **subclass of** (P279): all instances of these items are instances of those items; this item is a class (subset) of that item
  - “big city” is a subclass of “city”, which is a subclass of “human settlement”

Source: https://www.kaggle.com/kenshoresearch/kdwd-wikidata-introduction

Complexity is reduced using “aliases”





Acetaminophen ([Q57055]) has 586 aliases: Paracetamol, Tylenol, Paracet.

Inception ([P571]) has 56 aliases: date founded, created at, date formed.

aida-conll-yago dataset

Highlights

- Assignments of entities to the mentions of named entities annotated for the CoNLL 2003 entity recognition task
- Created by experiments in the EMNLP paper: Robust Disambiguation of Named Entities in Text: uses both popularity and context similarity
- Total size of ~800 MB
    - 176,615 tokens from the original document
    - 12.6% of tokens mapped to entities

Data Structure

	token	mention	full_mention	YAGO2	wikipedia_URL	wikipedia_ID	freebase
0	EU	B	EU	--NME--	None	None	None
1	rejects	None	None	None	None	None	None
2	German	B	German	Germany	http://en.wikipedia.org/wiki/Germany	11867	/m/0345h
3	call	None	None	None	None	None	None
4	to	None	None	None	None	None	None

Matches, Usage

- Yago2: another knowledge graph, combines wikidata and schema.org
  - 12.6%
- Wikipedia: Both URL and ID
  - 12.6%
- Freebase: former community contributed knowledge graph, was moved to wikidata
  - 12.6%

README: instructions on how to recreate (add congruence)

Provide ‘full_mention’ data to train

KWNLP dataset

Highlights

- Collection of anchored texts (texts with hyperlinks) and their corresponding wikipedia pages
- Created by Kensho in Kaggle Notebook URL (In linked_annotated_texts.jsonl file, loop through and collect (anchor, page) tuples)
- Total size of ~880 MB
    - 6,189,965 wikipedia pages
    - 15,269,229 (anchored text, page) tuples

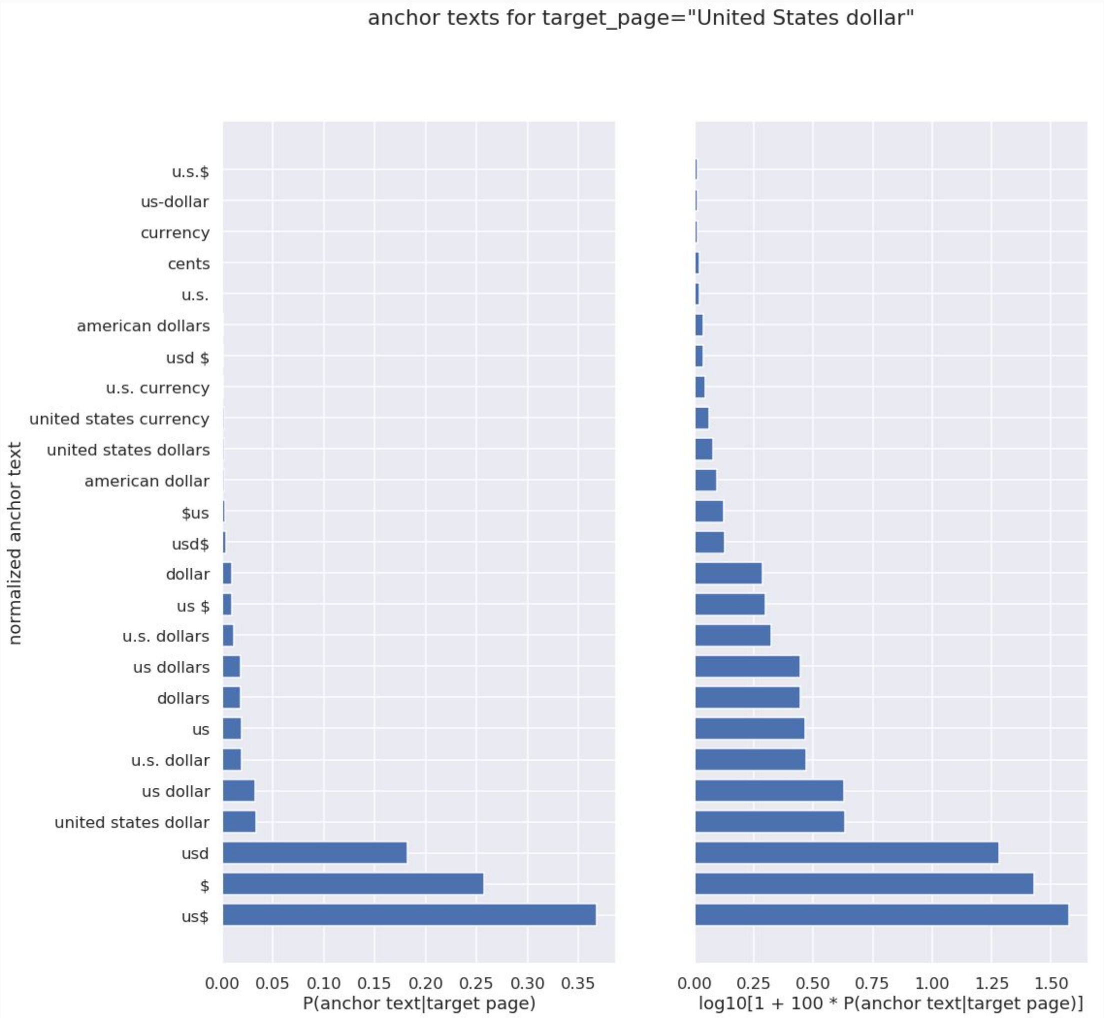
Data Structure

	page_id	item_id	page_title	views	len_artic
0	12	6199	Anarchism	35558	
1	25	38404	Autism	40081	
2	39	101038	Albedo	10770	
3	290	9659	A	29398	
4	303	173	Alabama	46680	

	anchor_text	target_page_id	count
0	United States	3434750	152451
1	World War II	32927	133668
2	India	14533	112069
3	France	5843419	109669
4	footballer	10568	101027

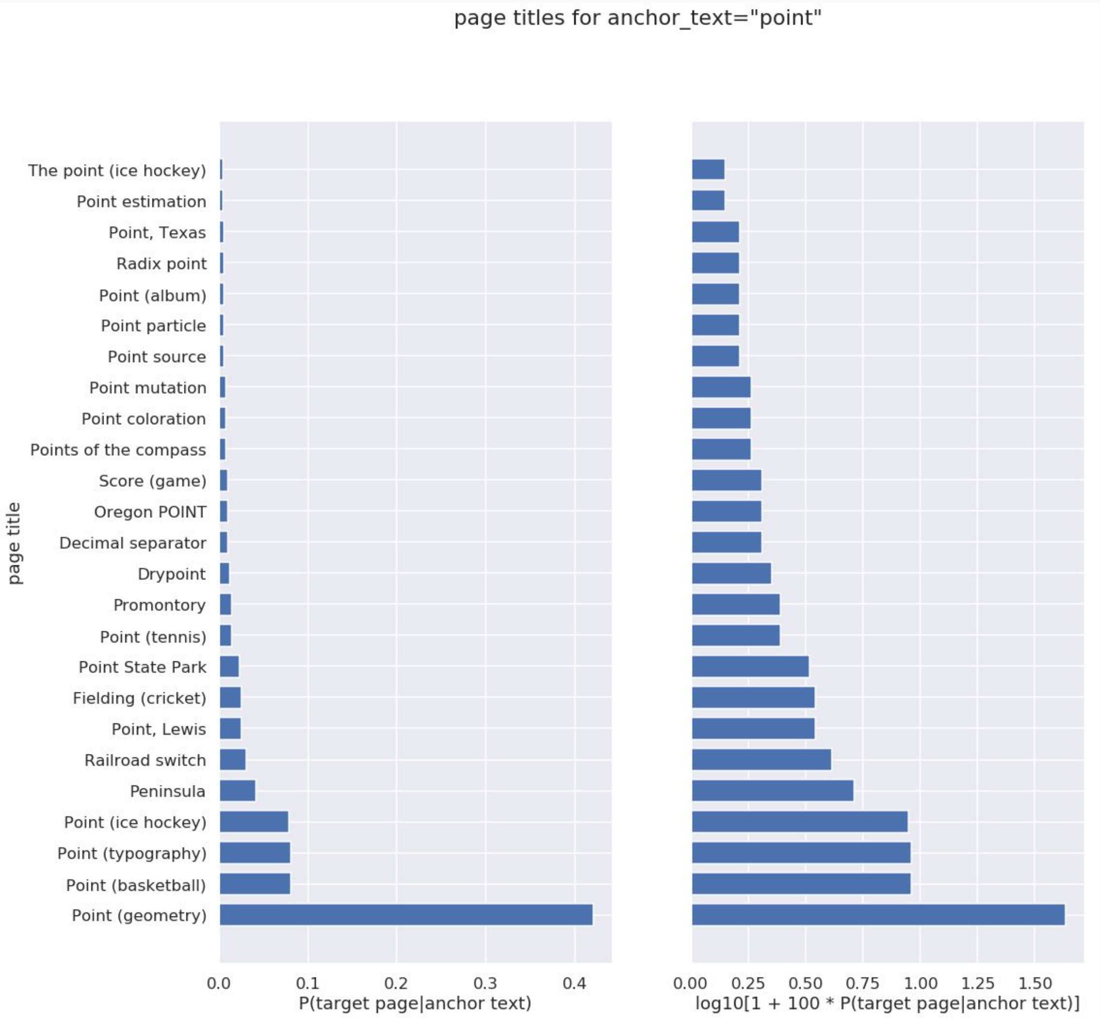
Visuals

- Provide similar candidates
- Good for generating popularity metrics
- Good for calculating conditional probabilities
  - P(anchor|text)
  - P(text|anchor)



Visuals

- Provide similar candidates
- Good for generating popularity metrics
- Good for calculating conditional probabilities
  - P(anchor|text)
  - P(text|anchor)



Usage

- Provide more data on popularity of mentions
- Context data, similarity candidates
- More data on conditional probabilities, serve as a prior for our model

Thanks!

Github repo:
https://github.com/TheDigitalFrontier/entity-disambiguation