tabularisi

Entity Linking to Knowledge Graphs

Avijit Thawani, Minda Hu, Erdong Hu,

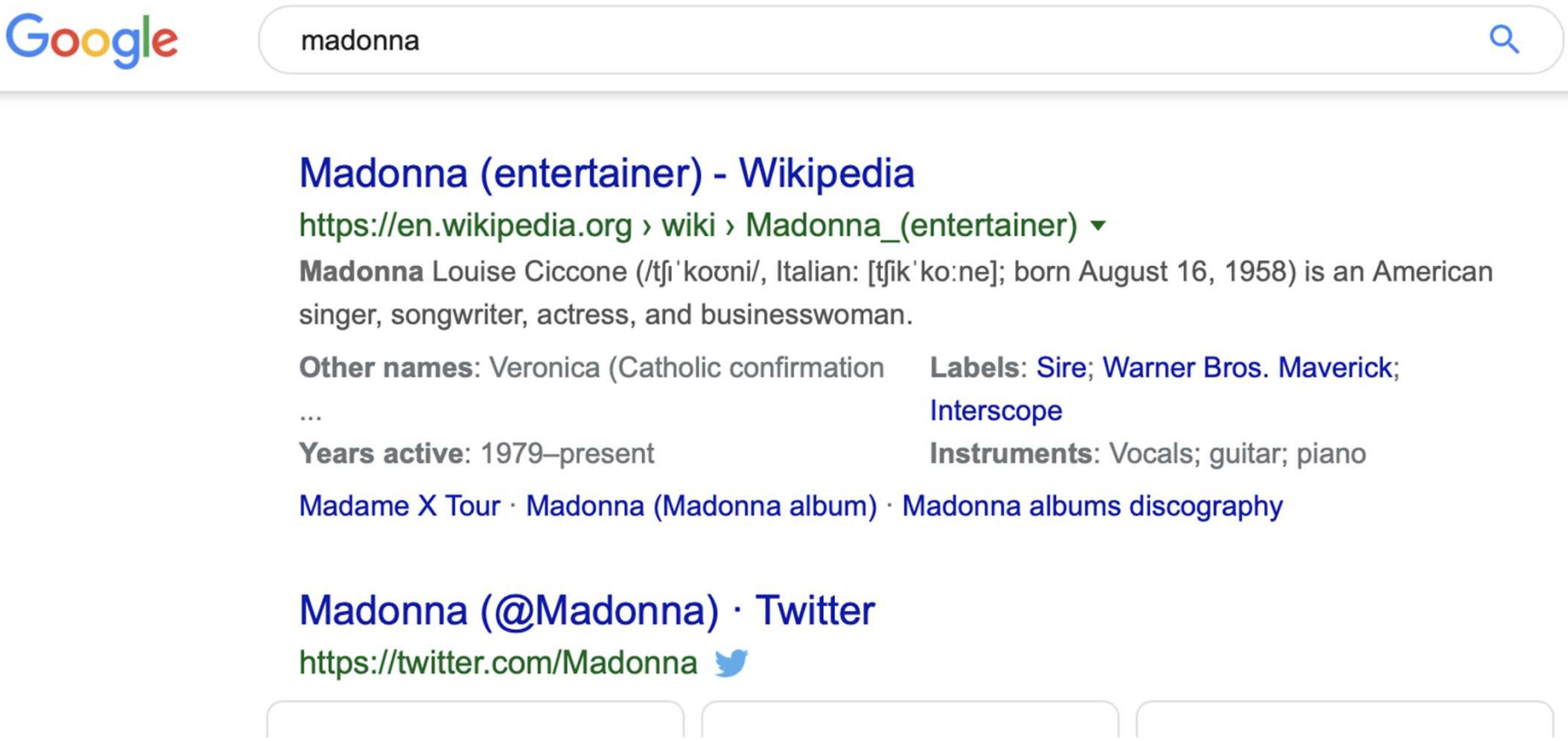
Husain Zafar, Naren Teja Divvala,

Amandeep Singh, Ehsan Qasemi, Pedro

Szekely, and Jay Pujara

The challenge

Column Header = ?
...
Madonna
...


Google
madonna
Madonna (entertainer) - Wikipedia
https://en.wikipedia.org › wiki › Madonna_(entertainer)
Madonna Louise Ciccone (/tʃɪˈkoʊni/, Italian: [tʃikˈkoːne]; born August 16, 1958) is an American singer, songwriter, actress, and businesswoman.
Other names: Veronica (Catholic confirmation ...
Years active: 1979–present
Labels: Sire; Warner Bros. Maverick; Interscope
Instruments: Vocals; guitar; piano
Madame X Tour · Madonna (Madonna album) · Madonna albums discography
Madonna (@Madonna) · Twitter
https://twitter.com/Madonna

The challenge

Column Header = Singer
...
dbpedia.org/page/Madonna_entertainer
...

The challenge

Column Header = ?
St Fernandez
Madonna
Pope John

The challenge

Column Header = Saints
dbpedia.org/page/St._Fernandez
dbpedia.org/page/Saint_Madonna
dbpedia.org/page/Pope_John

The insight

1. Generate candidates
2. Annotate easy cells
3. Figure out what's common
4. Apply on hard cells

Labels	Candidates	
John	dbr:Pope_John dbo:Person dbo:Saint dbp:Religion	
Saint Francis	dbr:St_Francis dbo:Person dbo:Saint dbp:Religion	dbr:Pope_Francis dbo:Person dbo:Saint dbp:Religion
Madonna	dbr:Madonna dbo:Person dbo:Singer dbp:Albums	dbr:Saint_Madonna dbo:Person dbo:Saint dbp:Religion
Adalgott	dbr:St_Adalgott dbo:Person dbo:Saint dbp:Religion	
St Mary	dbr:Mary dbo:Person dbo:Saint dbp:Religion	

The insight

1. Generate candidates
2. Annotate easy cells
3. Figure out what's common
4. Apply on hard cells

Labels	Candidates	
John	dbr:Pope_John dbo:Person dbo:Saint dbp:Religion	
Saint Francis	dbr:St_Francis dbo:Person dbo:Saint dbp:Religion	dbr:Pope_Francis dbo:Person dbo:Saint dbp:Religion
Madonna	dbr:Madonna dbo:Person dbo:Singer dbp:Albums	dbr:Saint_Madonna dbo:Person dbo:Saint dbp:Religion
Adalgott	dbr:St_Adalgott dbo:Person dbo:Saint dbp:Religion	
St Mary	dbr:Mary dbo:Person dbo:Saint dbp:Religion	

Labels	Candidates	
John EASY!	dbr:Pope_John dbo:Person dbo:Saint dbp:Religion	
Saint Francis	dbr:St_Francis dbo:Person dbo:Saint dbp:Religion	dbr:Pope_Francis dbo:Person dbo:Saint dbp:Religion
Madonna	dbr:Madonna dbo:Person dbo:Singer dbp:Albums	dbr:Saint_Madonna dbo:Person dbo:Saint dbp:Religion
Adalgott EASY!	dbr:St_Adalgott dbo:Person dbo:Saint dbp:Religion	
St Mary EASY!	dbr:Mary dbo:Person dbo:Saint dbp:Religion	

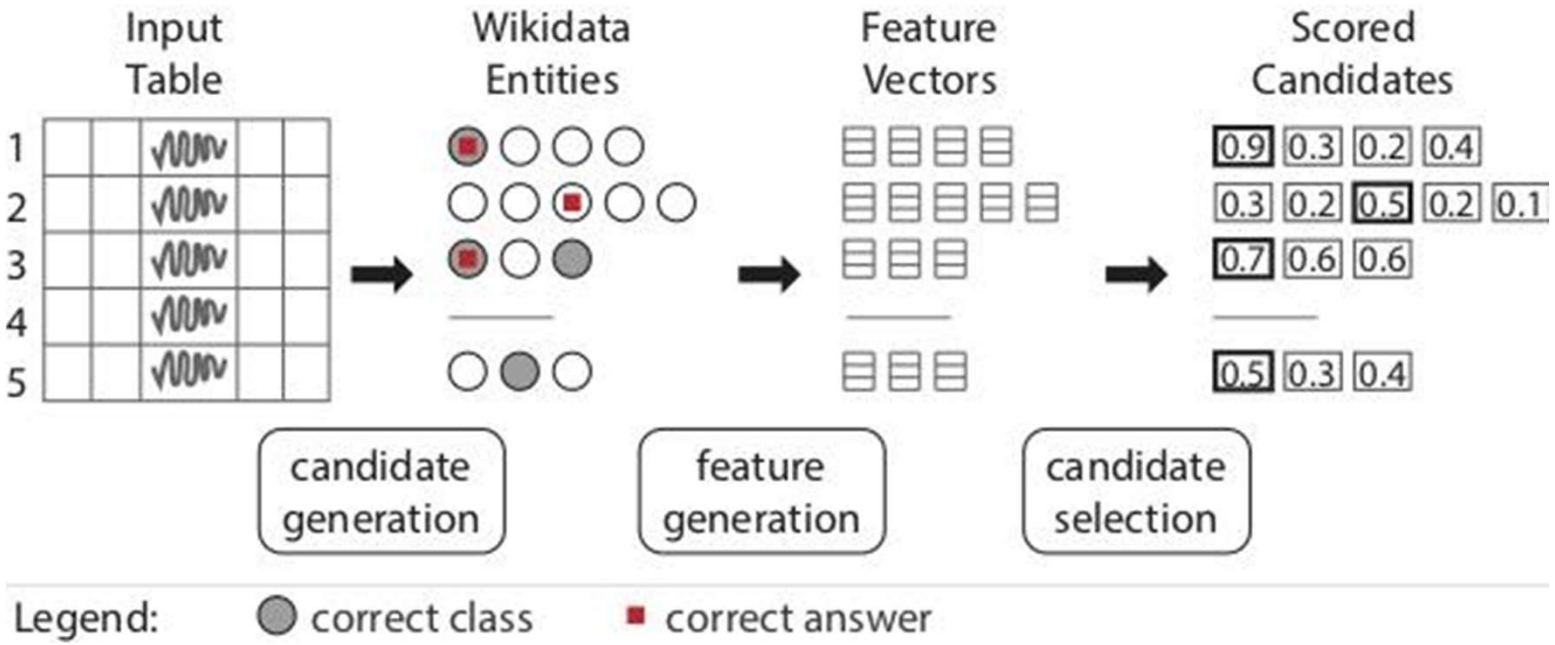
Labels	Candidates	
John EASY!	dbr:Pope_John dbo:Person dbo:Saint dbp:Religion	
Saint Francis	dbr:St_Francis dbo:Person dbo:Saint dbp:Religion	dbr:Pope_Francis dbo:Person dbo:Saint dbp:Religion
Madonna	dbr:Madonna dbo:Person dbo:Singer dbp:Albums	dbr:Saint_Madonna dbo:Person dbo:Saint dbp:Religion
Adalgott EASY!	dbr:St_Adalgott dbo:Person dbo:Saint dbp:Religion	
St Mary EASY!	dbr:Mary dbo:Person dbo:Saint dbp:Religion	

**Easy* can be redefined.

Labels	Candidates	
John	dbr:Pope_John dbo:Person dbo:Saint dbp:Religion	
Saint Francis	dbr:St_Francis dbo:Person dbo:Saint dbp:Religion	dbr:Pope_Francis dbo:Person dbo:Saint dbp:Religion
Madonna	dbr:Madonna dbo:Person dbo:Singer dbp:Albums	dbr:Saint_Madonna dbo:Person dbo:Saint dbp:Religion
Adalgott	dbr:St_Adalgott dbo:Person dbo:Saint dbp:Religion	
St Mary	dbr:Mary dbo:Person dbo:Saint dbp:Religion	

Semantic Feature	dbo: Person	dbo: Saint	dbo: Singer	dbp: Religion	dbp: Albums
dbr: madonna	1	0	1	0	1
dbr: saint_ madonna	1	1	0	1	0

The approach



The insight

1. Generate candidates
2. Annotate easy cells
3. Figure out what's common
4. Apply on hard cells

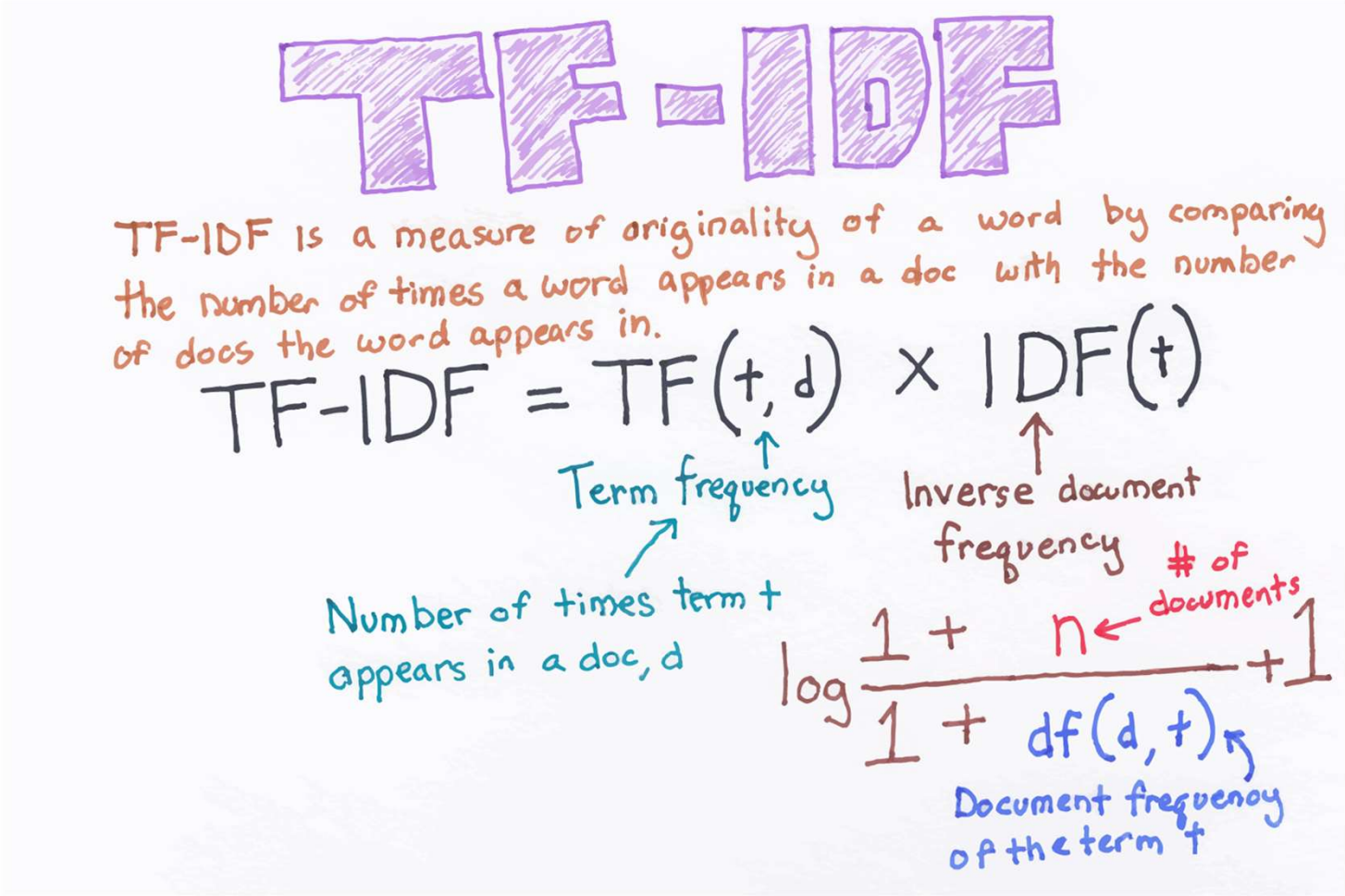


Image Source: becominghuman.ai blog

Saliency

Each class is a feature:	(Saint,	Person,	Singer,	...)
Saint Madonna:	(1,	1,	0,	...)
Madonna - the singer:	(0,	1,	1,	...)
TF (term frequency)	(3,	3,	1,	...)
IDF (inverse document frequency)	(0.48,	0.25,	0.95,	...)
Saliency of feature = TF x IDF	(1.5,	0.75,	0.95,	...)

Labels	Candidates	
John	dbr:Pope_John dbo:Person dbo:Saint dbp:Religion	
Saint Francis	dbr:St_Francis dbo:Person dbo:Saint dbp:Religion	dbr:Pope_Francis dbo:Person dbo:Saint dbp:Religion
Madonna	dbr:Madonna dbo:Person dbo:Singer dbp:Albums	dbr:Saint_Madonna dbo:Person dbo:Saint dbp:Religion
Adalgott	dbr:St_Adalgott dbo:Person dbo:Saint dbp:Religion	
St Mary	dbr:Mary dbo:Person dbo:Saint dbp:Religion	

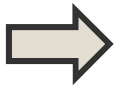
Semantic Feature	dbo: Person	dbo: Saint	dbo: Singer	dbp: Religion	dbp: Albums
dbr: madonna	1	0	1	0	1
dbr: saint_ madonna	1	1	0	1	0

Labels	Candidates	
John	dbr:Pope_John dbo:Person dbo:Saint dbp:Religion	
Saint Francis	dbr:St_Francis dbo:Person dbo:Saint dbp:Religion	dbr:Pope_Francis dbo:Person dbo:Saint dbp:Religion
Madonna	dbr:Madonna dbo:Person dbo:Singer dbp:Albums	dbr:Saint_Madonna dbo:Person dbo:Saint dbp:Religion
Adalgott	dbr:St_Adalgott dbo:Person dbo:Saint dbp:Religion	
St Mary	dbr:Mary dbo:Person dbo:Saint dbp:Religion	

Semantic Feature	dbo: Person	dbo: Saint	dbo: Singer	dbp: Religion	dbp: Albums	
(**TF**) Term Frequency	5	4	1	4	1	
Document Frequency	8	6	2	6	2	
(**IDF**) Inverse Document Frequency	0.05	0.18	0.65	0.18	0.65	
(**TFIDF**) TF ⊗ IDF	0.25	0.72	0.65	0.72	0.65	**Score** $\text{TFIDF} \cdot v_i^T$
($\mathbf{v_1}$) dbr: madonna	1	0	1	0	1	1.55
($\mathbf{v_2}$) dbr: saint_ madonna	1	1	0	1	0	1.69

Entity Linking

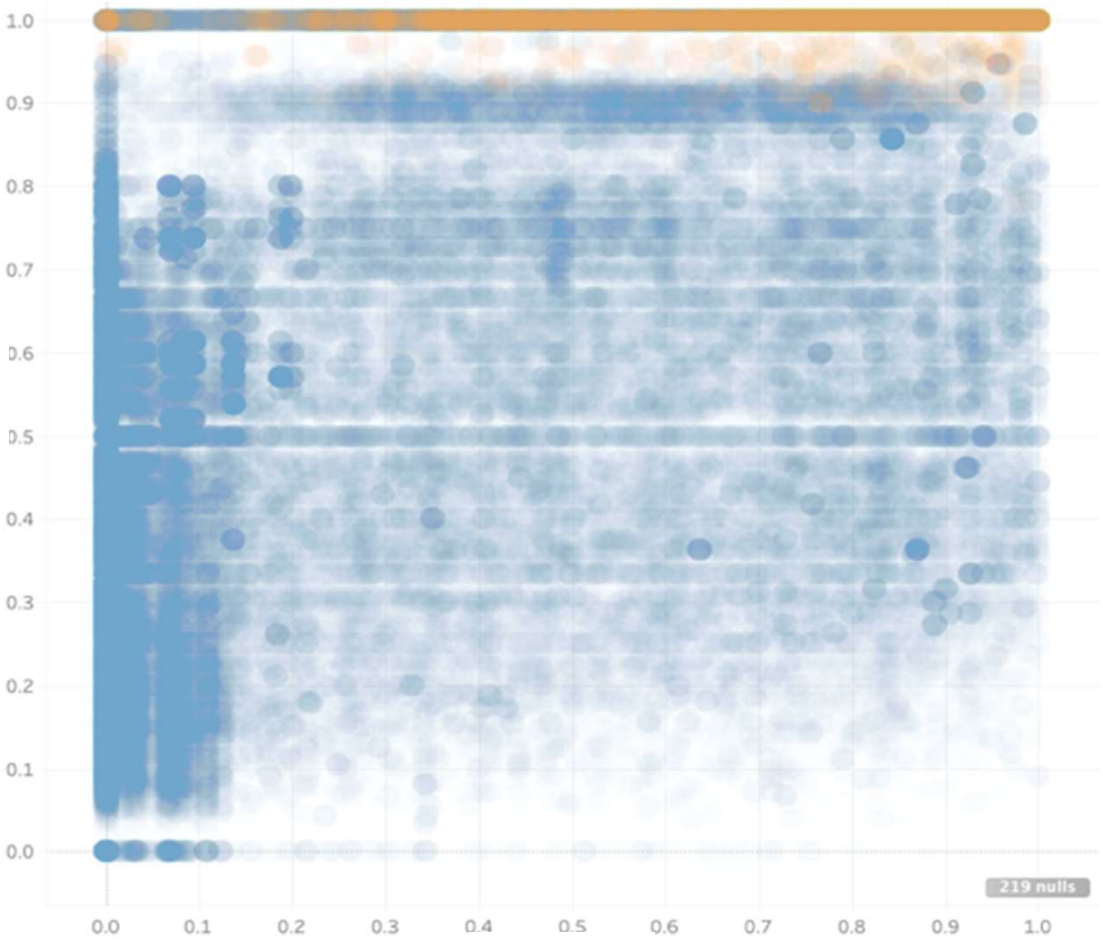
Column Header = ?
St Fernandez
Madonna
Pope John



Column Header = ?
dbpedia.org/page/St._Fernandez
dbpedia.org/page/Saint_Madonna
dbpedia.org/page/Pope_John

Analysis

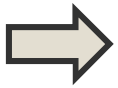
Levenshtein Similarity



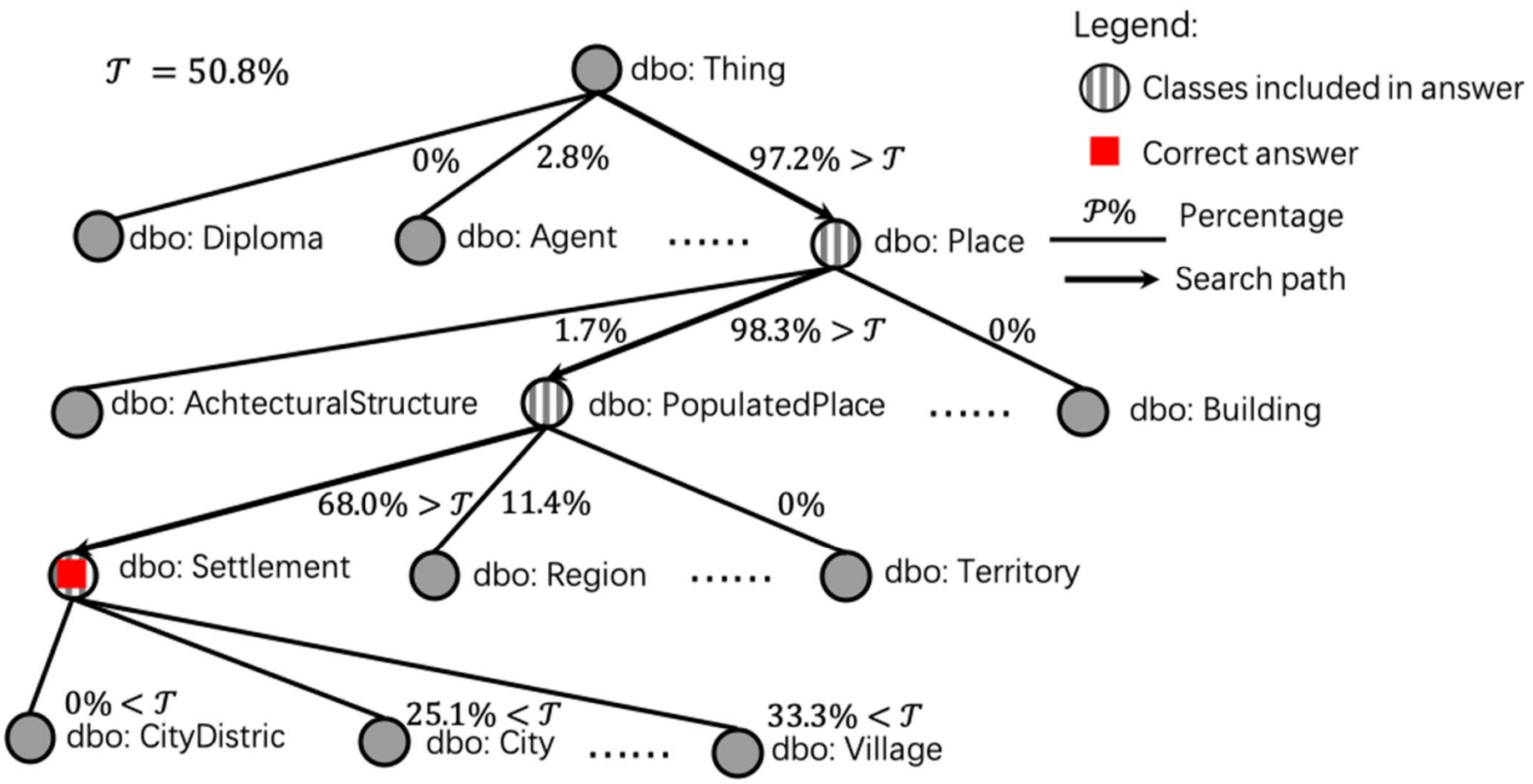
tf-idf

Semantic Labelling

Column Header = ?
dbpedia.org/page/St._Fernandez
dbpedia.org/page/Madonna_entertainer
dbpedia.org/page/Pope_John



Column Header = Saints
dbpedia.org/page/St._Fernandez
dbpedia.org/page/Saint_Madonna
dbpedia.org/page/Pope_John


$\mathcal{T}$ = 50.8%
dbo: Thing
0%
2.8%
97.2% > $\mathcal{T}$
dbo: Diploma
dbo: Agent
.....
dbo: Place
1.7%
98.3% > $\mathcal{T}$
0%
dbo: AchtecturalStructure
dbo: PopulatedPlace
.....
dbo: Building
68.0% > $\mathcal{T}$
11.4%
0%
dbo: Settlement
dbo: Region
.....
dbo: Territory
0% < $\mathcal{T}$
dbo: CityDistric
25.1% < $\mathcal{T}$
dbo: City
.....
33.3% < $\mathcal{T}$
dbo: Village
Legend:
Classes included in answer
Correct answer
$\mathcal{P}$%
Percentage
Search path

Hands On

```
git clone https://github.com/usc-isi-i2/wikidata-wikifier

cd wikidata-wikifier/wikifier/sample_files

curl -X POST -F 'file=@clubs.csv' -F columns=clubs
http://sitaware.isi.edu/wikify -o 'clubs_result.csv'

curl -X POST -F 'file=@countries.csv' -F columns=countries
http://sitaware.isi.edu/wikify -o 'countries_result.csv'

curl -X POST -F 'file=@footballers.csv' -F columns=footballers
http://sitaware.isi.edu/wikify -o 'footballers_result.csv'
```