

EECS 395/495 Project: Word Representations

Integrating Extracted KBs

- Several extracted KBs exist:
 - Google Knowledge Graph
 - NELL
 - TextRunner
 - WikiTables
- Problem: KBs are separate silos, with different knowledge formalisms (or lack thereof)

If two KBs contain different sets of knowledge, represented in different ways, how can the KBs share knowledge with each other?

Idea: Use natural language.

Each KB K estimates $P_K(\mathbf{w})$ over sentences \mathbf{w}
e.g., $\text{City}(\mathbf{x}) \Rightarrow \text{increased } P(\text{“cities such as } \mathbf{x}\text{”})$

Example

Cities containing skyscrapers?



Language Probability Query

$P(\text{cities such as } x) * P(\text{skyscraper in } x)$

KB2

Text
Corpus

Response

$P(x) = \{(\text{Shanghai}, 0.6), (\text{Montreal}, 0.2), (\text{New York City}, 0.2)\}$

KB1

Similarity Query

Distributionally similar to
Shanghai, Montreal, NYC?

Response

New York City
London
Singapore
Hong Kong
Paris
Tokyo
Zurich
Washington, D.C.
Chicago

KB3

Wiki tables Shanghai, New York City, Montreal

Wikipedia Table: **Global City Competitiveness Index** from **Glob...**
[-<see other tables](#)

Rank	City	Score	PWC 2008
1	New York City	71.4	1406
2	London	70.4	565
3	Singapore	70.0	215
4	Hong Kong	69.3	320
4	Paris	69.3	564
6	Tokyo	68.0	1479
7	Zurich	66.8	
8	Washington, D.C.	66.1	
9	Chicago	65.0	574

Project

- Create *word representations* from KBs
 - Vectors of numbers
 - E.g., the probability distributions $P(z \mid \text{word})$ learned in hmwk #4
 - For use in “similarity queries”
- Must make *some* use of ideas from course

Possibilities

- KB relations
 - Build graphical model as in homework #4?
 - Latent Dirichlet Allocation (we'll discuss this)
- Tables
 - numeric cells down to smaller space?
- Text
 - NLP tools?

Goals for today

- Form teams with good distribution of skills
 - Each team needs some CS people for
 - Matching our dev/test sets with resources
 - Programming any models required
- Develop a reasonable “plan of action”
 - Including who will do what