

# SHORT TEXT CONCEPTUALIZATION USING A PROBABILISTIC KNOWLEDGEBASE

---

Yangqiu Song, Haixun Wang, Zhongyuan Wang,  
Hongsong Li, Weizhu Chen

Web Search and Mining  
Microsoft Research Asia

# Semantic Analysis of Text

- Implicit semantic analysis (topic modeling): [Blei et al., 2003]
  - Based on multiple document statistics
  - Topic as bag-of-words; difficult to interpret
  - Require sufficient statistical information (long text documents)
- Explicit semantic analysis: [Gabrilovich and Markovitch, 2005]
  - Wikipedia titles as concepts
  - Map text to Wikipedia titles
  - Bag of concepts cannot model concepts in human's mental world
- Probabilistic knowledgebase: **Probase**
  - Mining concepts/instances/attributes from the Web
  - Using probabilities to represent popularities of concepts/instances/attributes
  - Rich concepts with different granularities

**1** More than 2.7 million concepts (core)  
automatically harnessed from 1.68 billion documents

**2** Computation/Reasoning  
enabled by scoring:

Consensus:

e.g., is there a company called Apple?

Typicality:

e.g. how likely you think of Apple when  
you think about companies?

Ambiguity:

e.g., does the word *Apple*, sans any  
context, represent *Apple the company*?

Similarity:

e.g., how likely is an actor also a celebrity?

Freshness:

e.g., *Pluto as a dwarf planet* is a claim more  
fresh than *Pluto as a planet*.

...

Capture  
concepts  
in human  
mind

Represent  
them in a  
*computable*  
form

Transform  
them to  
machines

Machines  
have better  
understanding  
of human  
world

**4** A little knowledge  
goes a long way after  
machines  
acquire a  
human  
touch

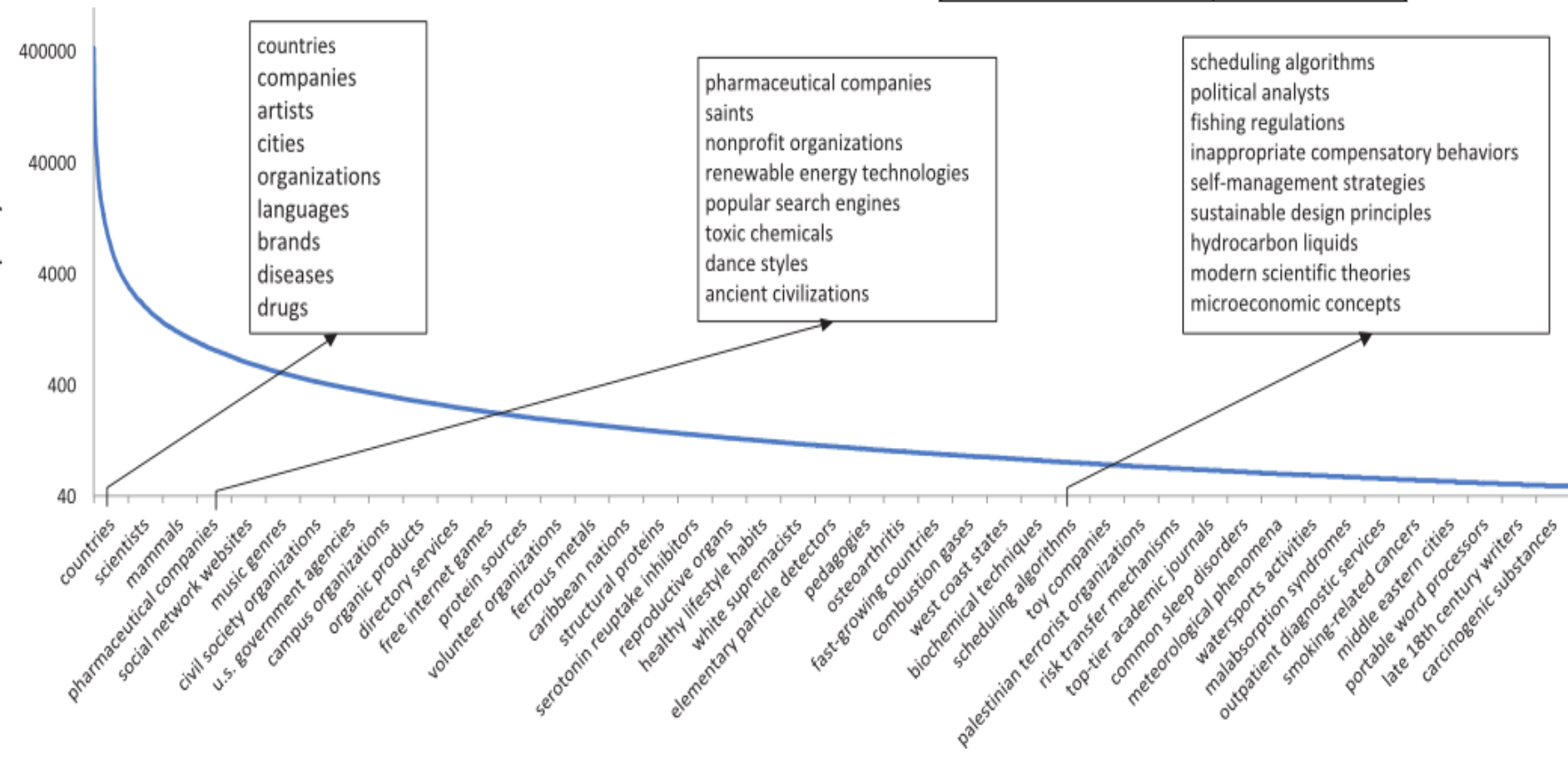
**3** Give machines a new CPU  
(Commonsense Processing Unit)  
powered by a distributed graph engine called Trinity.

# Data Sources

- Text data
  - 1.68 billion Web documents from Bing
- Patterns for single statements
  - Concept-instance: Hearst pattern [Hearst, 1992] (“A such as B, C and D”, etc.)
  - Instance-attributes: “What is A of B?”, etc.
  - Examples:
    - Good: “countries such as **USA** and **Japan** ...”
    - Tough: “animals other than **cats** such as **dogs** ...”
    - Hopeless: “At Berklee, I was playing with **cats** such as *Jeff Berlin*, *Mike Stern*, *Bill Frisell*, and *Neil Stubenhaus*.”
- Incorporated with other knowledgebases
  - Wikipedia
  - Freebase
- Post-processing/refining
  - Denoising
  - Taxonomy building
  - ...

# Rich Concepts in Probase

|                   | # of concepts    |
|-------------------|------------------|
| Freebase [6]      | 1,450            |
| WordNet [13]      | 25,229           |
| WikiTaxonomy [21] | < 127,325        |
| YAGO [28]         | 149,162          |
| DBPedia [4]       | 259              |
| ResearchCyc [15]  | ≈ 120,000        |
| KnowItAll [12]    | N/A              |
| TextRunner [5]    | N/A              |
| OMCS [25]         | N/A              |
| NELL [8]          | 123              |
| <b>Probase</b>    | <b>2,653,872</b> |



# Comparison of Knowledgebases

|          | WordNet<br>Hypernyms                                                                                                                                                                                                                      | Wikipedia<br>Categories                                                                                                                                                  | Freebase Types                                                                                                                                                                                                                                                                                    | Probase Concepts                                                                                                                                                                                                                                                                                               |
|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Cat      | Feline; Felid; Adult male;<br><b>Man; Gossip; Gossiper;</b><br>Gossipmonger;<br>Rumormonger;<br>Rumourmonger;<br>Newsmonger; <b>Woman;</b><br><b>Adult female; Stimulant;</b><br><b>Stimulant drug;</b> Excitant;<br>Tracked vehicle; ... | Domesticated animals;<br>Cats; Felines; Invasive<br>animal species;<br>Cosmopolitan species;<br><b>Sequenced genomes;</b><br><b>Animals described in</b><br><b>1758;</b> | <b>TV episode; Creative work;</b><br><b>Musical recording;</b> Organism<br>classification; Dated location;<br>Musical release; <b>Book; Musical</b><br><b>album; Film character;</b><br><b>Publication; Character</b><br><b>species;</b> Top level domain;<br>Animal; Domesticated animal;<br>... | Animal; Pet; Species;<br>Mammal; Small animal;<br>Thing; Mammalian<br>species; Small pet;<br>Animal species;<br>Carnivore; Domesticated<br>animal; Companion<br>animal; Exotic pet;<br>Vertebrate; ...                                                                                                         |
| Language | Communication; Auditory<br>communication; Word;<br>Higher cognitive process;<br><b>Faculty; Mental faculty;</b><br><b>Module; Text;</b> Textual<br>matter;                                                                                | Languages; Linguistics;<br>Human communication;<br>Human skills; <b>Wikipedia</b><br><b>articles with ASCII art</b>                                                      | Employer; Written work;<br>Musical recording; Musical<br>artist; Musical album;<br>Literature subject; Query;<br>Periodical; Type profile;<br>Journal; Quotation subject;<br>Type/domain equivalent topic;<br>Broadcast genre; Periodical<br>subject; Video game content<br>descriptor; ...       | <b>Instance of:</b> Cognitive<br>function; Knowledge;<br>Cultural factor; Cultural<br>barrier; Cognitive<br>process; Cognitive ability;<br>Cultural difference; Ability;<br>Characteristic;<br><b>Attribute of:</b> Film; Area;<br>Book; Publication;<br>Magazine; Country; Work;<br>Program; Media; City; ... |

# Short Text Example

- Example. Given three tweets, how to compute the similarities?

animal, species, mammal,  
wild or domestic  
animal, family member;

- T1. “HA! just came back from walking the **dog**... man... It isn’t **easy** i guarantee!”
- T2. “**Microsoft Visual Studio** 2010 Express for **Windows** P. The same powerful and **easy** to use **Visual Studio**. ”
- T3. “Dear **Yahoo!**, It’s so cute how you keep trying... Sincerely, **Google**. ”

search engine,  
company, site,  
IT company,  
large company

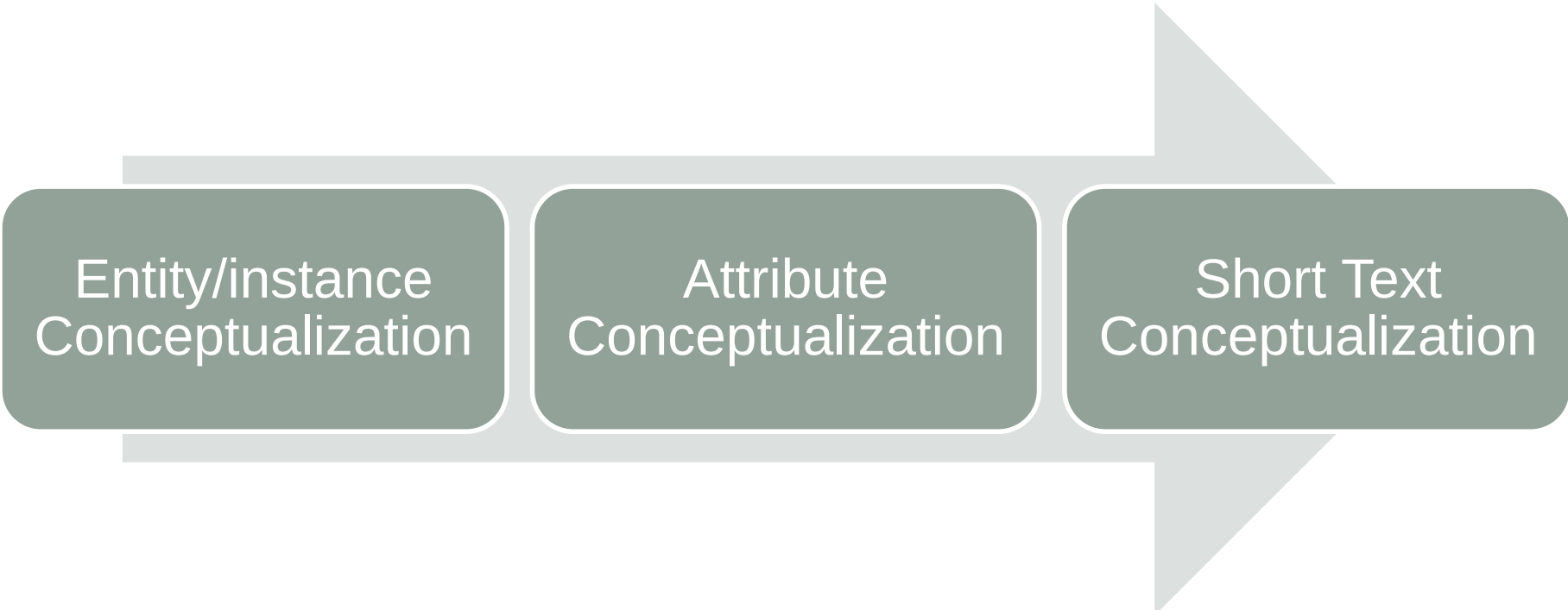
Application,  
development tool,  
code editor, product

Company, corporation,  
large company,  
software company,  
industry leader

System, platform,  
product, popular  
operating system

- Clustering/classification of short text can be done on the concept distribution after conceptualization

# Short Text Conceptualization: Step by Step



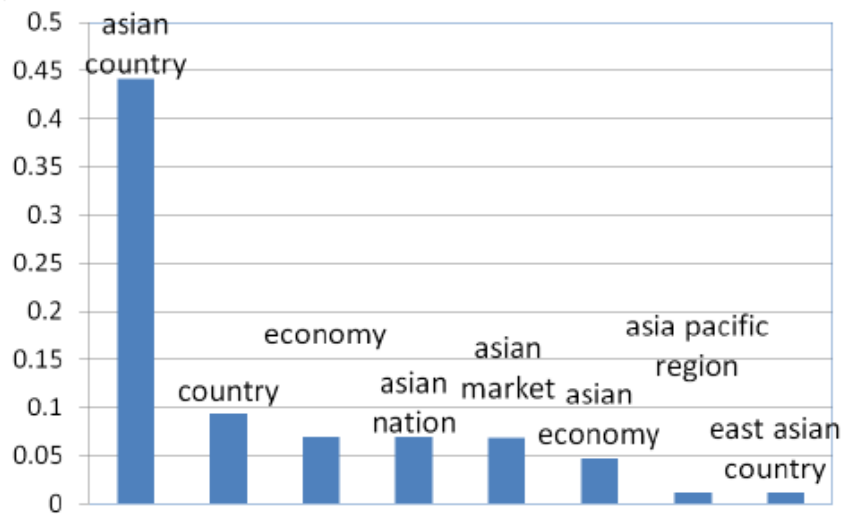
Entity/instance  
Conceptualization

Attribute  
Conceptualization

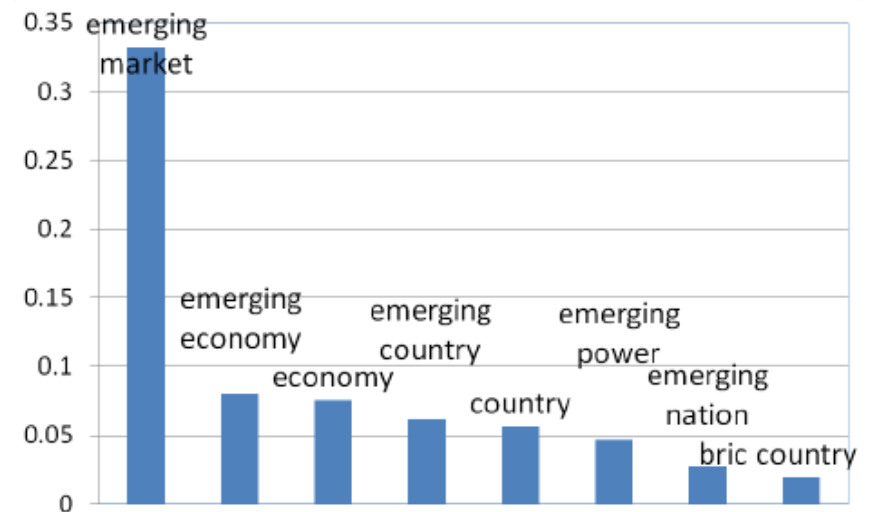
Short Text  
Conceptualization

# Entity Conceptualization

- Given “China (I), India (I), Japan (I), Singapore (I)”



- Given “China (I), Russia (I), India (I), Brazil (I)”



# Entity Conceptualization

- Given a set of instances

$$E = \{e_i, i \in 1, \dots, M\}$$

- Target Concept (Naïve Bayes Rule)

$$P(c_k|E) = \frac{P(E|c_k)P(c_k)}{P(E)} \propto P(c_k) \prod_{i=1}^M P(e_i|c_k).$$

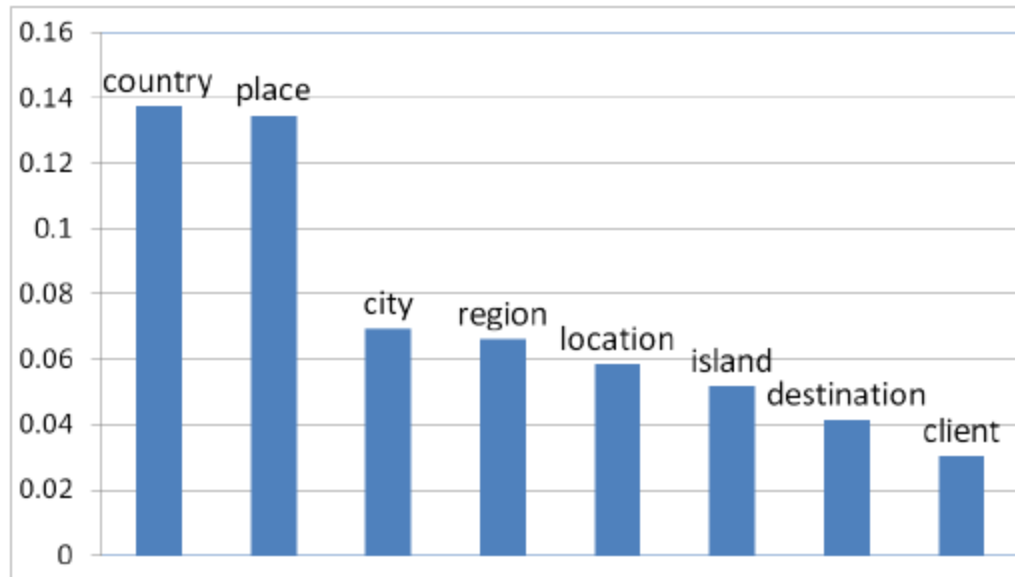
- Where  $c_k$  is a concept, and

$$P(e_i|c_k) = \frac{P(e_i, c_k)}{P(c_k)}$$

- is computed based on the concept-entity co-occurrence

# Attribute Conceptualization

- Given “population (A), location (A), president (A)”



# How to Infer Concepts from Attributes?

- Given a set of attributes  $A = \{a_j, j \in 1, \dots, N\}$ .
- The Naïve Bayes Rule gives

$$P(c_k|A) = \frac{P(A|c_k)P(c_k)}{P(A)} \propto P(c_k) \prod_{j=1}^N P(a_j|c_k),$$

- where

$$P(a_j|c_k) = \sum_{i: e_i \in E} P(a_j|e_i)P(e_i|c_k),$$

(university, florida state university, 75)  
(university, harvard university, 388)  
(university, university of california, 142)  
(country, china, 97346)  
(country, the united states , 91083)  
(country, india , 80351)  
(country, canada , 74481)



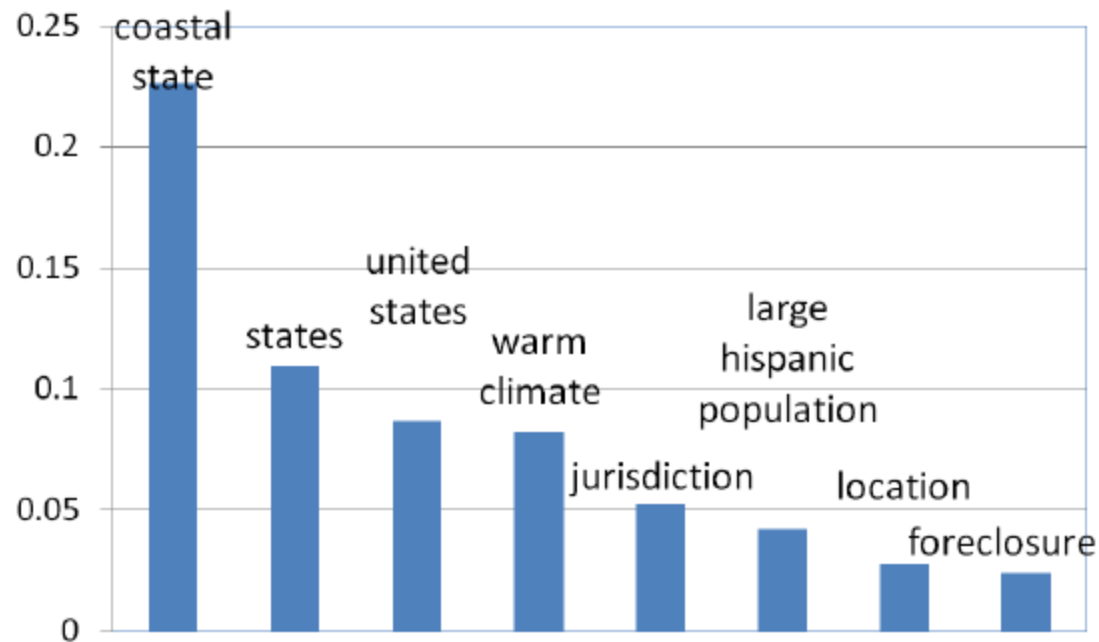
(florida state university, website, 34)  
(harvard university, website, 38)  
(university of california, city, 12)  
(china, capital, 43)  
(the united states , capital, 32)  
(india , population, 35)  
(canada , population, 21)



(university, website, 4568)  
(university, city, 2343)  
(country, capital, 4345)  
(country, population, 3234)  
.....

# When Type of Term is Unknown

- Given “California (U), Florida (U), population (U)”



# Mixture Models

- Given a set of terms with unknown types  $T = \{t_l\}, l = 1, \dots, L$

- Generative model**

$$P(t_l|c_k) = P(t_l|z_l = 1, c_k)P(z_l = 1|c_k) + P(t_l|z_l = 0, c_k)P(z_l = 0|c_k)$$

Using Naive Bayesian rule gives:

$$P(c_k|T) = \frac{P(T|c_k)P(c_k)}{P(T)} \propto P(c_k) \prod_l^L P(t_l|c_k)$$

- Discriminative model** (Noisy-OR)

$$P(c_k|t_l) = 1 - (1 - P(c_k|t_l, z_l = 1))(1 - P(c_k|t_l, z_l = 0))$$

And using twice Bayesian rule gives:

$$P(c_k|T) \propto P(c_k) \prod_l^L P(t_l|c_k) \propto \frac{\prod_l P(c_k|t_l)}{P(c_k)^{L-1}}$$

- where  $z_l = 1$  indicates “instance” and  $z_l = 0$  indicates “attribute”

# Short Text Clustering Settings

- Problem 1 (unique concepts): use the hashtag keywords to retrieve tweets in 3 categories:
  - 1. Microsoft, Yahoo, Google, IBM, Facebook
  - 2. cat, dog, fish, pet, bird
  - 3. Brazil, China, Russia, India
- Problem 2 (concepts with subtle differences): use hashtag keywords to retrieve tweets in 4 categories:
  - 1. United states, American, Canada
  - 2. Malaysia, China, Singapore, India, Thailand, Korea
  - 3. Angola, Egypt, Sudan, Zambia, Chad, Gambia, Congo
  - 4. Belgium, Finland, France, Germany, Greece, Spain, Switzerland

# Comparison Results

Clustering NMI scores on Twitter data.

| Method                        | @Problem1                           | @Problem2                           |
|-------------------------------|-------------------------------------|-------------------------------------|
| Original Data                 | $0.215 \pm 0.010$                   | $0.452 \pm 0.076$                   |
| LDA ( $1 \times$ Cluster Num) | $0.161 \pm 0.065$                   | $0.114 \pm 0.037$                   |
| LDA ( $2 \times$ Cluster Num) | $0.067 \pm 0.022$                   | $0.069 \pm 0.024$                   |
| WordNet                       | $0.195 \pm 0.070$                   | $0.074 \pm 0.074$                   |
| Freebase                      | $0.531 \pm 0.164$                   | $0.204 \pm 0.037$                   |
| Wikipedia (Category-Link)     | $0.540 \pm 0.077$                   | $0.336 \pm 0.089$                   |
| Wikipedia (ESA)               | $0.351 \pm 0.132$                   | $0.340 \pm 0.800$                   |
| Probase (Top 10)              | $0.318 \pm 0.110$                   | $0.490 \pm 0.029$                   |
| Probase (Top 20)              | $0.479 \pm 0.111$                   | $0.555 \pm 0.019$                   |
| Probase (Top 50)              | $0.559 \pm 0.123$                   | <b><math>0.632 \pm 0.066</math></b> |
| Probase (Top 500)             | <b><math>0.826 \pm 0.062</math></b> | $0.301 \pm 0.189$                   |
| Probase (Top 5000)            | $0.690 \pm 0.176$                   | $0.095 \pm 0.084$                   |

# Conclusion and Future Work

- Probase
  - Open information extracted from the Web
  - Rich concept coverage as human's mental world
  - Concepts with different granularities
- Short text analysis
  - Model short text semantics as concept distribution
  - Knowledgebase can help short text understanding
- Further works
  - De-noising Probase
  - Better taxonomy building and inference
  - More relationships
- Further applications
  - Conceptualization of
    - Query / URL / Document / Image, video with tags and surrounding text
  - Semantic relatedness / similarities

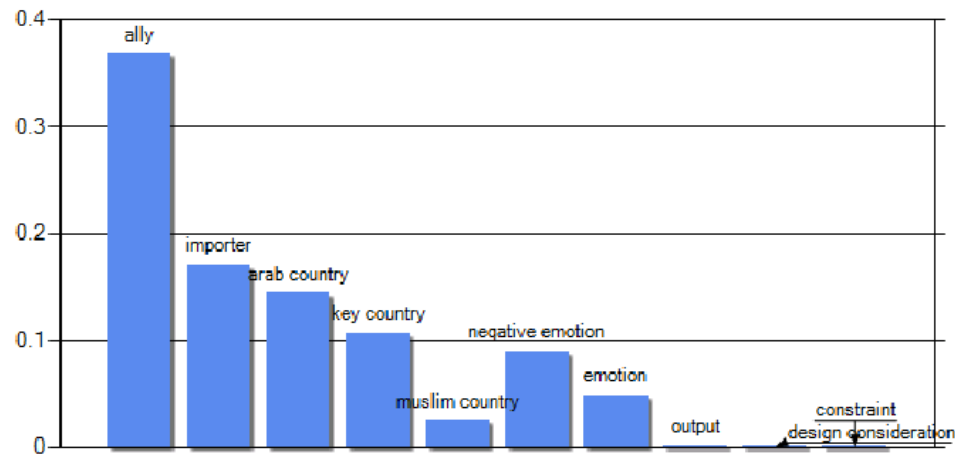
Thank you for your attention!

# Short Text Conceptualization

Please input your query/text

Oil rose by more than \$2 on Thursday as tensions in Saudi Arabia and Bahrain fueled fears of further supply disruption while investors weighed the impact on energy demand from quake-hit Japan.

Parse



|           |                               |
|-----------|-------------------------------|
| Location: | Saudi Arabia; Bahrain; Japan; |
| Date:     | Thursday;                     |
| Money:    | \$2;                          |

|         | Term group                                                                                       | Concept group                                                                                                                                                                                                            |
|---------|--------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Group 0 | japan; saudi arabia; bahrain; impact; investor; tension; fear; supply disruption; energy demand; | ally; importer; arab country; key country; muslim country; oil producer; force; strong emotion; u.s. ally; gulf country; middle eastern country; american ally; regional market; region; psychological problem; measure; |
| Group 1 | fear; tension; investor;                                                                         | negative emotion; emotion;                                                                                                                                                                                               |
| Group 2 | energy demand;                                                                                   | output; design consideration; constraint;                                                                                                                                                                                |

# Entity Abstraction Example

## Entity Abstraction

Turn the most likely concept which can generalize all the entities. The top entities in the concept are also returned. ("Russia", "India" and "Brazil", then click 'Abstract' and you can find something interesting.)

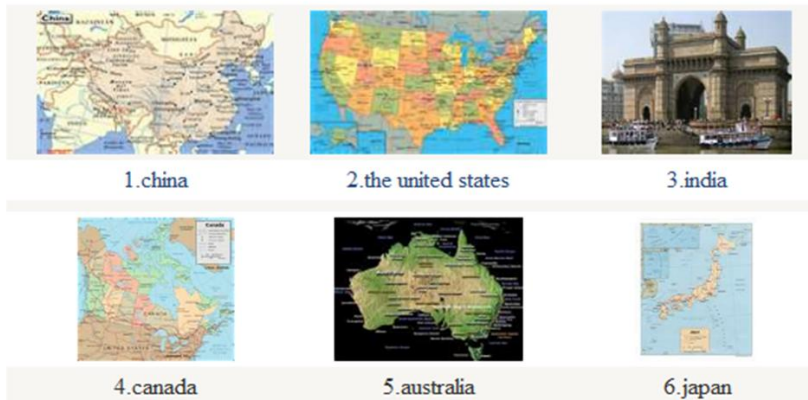
China

Russia

Abstract

I think you are talking about **country**

Entities in this concept include



## Entity Abstraction

Turn the most likely concept which can generalize all the entities. The top entities in the concept are also returned. ("Russia", "India" and "Brazil", then click 'Abstract' and you can find something interesting.)

China

Russia

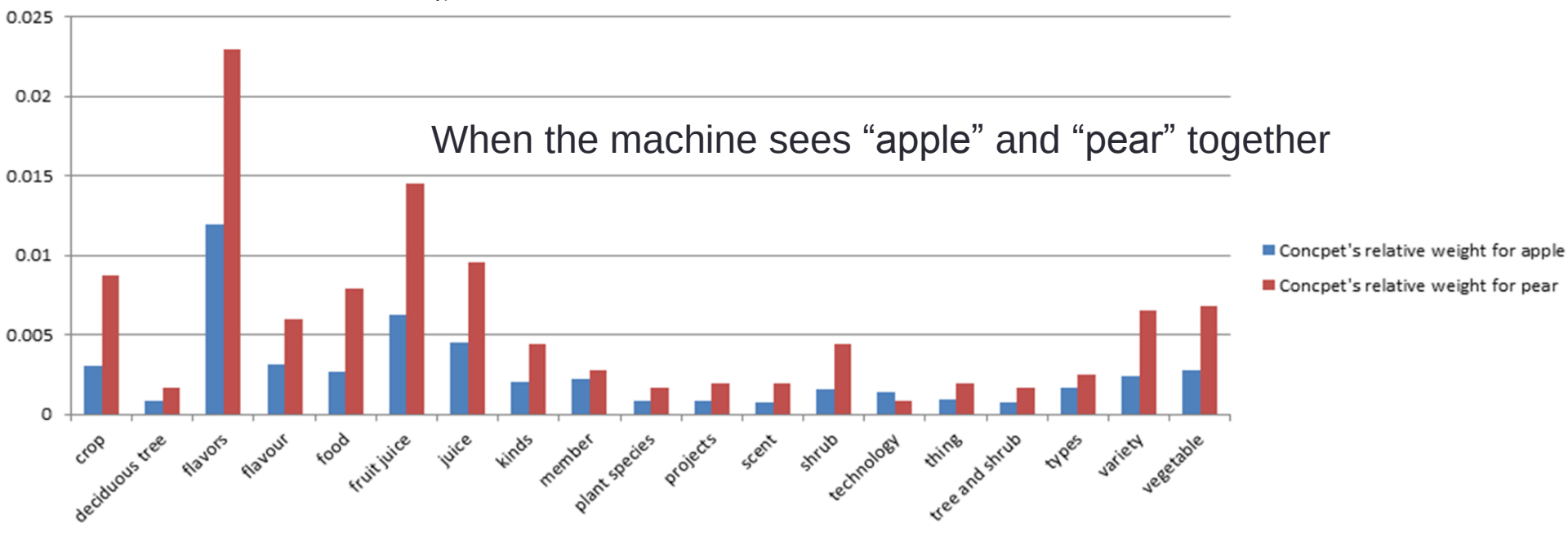
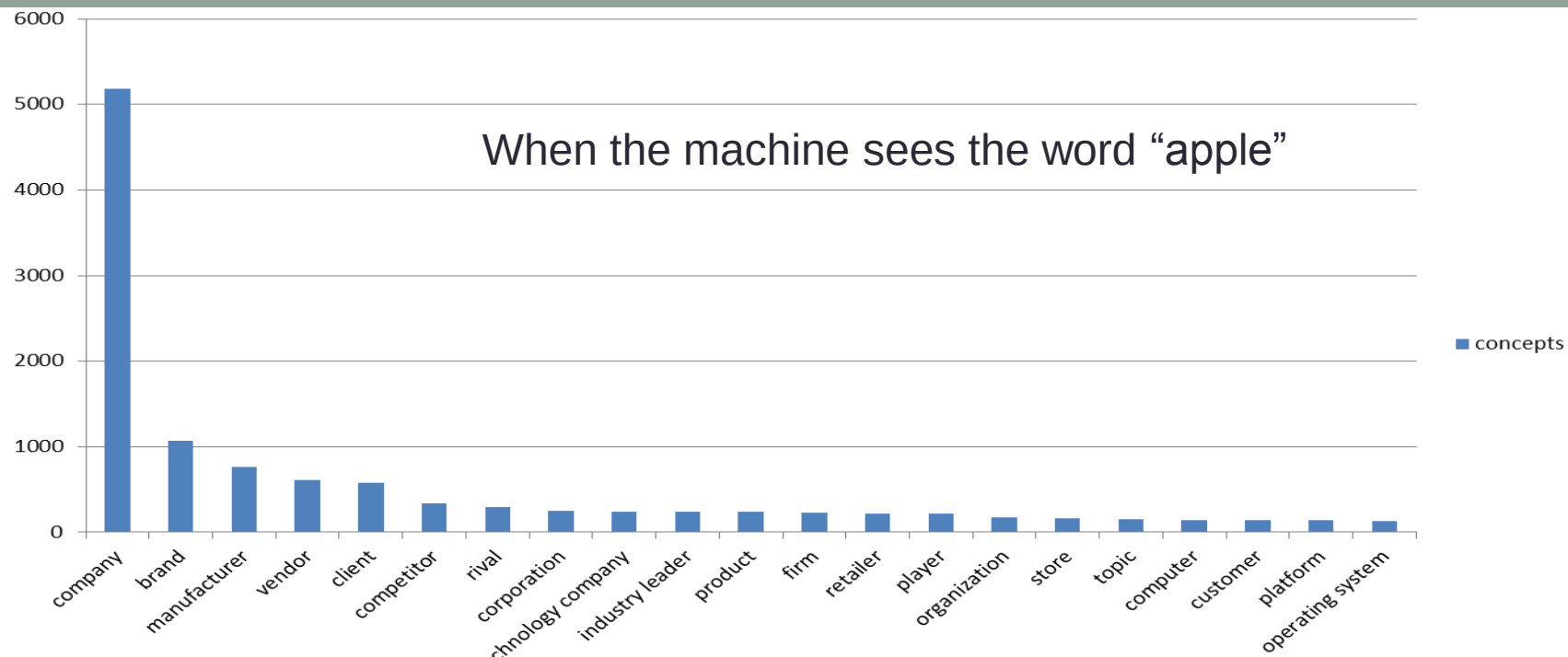
India

Abstract

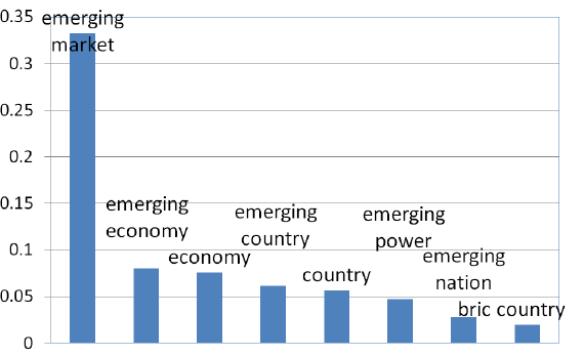
I think you are talking about **emerging market**

Entities in this concept include

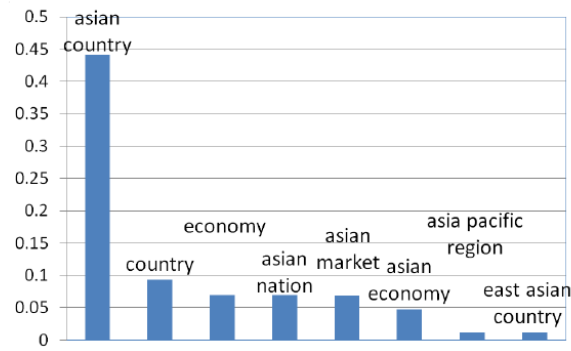




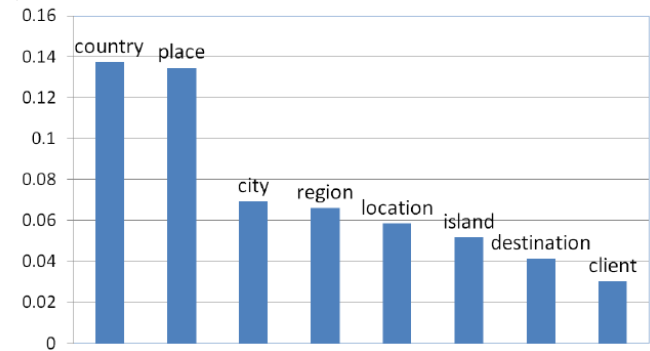
# Example



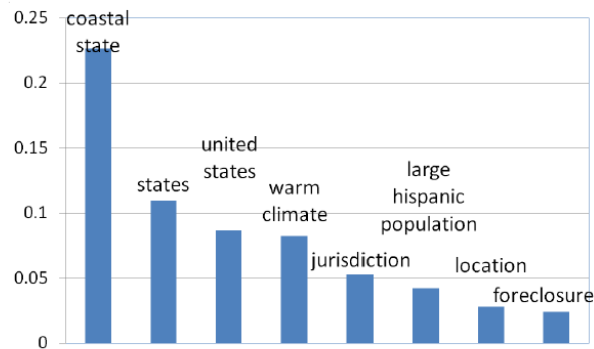
(a) China (I), Russia (I), India (I), Brazil (I)



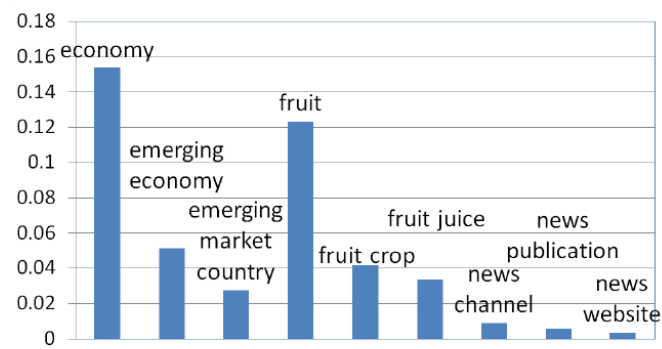
(b) China (I), India (I), Japan (I), Singapore (I)



(c) population (A), location (A), president (A)



(d) California (U), Florida (U), population (U)



(e) China (U), Brazil (U), Russia (U), apple (U), banana (U), BBC (U), New York Time (U)

# Probase Schema

