Master of Technology in Knowledge Engineering

Text Mining

Advanced Topics in Text Mining

Rajaraman Kanagasabai
Institute for Infocomm Research
email: kanagasa@i2r.a-star.edu.sg

Overview

- Introduction to Advanced Text mining
  - » The Semantic Gap in Text mining
  - » Traditional Approaches to Text Mining
    - ♦ Keywords, NLP, & Ontologies
  - » The Deep Learning Approach
- Text Mining with Deep Learning
  - » Background
  - » Major DL Approaches to Text Mining
- Case studies
- Workshop
  - » Hands-on Exercises with Word2Vec

What is Text Analytics?

- *“discovery of* ***new*** *previously unknown information, by* ***automatically*** *extracting information from a usually* **large amount** *of different unstructured textual resources”*

- Other definitions
  - » Use of **computational techniques** to extract high quality information from text
  - » Extract and discover knowledge hidden in text **automatically**

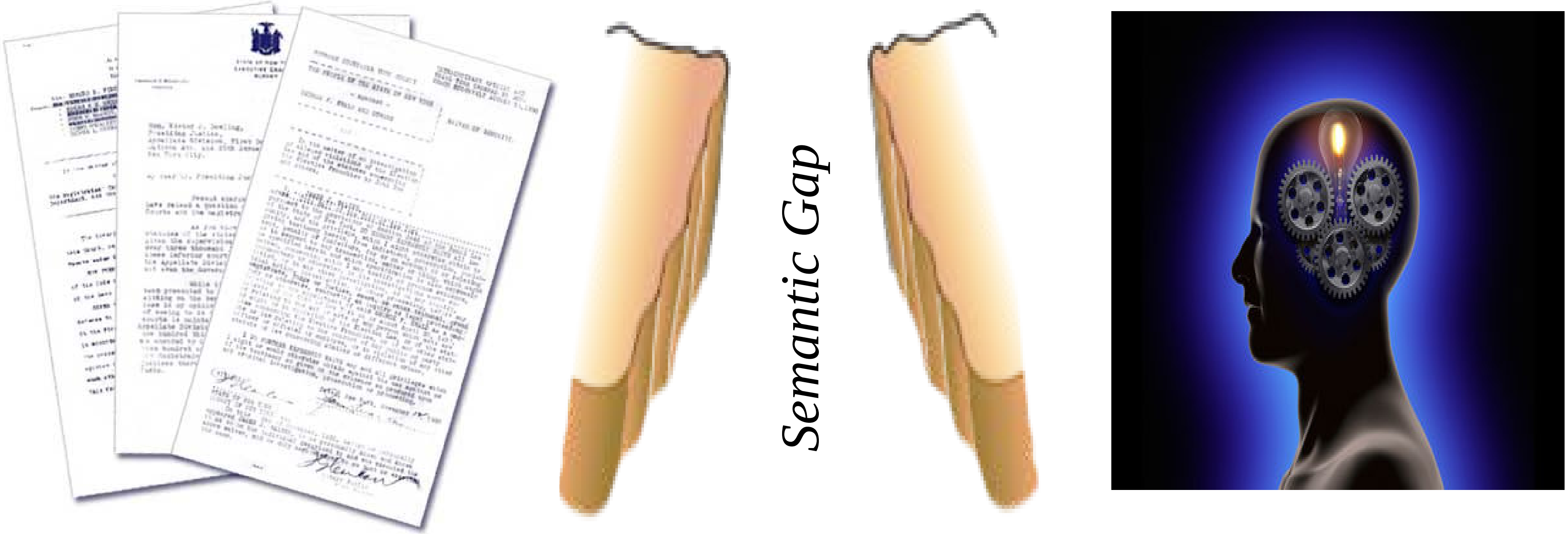
Key Text Analytics Problems

- Analyze Document Collections
  - » Information Retrieval
  - » Classification (Supervised Learning)
  - » Clustering (Unsupervised Learning)
- Analyze Document
  - » Summarization
  - » Information Extraction: Extract Names, Relations, Facts
- Analyze Sentence
  - » Sentiment Analysis
  - » Co-reference Resolution

Why not just apply standard DA?

> “Text expresses a vast, rich range of information, but encodes this information in a form that is difficult to decipher automatically.”
>
> -- Marti A. Hearst,
> “Untangling Text Data Mining,” 1999

Key Issue



Raw Text

Semantic Gap

Text Understanding

Example 1

- *It would be a great help if you can assist me, and I will appreciate the favor*

=> *Known as Synonymy*

Example 2

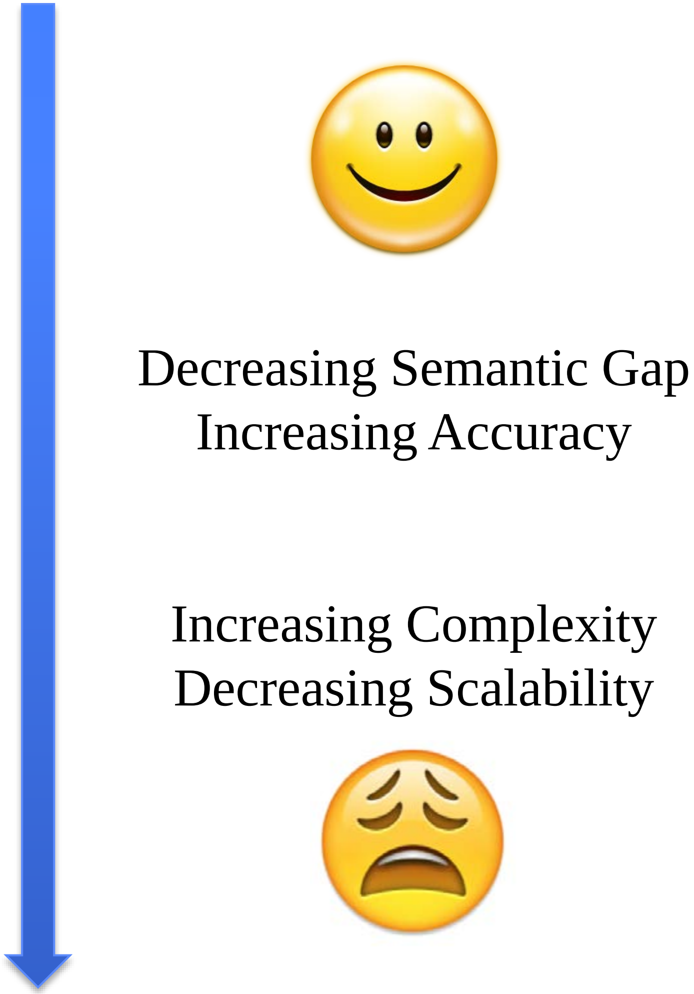
- *"I banked on him to meet at the Deutche Bank located near the river bank", said Mr. Banks.*

=> *Known as Polysemy*

Overview of Approaches

Approaches

- Keywords-based
  - » Statistical
  - » Zero semantics approach
- NLP based
  - » Language parsing
  - » Syntactic approach
- Ontology based
  - » Uses formal logic representations
- Deep Learning based
  - » Deep neural architectures & ML

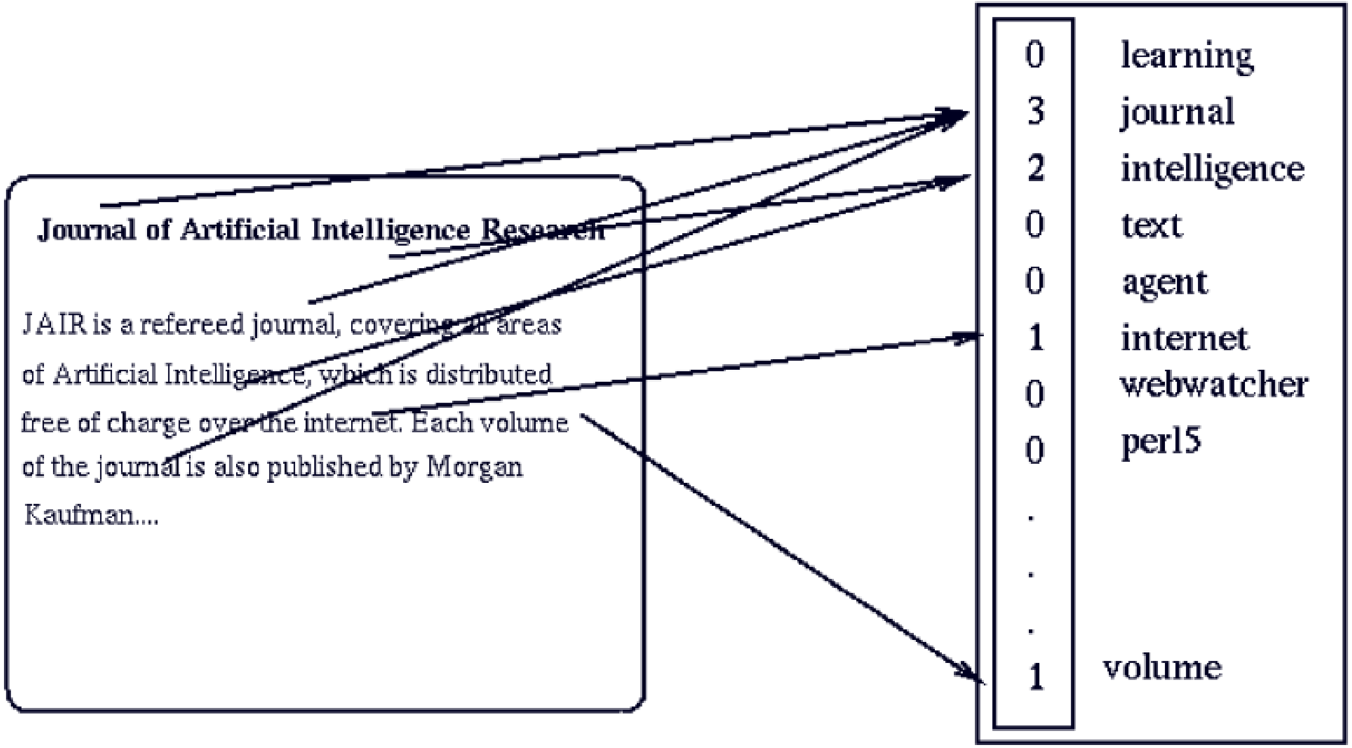


Decreasing Semantic Gap
Increasing Accuracy

Increasing Complexity
Decreasing Scalability

Keywords-based Approach

Bag-of-words document representation



Journal of Artificial Intelligence Research

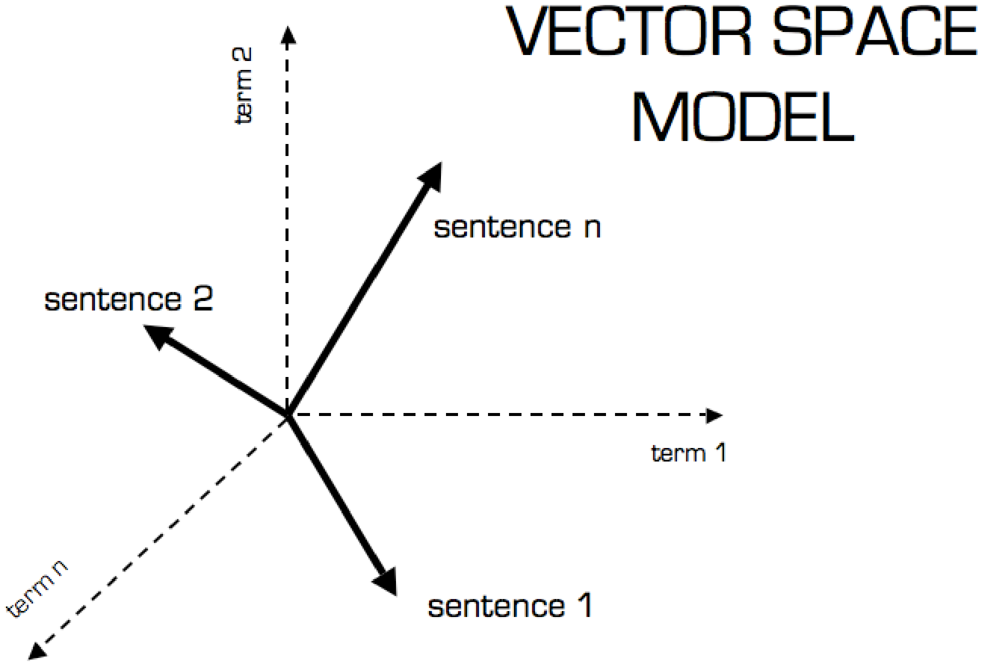
JAIR is a refereed journal, covering all areas of Artificial Intelligence, which is distributed free of charge over the internet. Each volume of the journal is also published by Morgan Kaufman....

0	learning
3	journal
2	intelligence
0	text
0	agent
1	internet
0	webwatcher
0	perl5
.	
.	
.	
1	volume

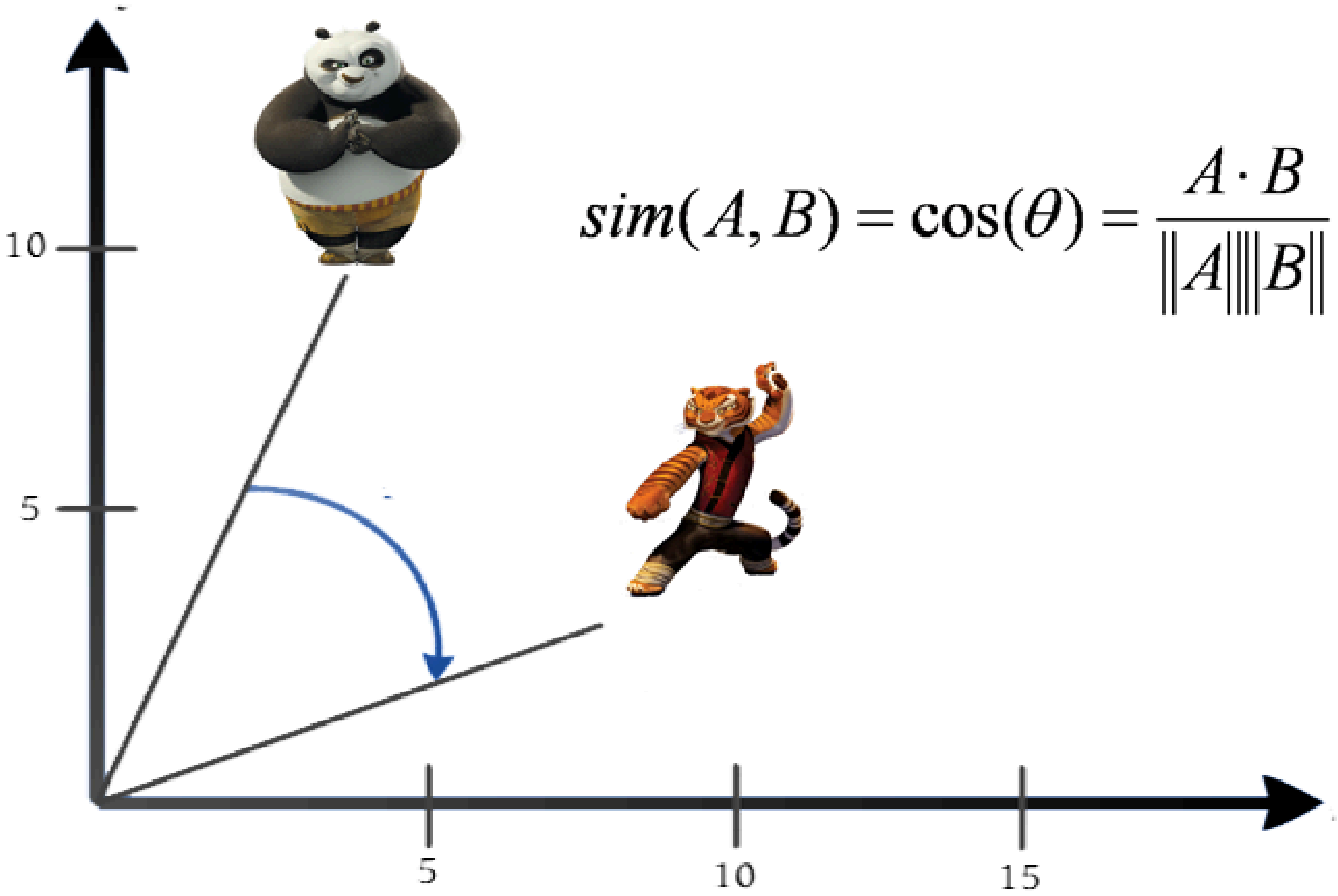
Bag of words Representation

- Texts are treated as a “bag” of words or terms

- Any document can be represented as a vector: a list of **terms** and their associated **weights**
  - » D= {(t_1,w_1),(t_2,w_2),…………
  - » t_i: i-th term
  - » w_i: weight for the i-th term



Cosine Similarity



$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Problems with Keywords

- Bag of words approach
  - Is both the **advantage** and **disadvantage**
- Often results in **huge semantic gap**
  - A positive or negative sentiment word may have opposite orientations in different application domains. (“*This camera sucks*.” -> negative; “*This vacuum cleaner really sucks*.” -> positive)
  - A sentence containing sentiment words may not express any sentiment. (e.g. “*Can you tell me which Sony camera is good*?”)
  - Sarcastic sentences with or without sentiment words are hard to deal with. (e.g. “*What a great car! It stopped working in two days.*”

NLP-based Approach

Natural Language Processing (NLP)

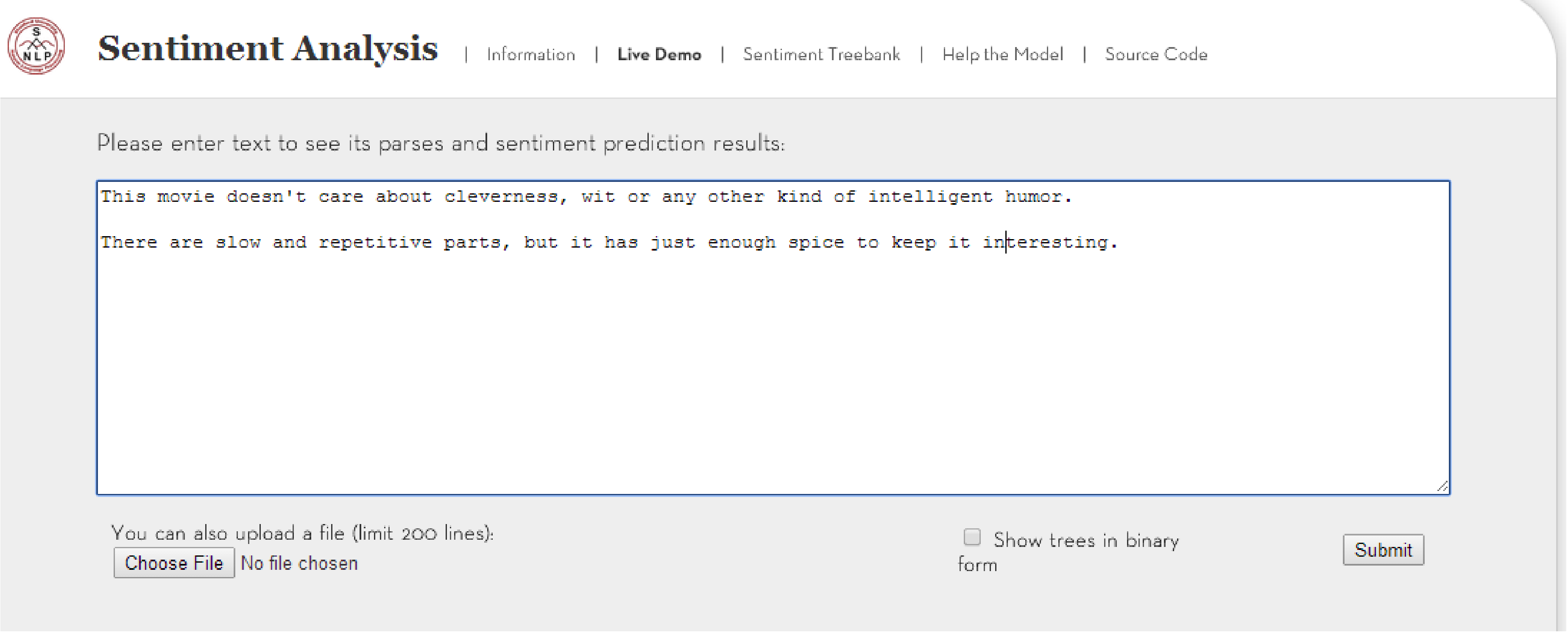
- **<u>NLP</u>**: is a field of computer science, artificial intelligence, and linguistics, concerned with the interactions between computers and human (natural) languages.
- Major NLP applications
  - » Part-of-speech tagging (POS tagging)
  - » Relationship extraction
  - » Sentiment analysis
  - » Topic segmentation and recognition
  - » Machine translation

Shallow NLP

- Part-of-Speech (POS)
  - » Identify Nouns, Verbs, Tenses, Prepositions, etc.
- Morphology
  - » Do stemming the right way
- Syntax
  - » Extract sentence structure

Demonstration: Sentence-level Sentiment – 1/3

- Stanford Sentiment Analyzer
  - » http://nlp.stanford.edu:8080/sentiment/rntnDemo.html


Sentiment Analysis | Information | Live Demo | Sentiment Treebank | Help the Model | Source Code

Please enter text to see its parses and sentiment prediction results:

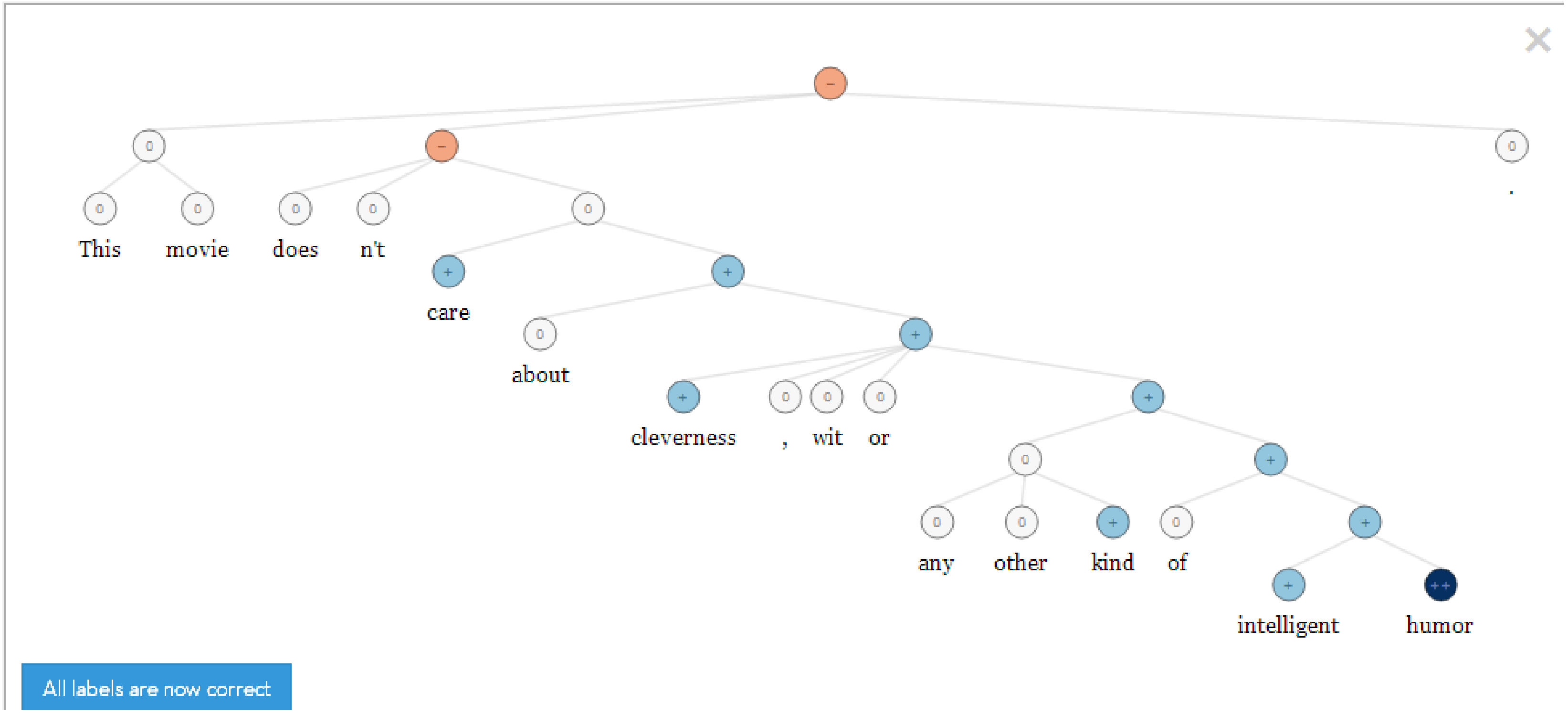
```
This movie doesn't care about cleverness, wit or any other kind of intelligent humor.

There are slow and repetitive parts, but it has just enough spice to keep it interesting.
```

You can also upload a file (limit 200 lines):

Choose File
No file chosen
□ Show trees in binary form
Submit

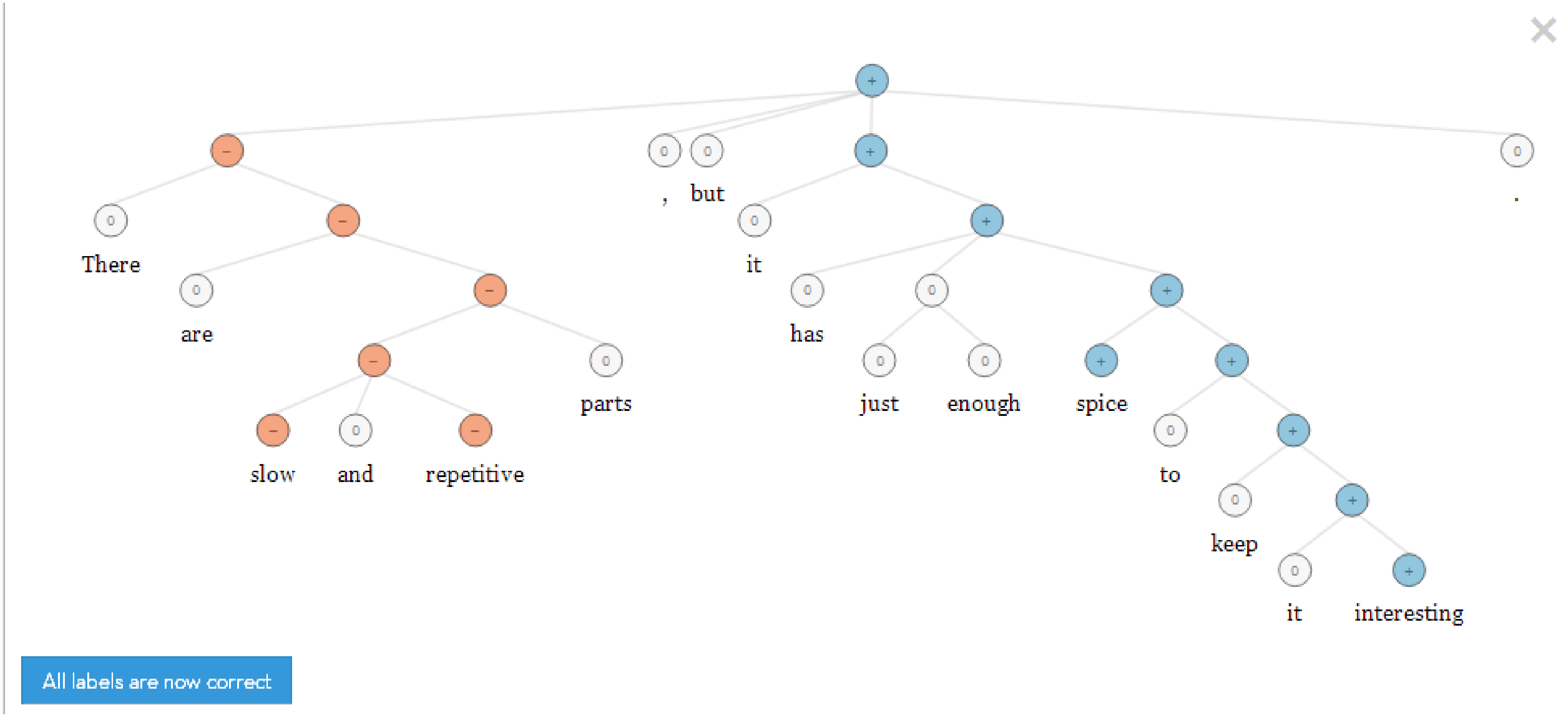
Demonstration: Sentence-level Sentiment – 2/3

- Review 1: This movie doesn't care about cleverness, wit or any other kind of intelligent humor. -> **Negative**



Demonstration: Sentence-level Sentiment – 3/3

- There are slow and repetitive parts, but it has just enough spice to keep it interesting. -> **Positive**



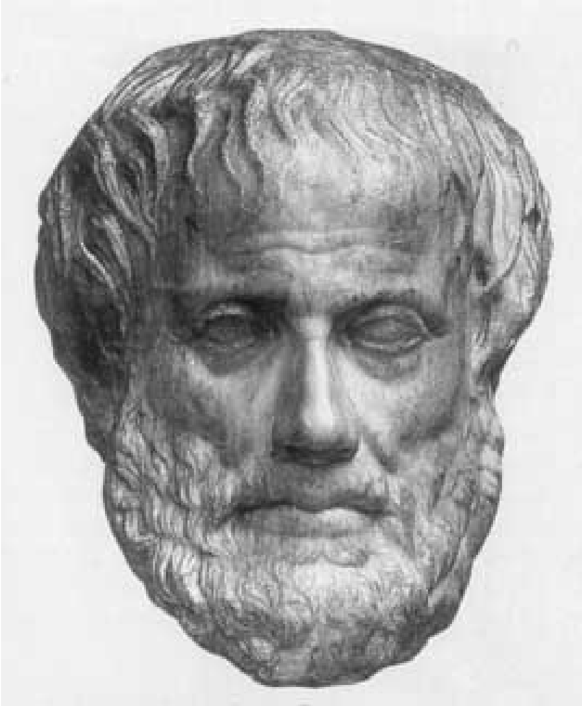
Problems with NLP

- Better than Bag-of-words but still not the best
  - » synonymy and polysemy not entirely overcome
  - » Basically, a syntactic approach
- Examples
  - » "I studied in Cambridge" (Which Cambridge?)
  - » "I live in Singapore" (Do you live in Asia?)

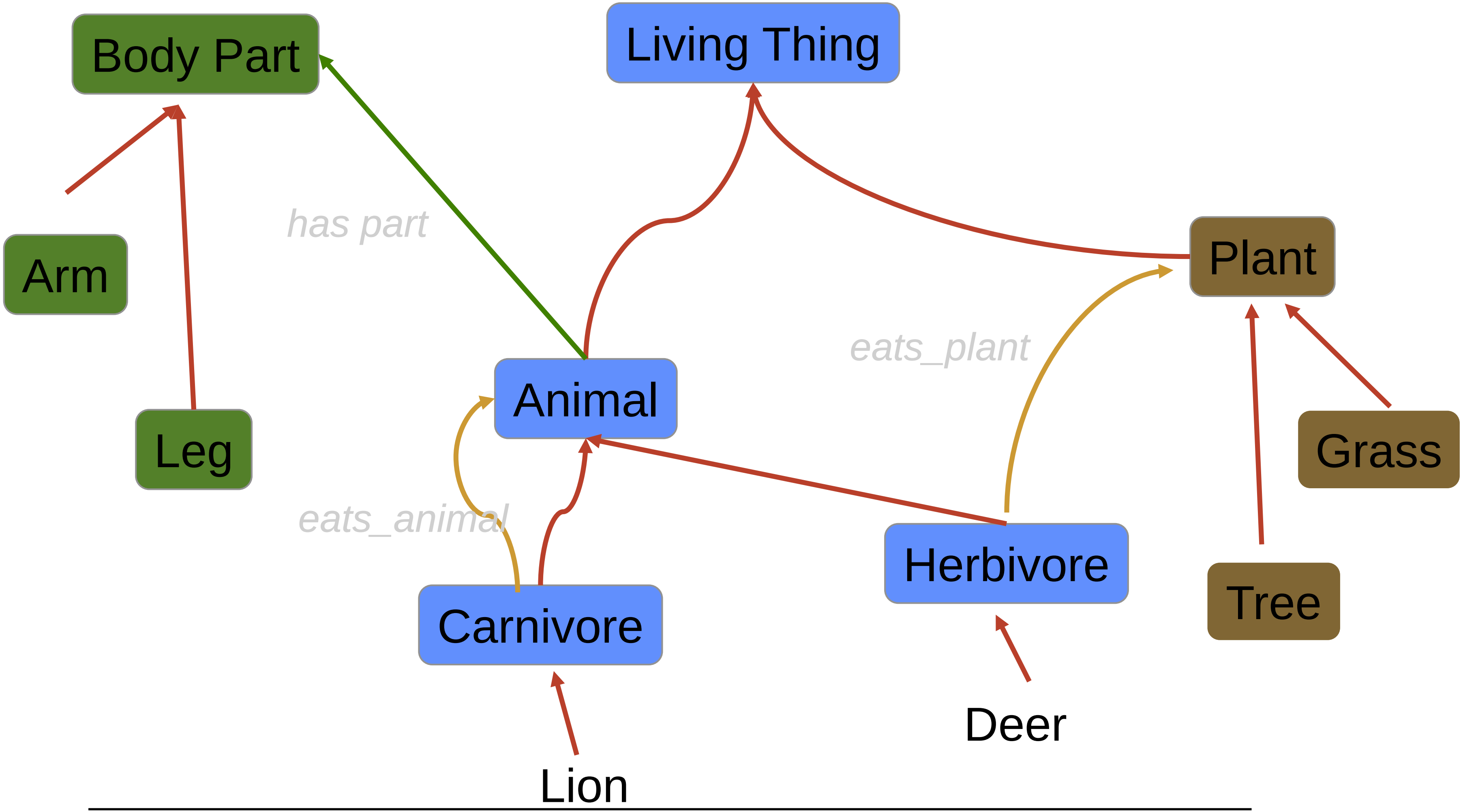
Ontology-based Approach

What Is An Ontology?

- Ontology (Socrates & Aristotle 400-360 BC)
- Word borrowed by computing for the explicit description of the conceptualisation of a domain:
  - » concepts
  - » properties and attributes of concepts
  - » constraints on properties and attributes
  - » Individuals (often, but not always)
- An ontology defines
  - » a common vocabulary
  - » a shared understanding
- E.g.: Wiki/Yahoo categories, WordNet
- Backbone of semantic web



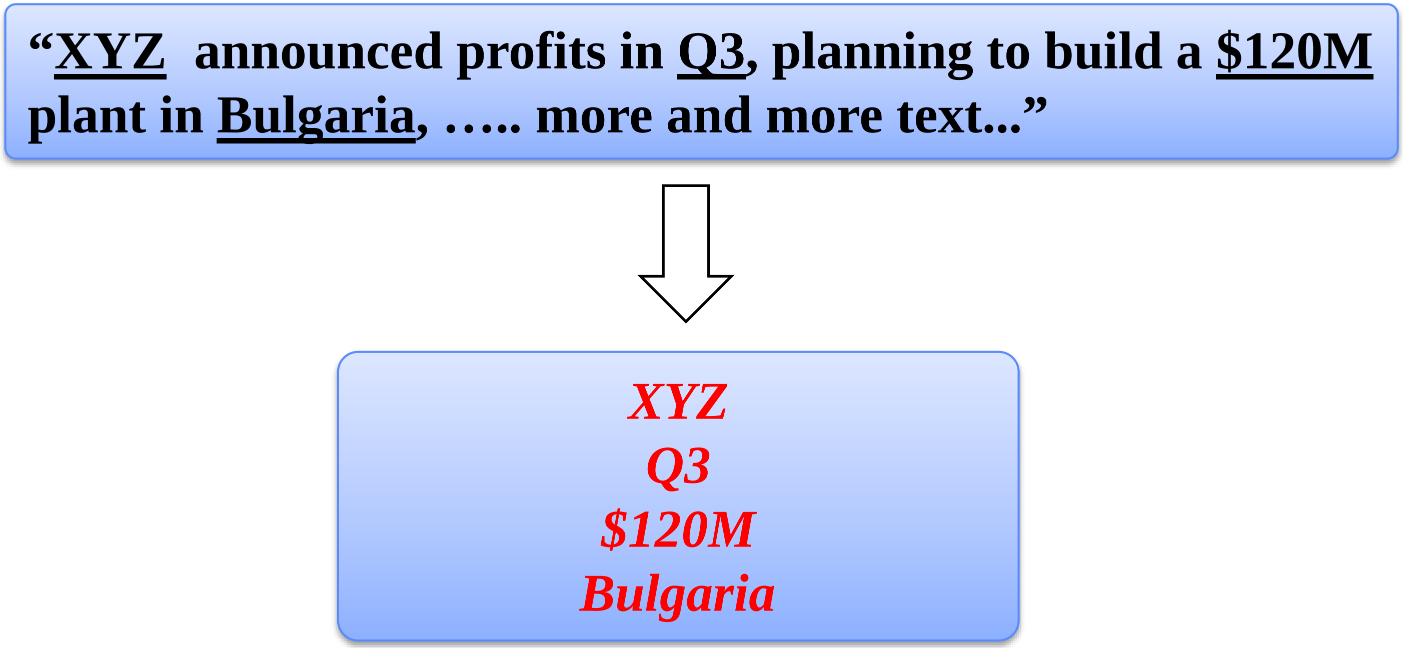
A simple ontology: Animals



How is it useful for Text Mining?

“XYZ announced profits in Q3, planning to build a $120M plant in Bulgaria, ….. more and more text...”

Information Extraction



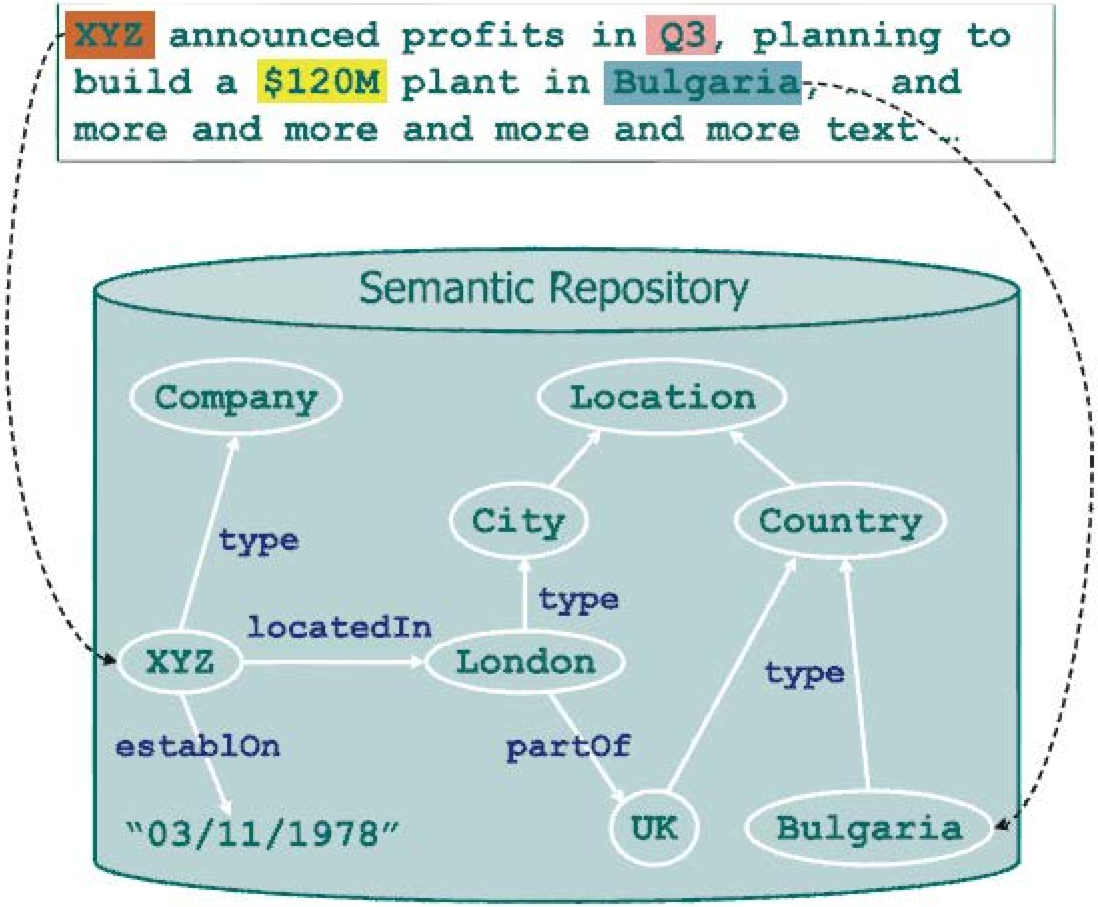
"XYZ announced profits in Q3, planning to build a $120M plant in Bulgaria, ….. more and more text..."

XYZ
Q3
$120M
Bulgaria

Loosely structured.. "Semantic Gaps"

OK for simple scenarios, but messy in larger applications!

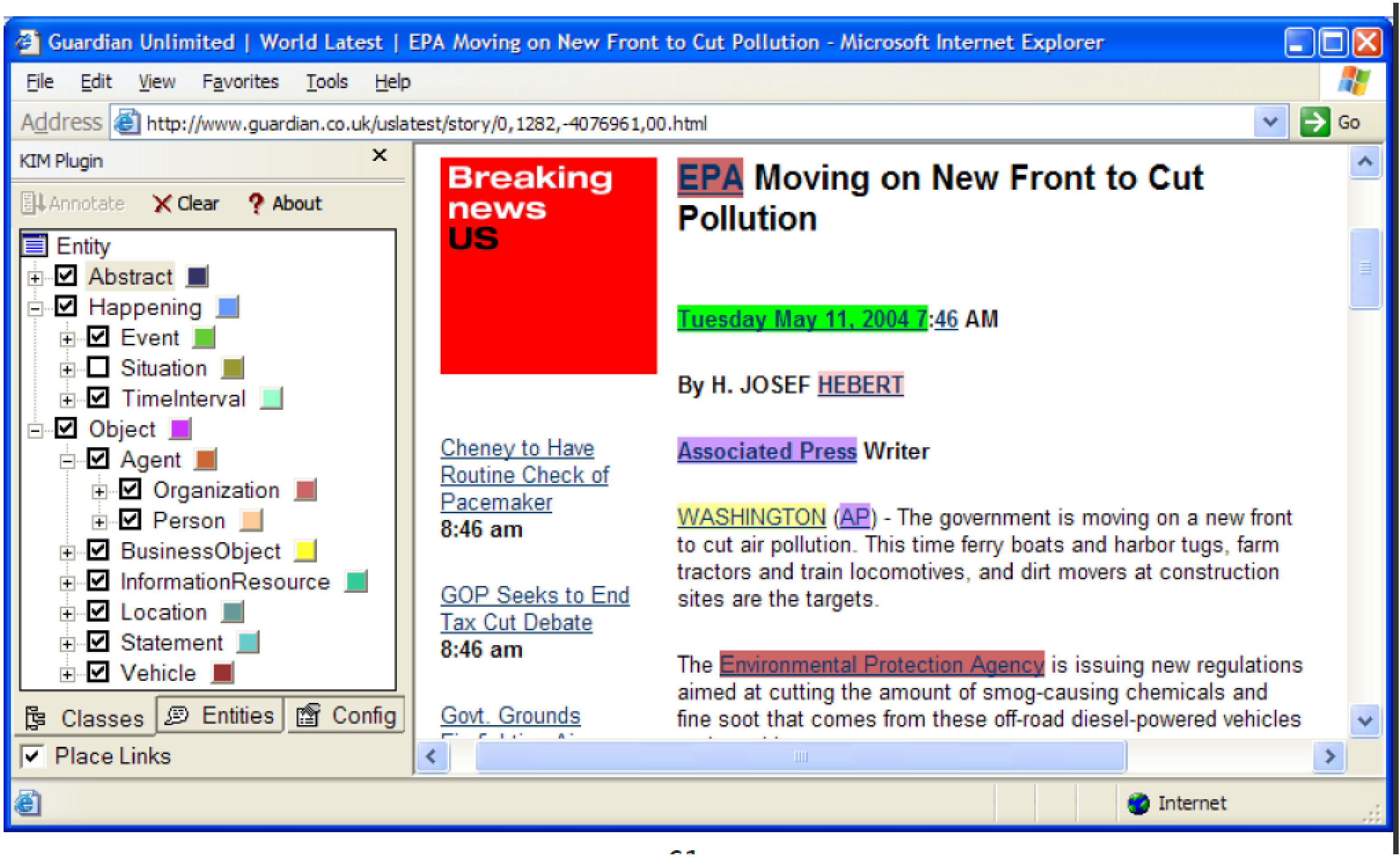
With Proper Semantics


XYZ announced profits in Q3, planning to build a $120M plant in Bulgaria, ... and more and more and more and more text ...
Semantic Repository
Company
Location
City
Country
type
type
locatedIn
XYZ
London
type
establOn
partOf
"03/11/1978"
UK
Bulgaria

What can we do with ontologies? (1/2)

- **Semantic Annotation**: represent metadata/keywords with proper semantics
  - » to represent ‘Cambridge’ as a UK location, link it to an ontology instance ‘Cambridge, UK’ rather than ‘Cambridge, Massachusetts’
  - » Link synonyms to the ontology , e.g. link ‘heart attack’ to ‘myocardial infarction’

Example



http://www.guardian.co.uk/uslatest/story/0,1282,-4076961,00.html

Cheney to Have Routine Check of Pacemaker
8:46 am

GOP Seeks to End Tax Cut Debate
8:46 am

Govt. Grounds

EPA Moving on New Front to Cut Pollution

Tuesday May 11, 2004 7:46 AM

By H. JOSEF HEBERT

Associated Press Writer

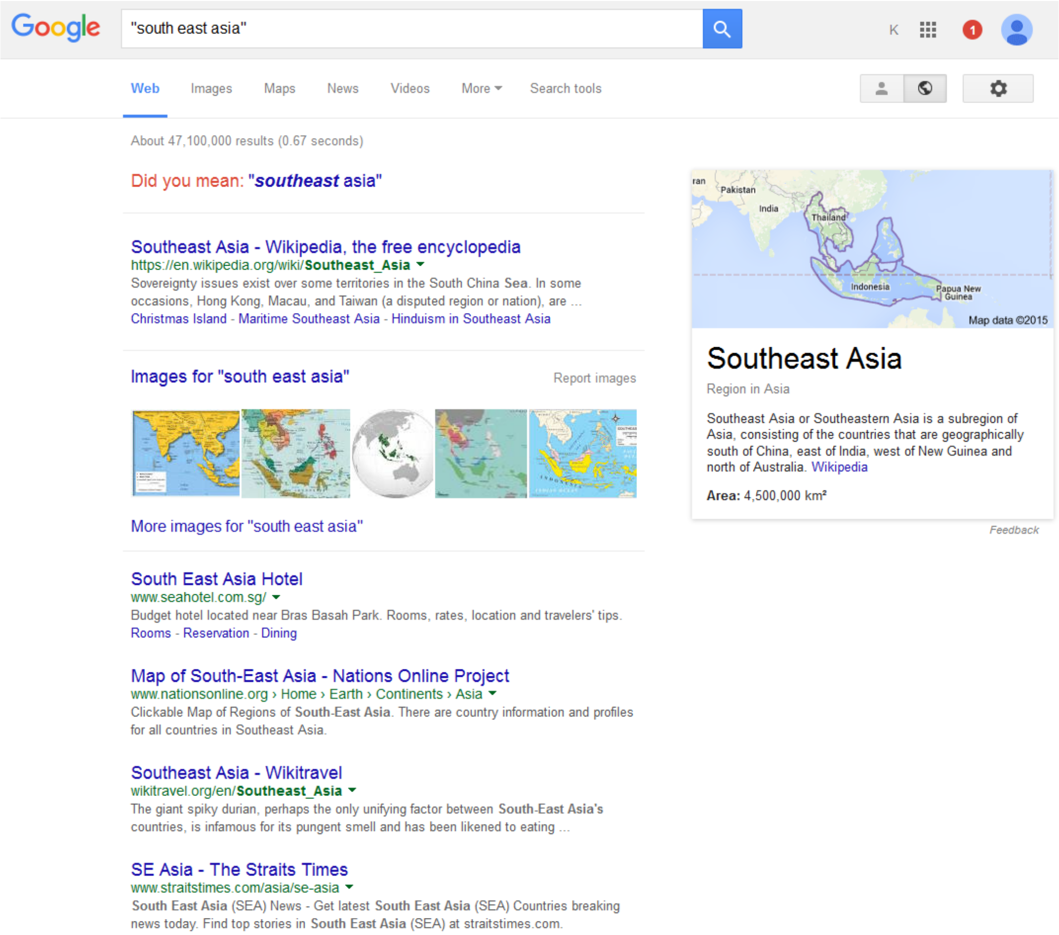
WASHINGTON (AP) - The government is moving on a new front to cut air pollution. This time ferry boats and harbor tugs, farm tractors and train locomotives, and dirt movers at construction sites are the targets.

The Environmental Protection Agency is issuing new regulations aimed at cutting the amount of smog-causing chemicals and fine soot that comes from these off-road diesel-powered vehicles

What can we do with ontologies? (2/2)

- **Semantic Search**: Use proper semantics captured in ontologies to retrieve more relevant results
  - » search for ‘heart attack’ retrieves documents containing ‘myocardial infarction’
  - » search for ‘South East Asia’ retrieves documents that contain ‘Singapore’ or ‘Thailand’

Google’s Semantic Search



"south east asia"

About 47,100,000 results (0.67 seconds)

Did you mean: "***southeast*** asia"

Southeast Asia - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Southeast_Asia**
Sovereignty issues exist over some territories in the South China Sea. In some occasions, Hong Kong, Macau, and Taiwan (a disputed region or nation), are ...
Christmas Island - Maritime Southeast Asia - Hinduism in Southeast Asia

Images for "south east asia" Report images

More images for "south east asia"

South East Asia Hotel
www.seahotel.com.sg/
Budget hotel located near Bras Basah Park. Rooms, rates, location and travelers' tips.
Rooms - Reservation - Dining

Map of South-East Asia - Nations Online Project
www.nationsonline.org › Home › Earth › Continents › Asia
Clickable Map of Regions of **South-East Asia**. There are country information and profiles for all countries in Southeast Asia.

Southeast Asia - Wikitravel
wikitravel.org/en/**Southeast_Asia**
The giant spiky durian, perhaps the only unifying factor between **South-East Asia's** countries, is infamous for its pungent smell and has been likened to eating ...

SE Asia - The Straits Times
www.straitstimes.com/asia/se-asia
South East Asia (SEA) News - Get latest **South East Asia** (SEA) Countries breaking news today. Find top stories in **South East Asia** (SEA) at straitstimes.com.

Map data ©2015

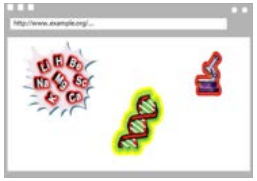
Southeast Asia

Region in Asia

Southeast Asia or Southeastern Asia is a subregion of Asia, consisting of the countries that are geographically south of China, east of India, west of New Guinea and north of Australia. Wikipedia

Area: 4,500,000 km²

Semantic Analytics Framework



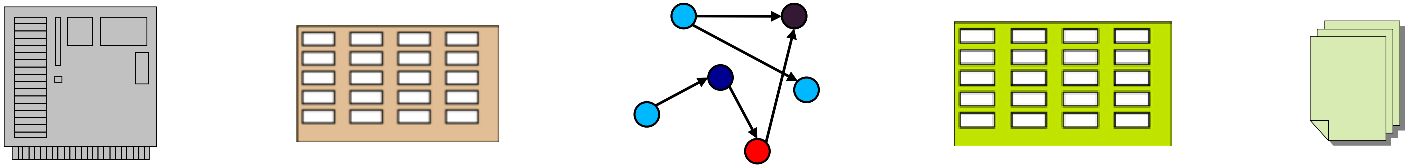
something said here on the bioinformatics data which is then described in human terms if somebody can read that, which is probably not the case...

Applications

Manipulate Query …

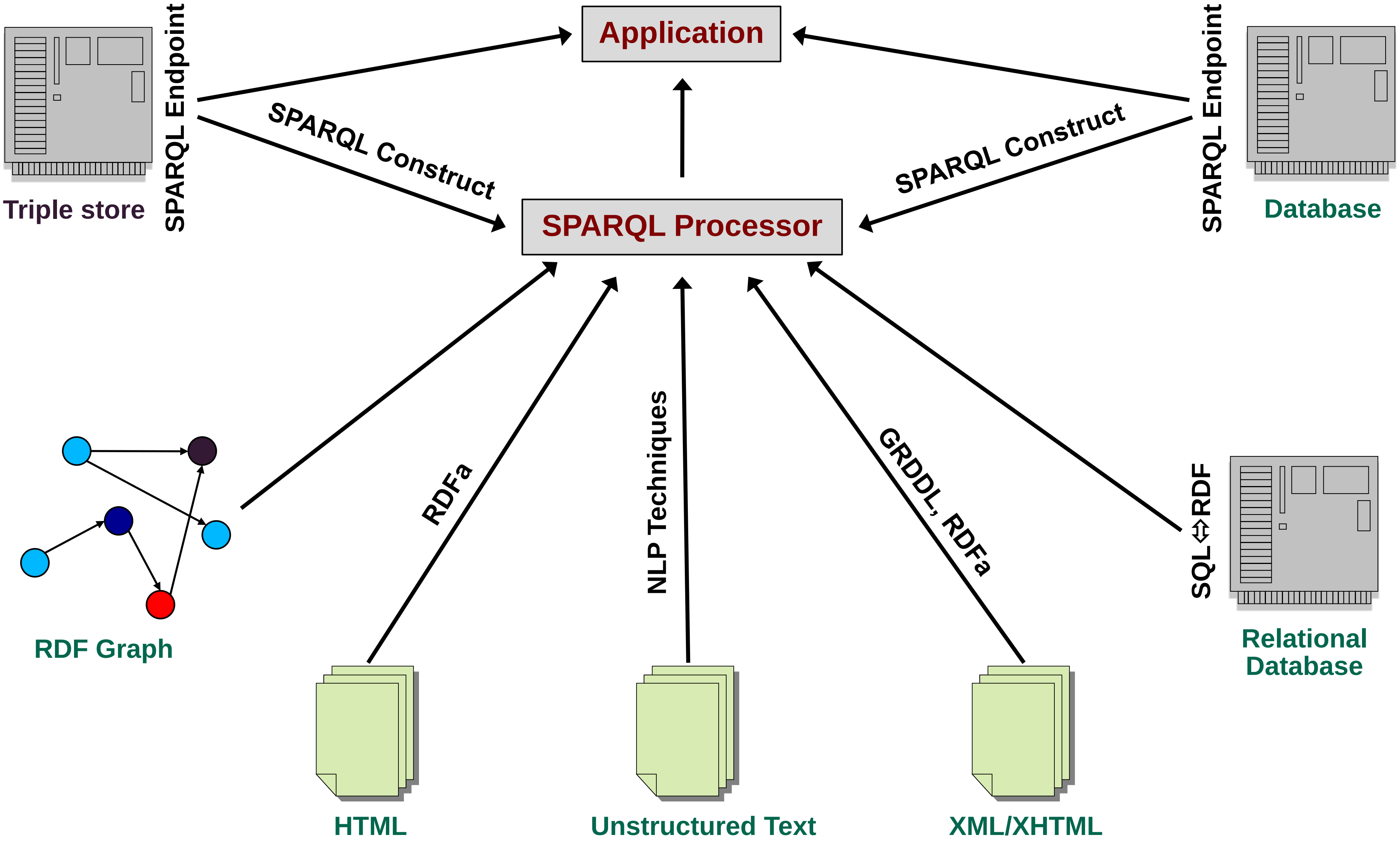
Data represented in abstract format

Map, Expose, …



Data in various formats

Data Linking & Integration



Summary

- Addressing the Semantic Gap is the key to effective Text Mining
- Keyword approach leads to a huge semantic gap, and hence often a big limitation in real life
- NLP addresses it somewhat but is still a syntactic approach
- Knowledge representations, as modeled by Ontologies, can bridge the Semantic gap but require a lot of manual effort in ontology construction