IBM Research

Semantic Concept Discovery Over Event Data

Oktie Hassanzadeh, Shari Trewin, Alfio Gliozzo
IBM Research AI

IBM

Use Case: Question Analysis

- High-Level Goal:

> *Given a question (set of entities & concepts), find the most relevant entities & concepts needed to generate a high-quality analysis report.*

- Example question:
  "What are the consequences of Brexit on London's financial markets?"

  Need to discover:
  - Key topics (e.g., financial markets, economy, Brexit, Brexit Divorce Bill)
  - Key people and organizations involved (e.g., The European Union, decision makers in the EU & UK, people involved in Brexit negotiations)
  - All the related events (e.g., Negotiation meetings, Parliamentary elections within the EU, etc.)

Concept Discovery for Deep Analysis

Goal: *Given an analysis question, find the most relevant concepts (topics, entities) needed to make a comprehensive and unbiased analysis*

Classic Solution:

Search over news articles, social media, past reports, other media (text, photo & videos)

Identify relevant concepts mentioned in unstructured sources to establish context

Perform Analysis tasks such as forecasting, hypothesis generation, and scenario analysis, given the context

Problem #1: Mostly manual process, and massive amount of information - so the outcome could be ***biased*** & ***incomplete***

Problem #2: Sources of **search** do not match the available structured data and the corresponding models used for **analytics**

Event Databases

- Structured data representing ”events” as reported on the media

GDELT https://www.gdeltproject.org/

ICEWS http://www.lockheedmartin.com/us/products/W-ICEWS.html

EventRegistry http://eventregistry.org/

Event Databases

- Structured data representing ”events” as reported on the media

GDELT

ICEWS

Political event databases:
“Event” is an action associated with up to two actors

EventRegistry → Generic event database:
“Event” is a collection of articles on the same topic

Event Databases

- Structured data representing ”events” as reported on the media

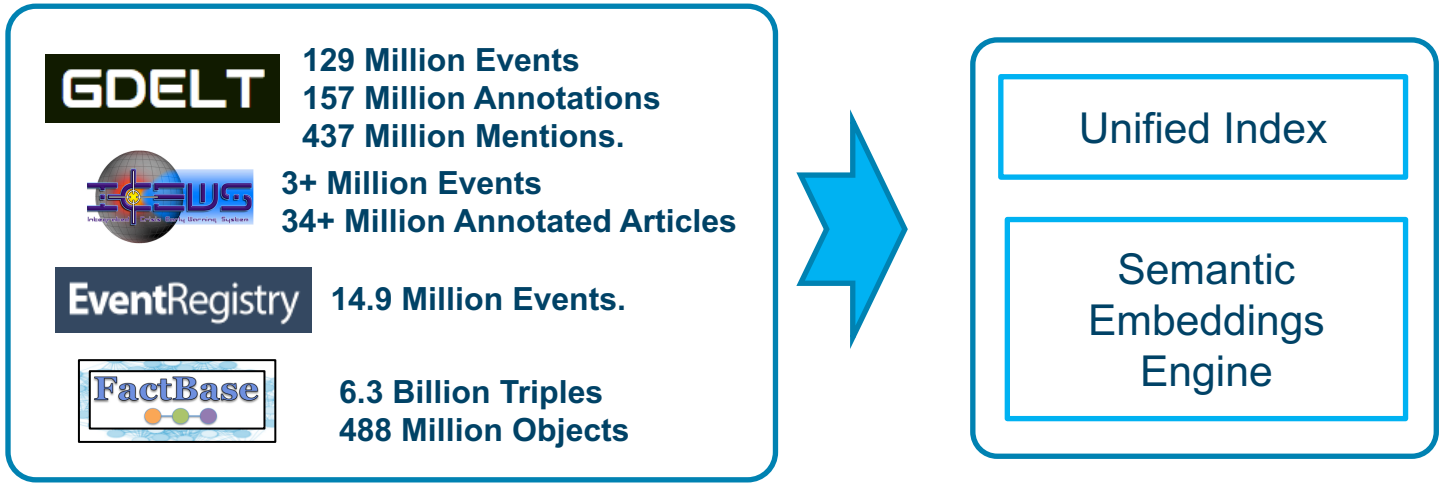
GDELT 2.0 → **129+ Million Events**
157+ Million Articles Annotated (GKG)
437+ Million Mentions (2+ years)

ICEWS → **14.9 Million Events (1995 to 2015)**

EventRegistry → **5+ Million Events**
180+ Million Annotated Articles (2+ years)

Concept Discovery for Deep Analysis: Our Solution

Semantic Data Curation Engine:

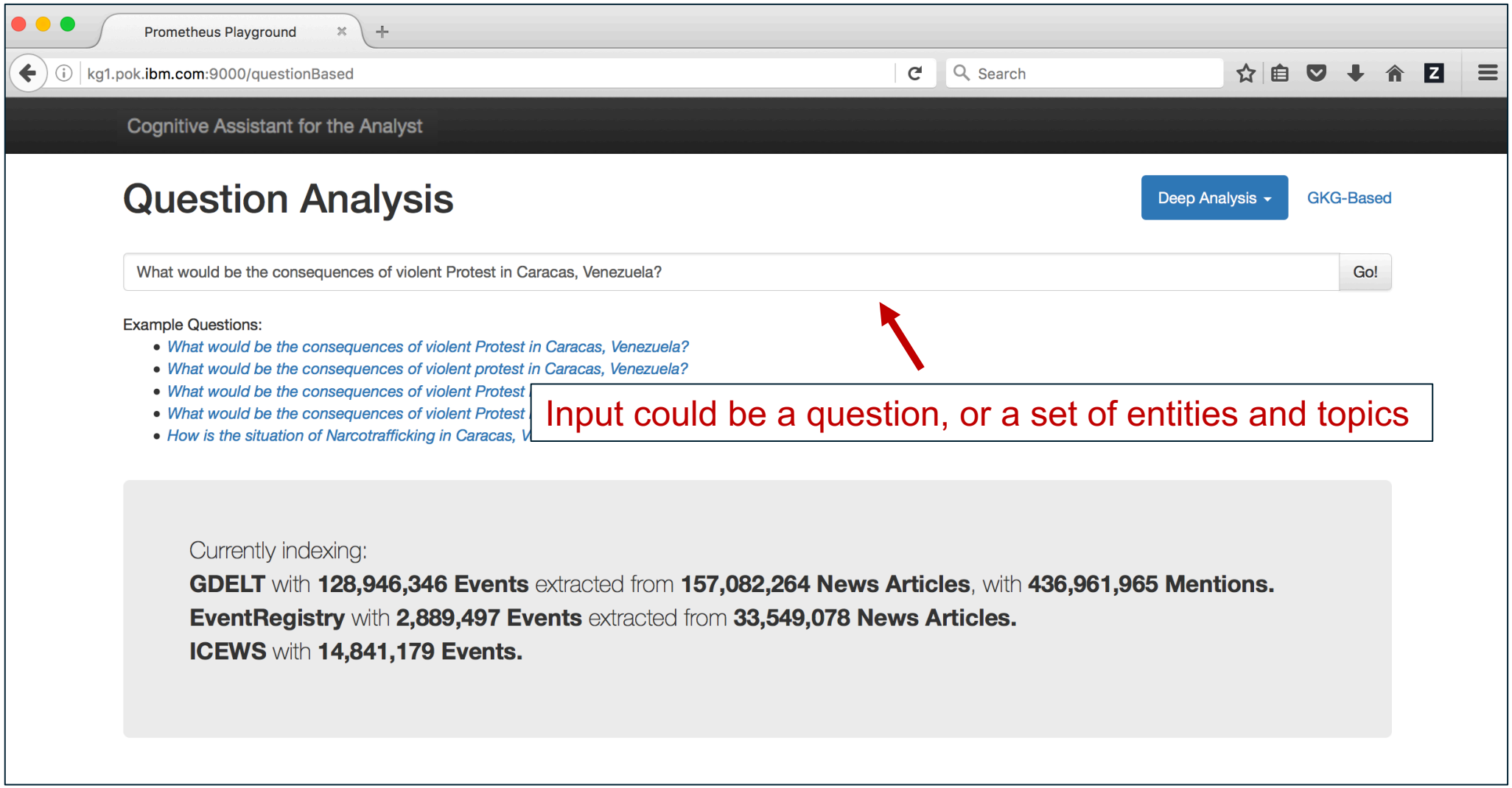


Analysis Question: *What is the likelihood of violent protest in Caracas, Venezuela?*

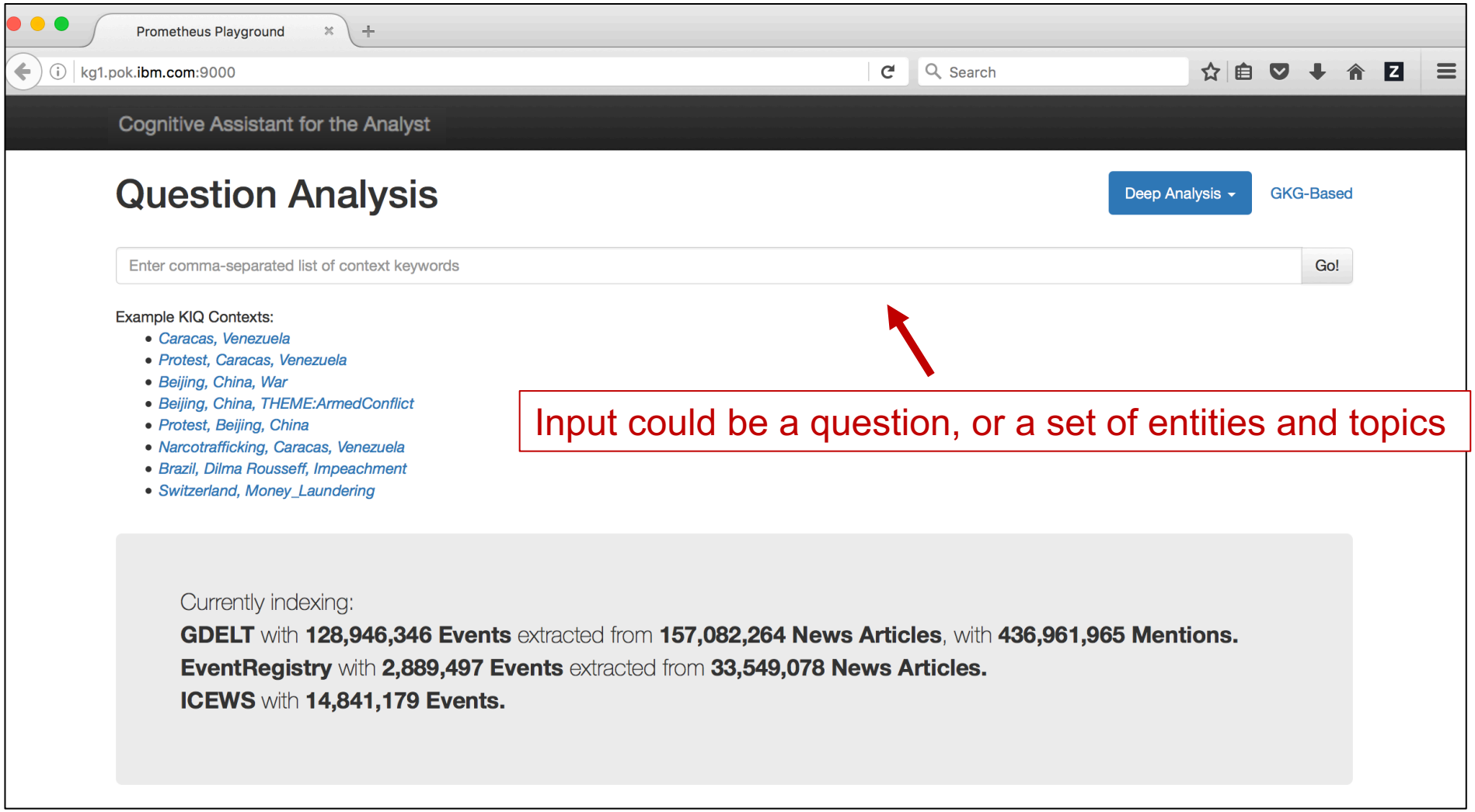
Natural Language Question Understanding: Topic: *Violent Protest* Location: Caracas, Venezuela

Semantic Concept Discovery: *Query DeepSim API with* **Topic**#Violent_Protest **Location**#Caracas_Venezuela

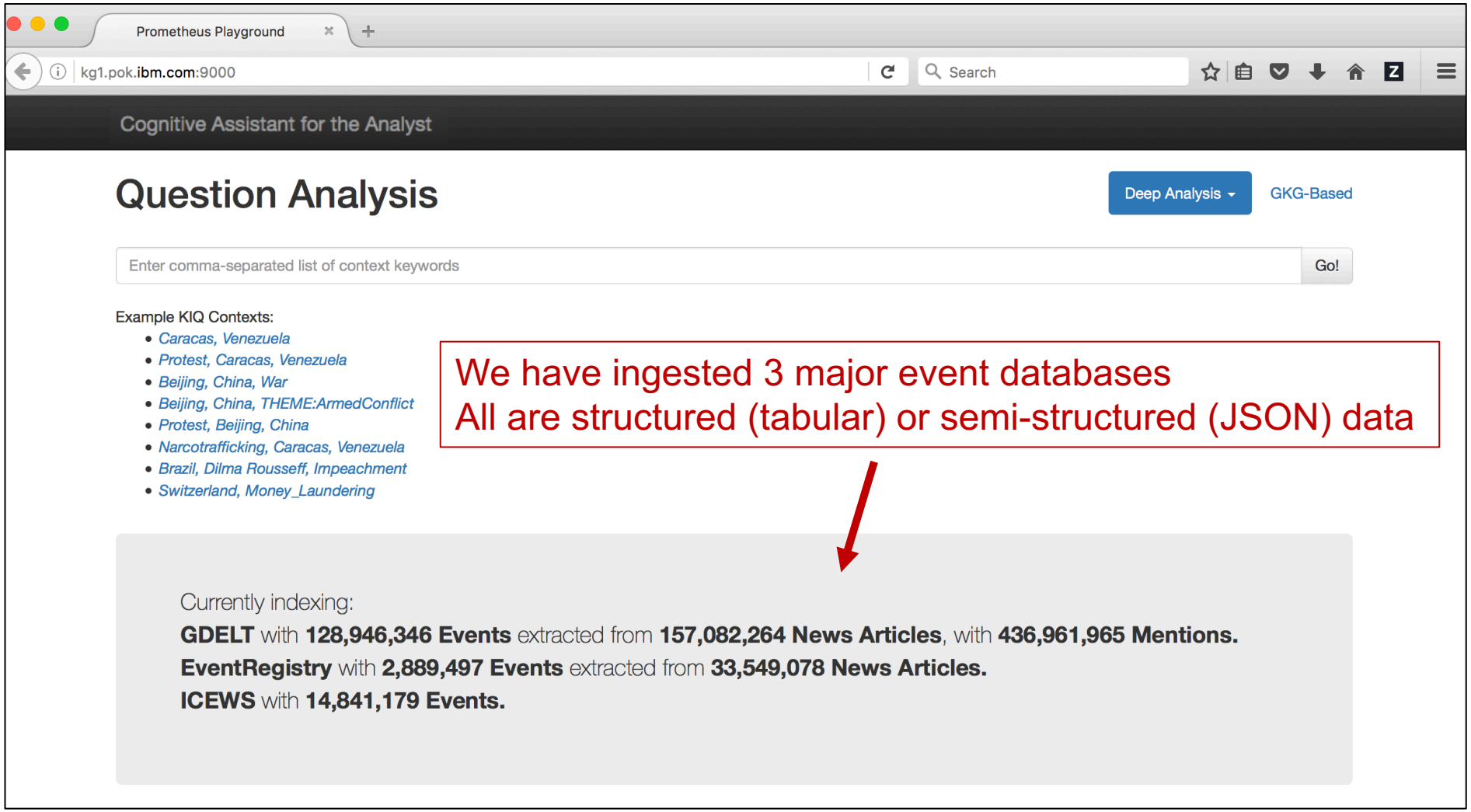
Question Analysis: Current Prototype



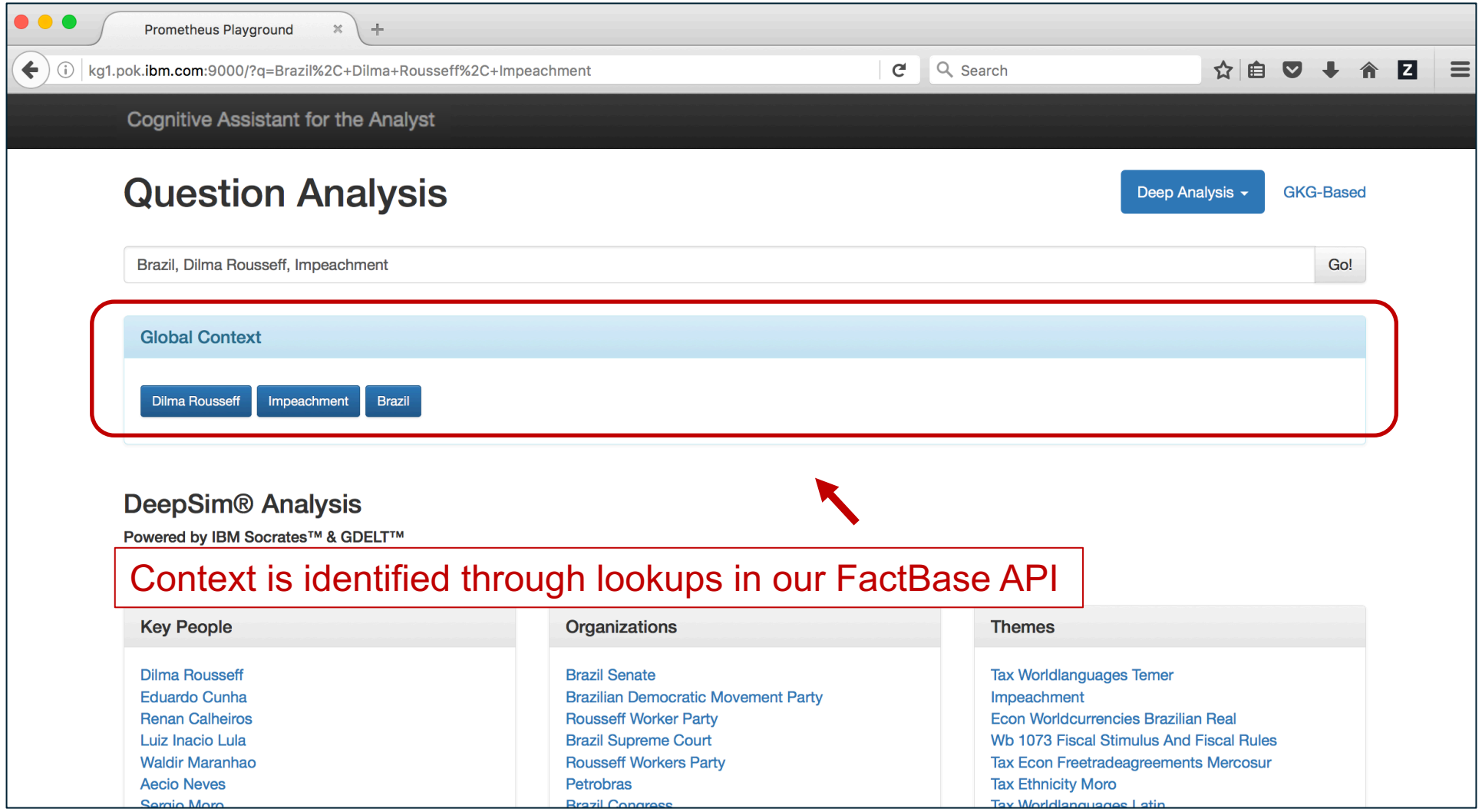
Question Analysis: Current Prototype



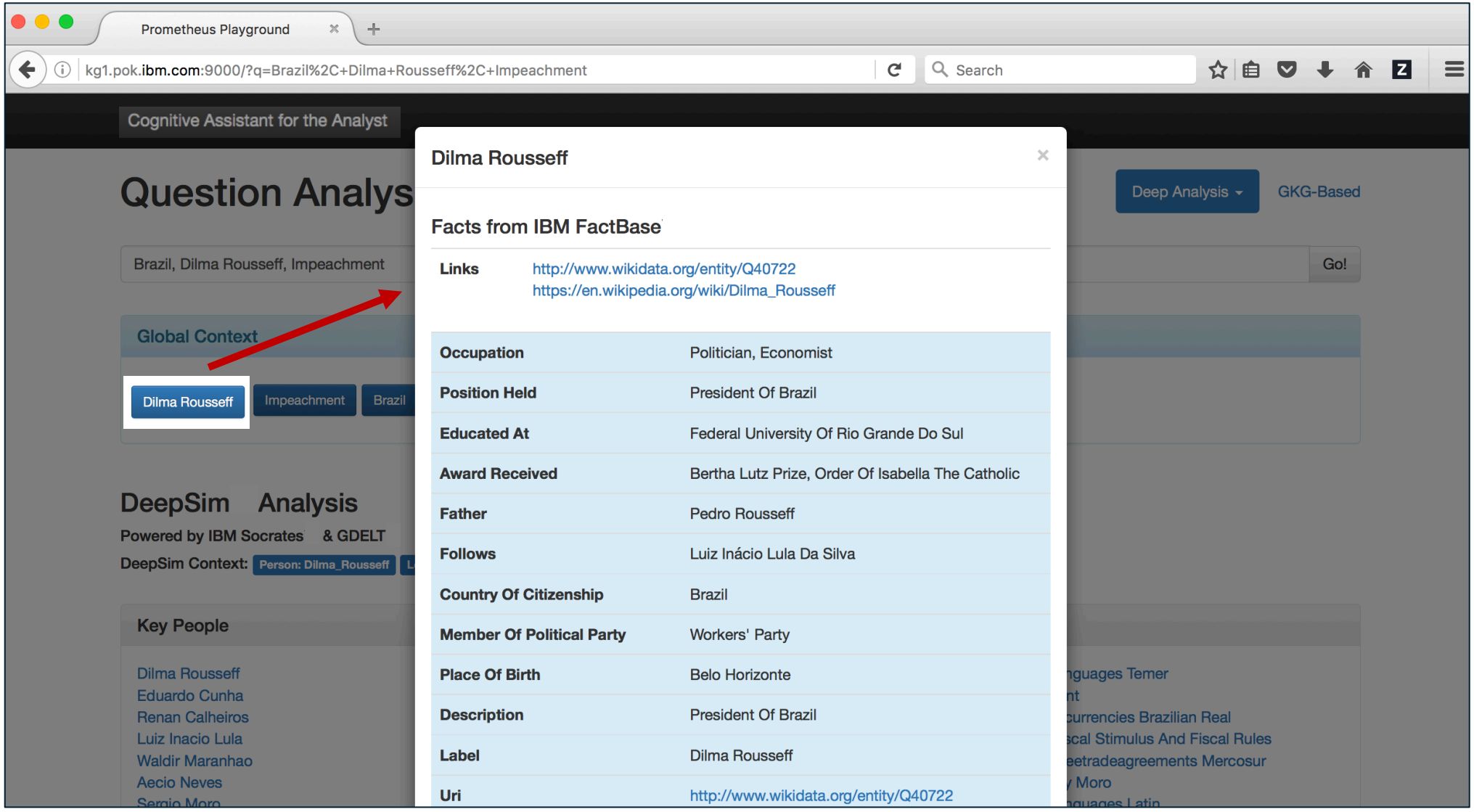
Question Analysis: Current Prototype



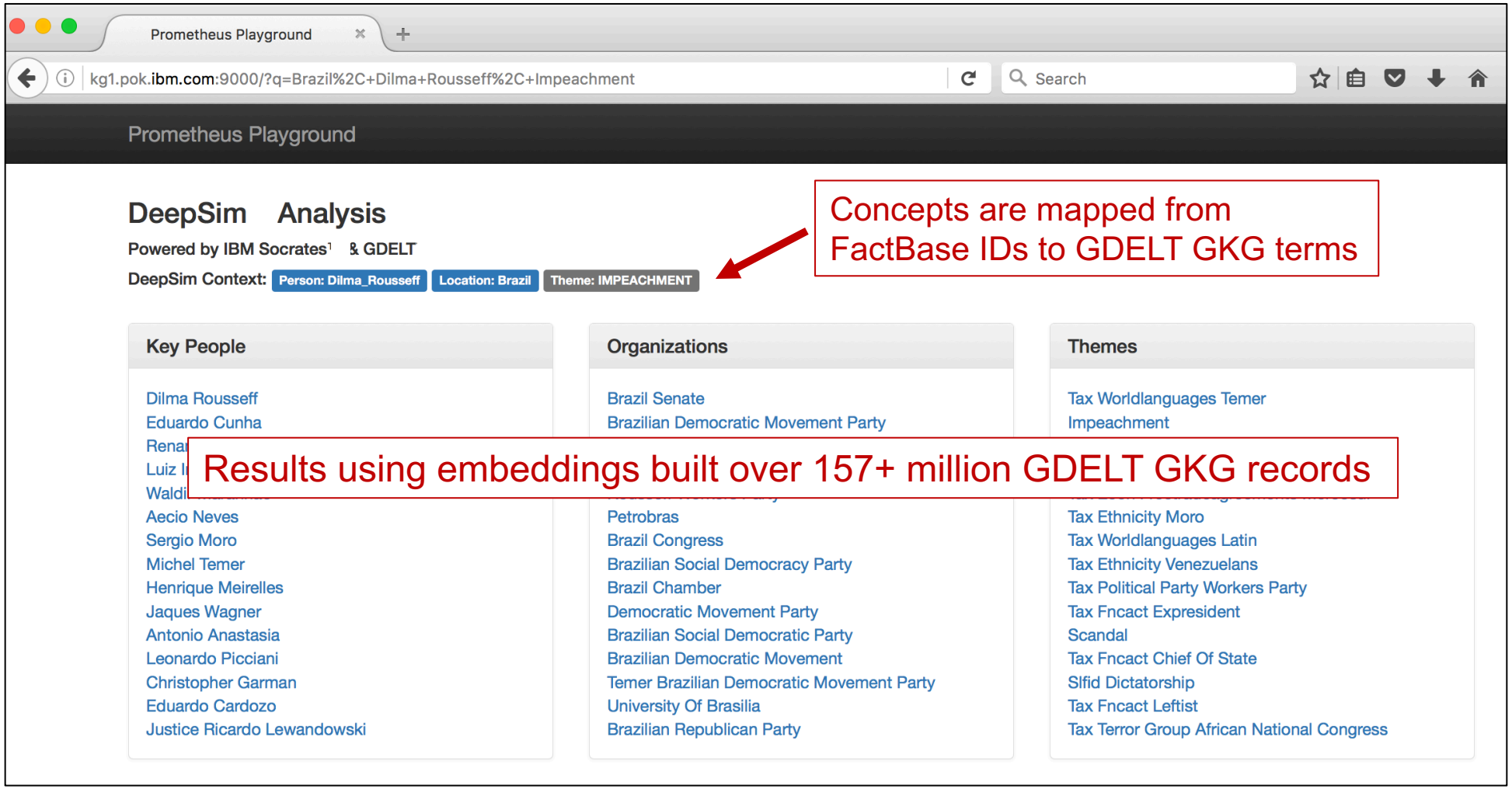
Question Understanding



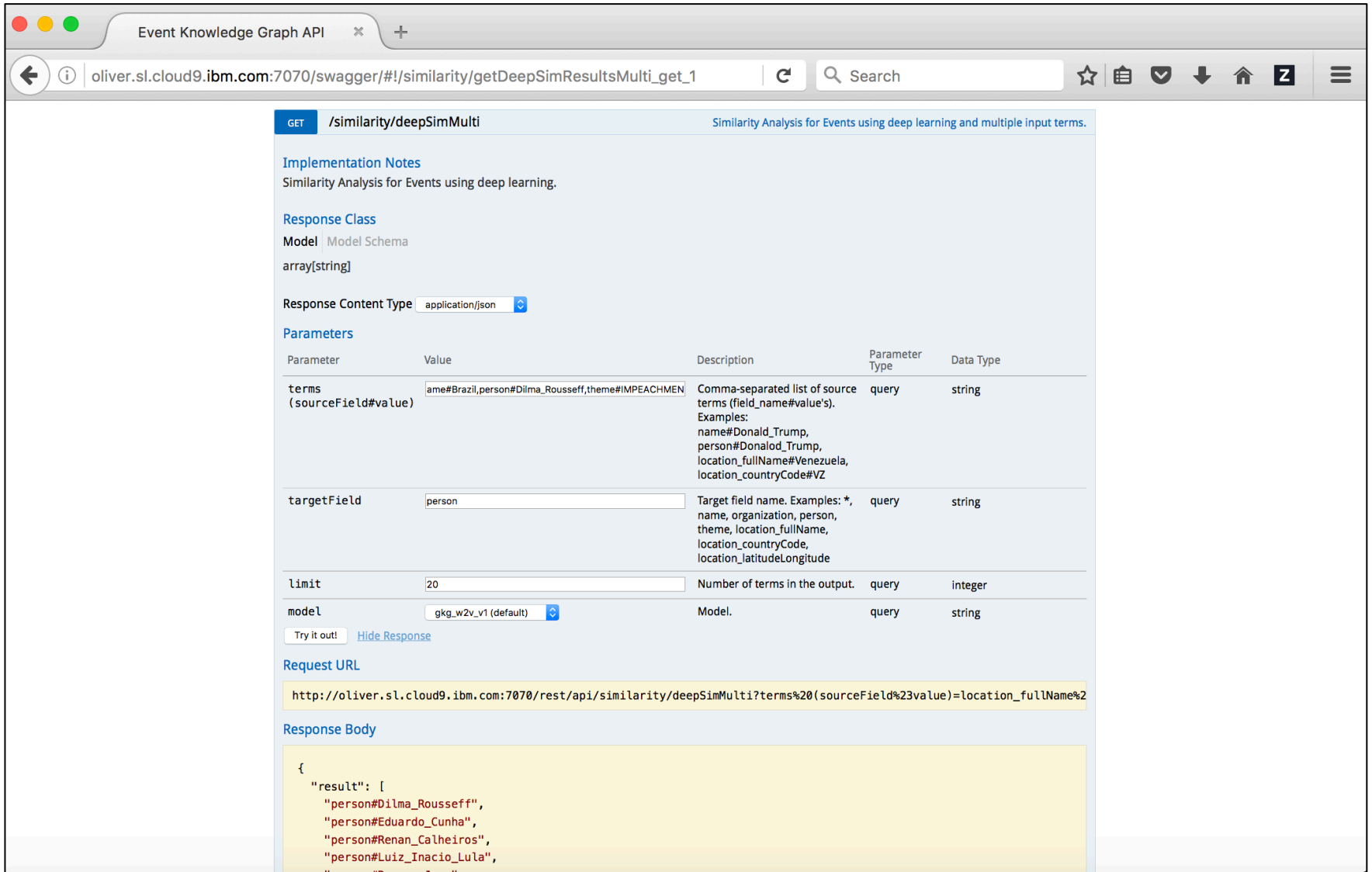
Question Understanding: Context Details



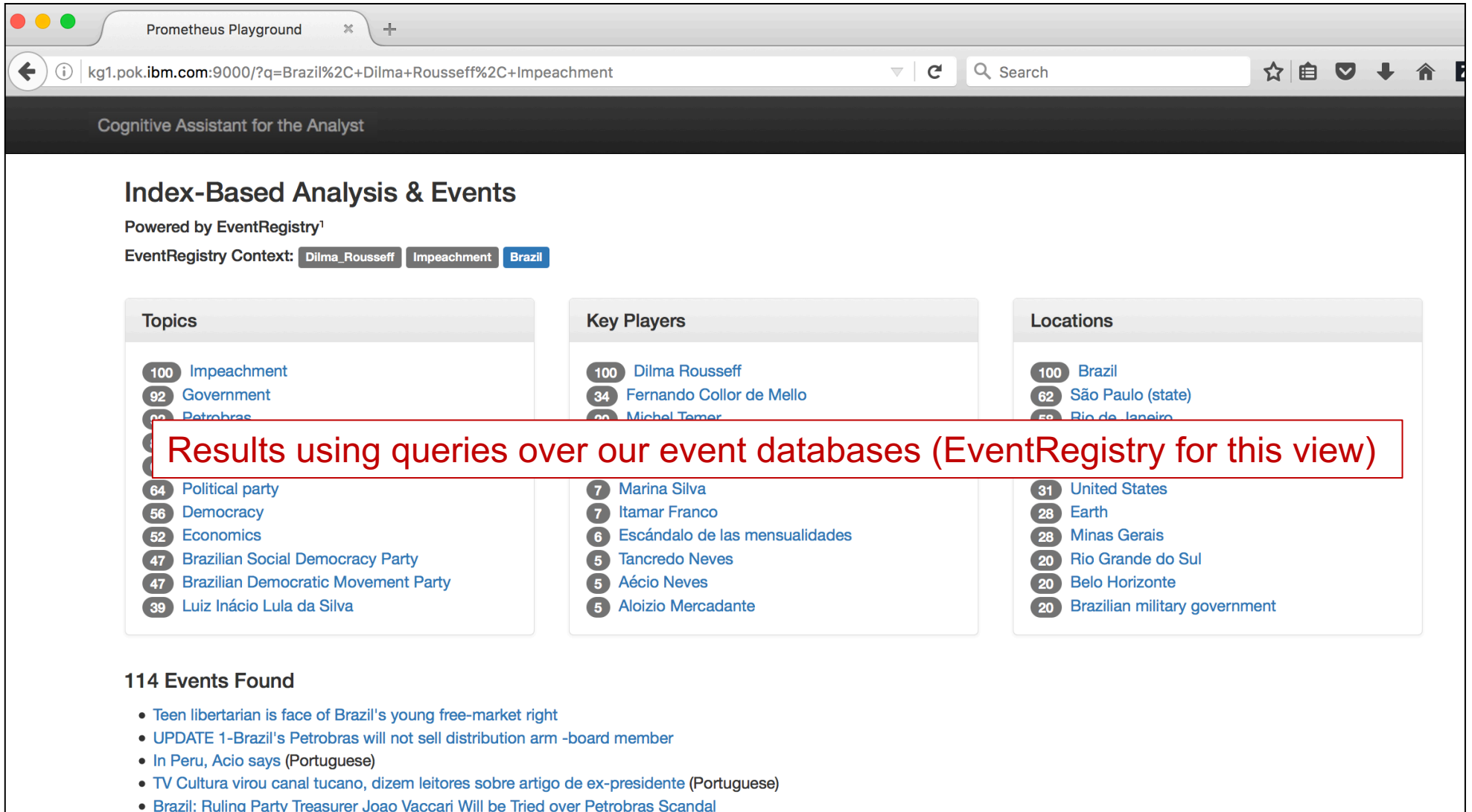
DeepSim Analysis – Using Semantic Similarity Anlysis over GDELT GKG



Swagger API behind DeepSim Analysis



Index-based Analysis (co-occurrence)



Swagger API behind Index-Based Analysis

concept-discovery

This is the concept discovery API. It provides a set of functions for expanding a query with related people, organizations and themes.

extract

Show/Hide | List Operations | Expand Operations

POST /extract/concepts — This api extracts concepts (person, organization, themese, etc.) from a given text string

query

Show/Hide | List Operations | Expand Operations

POST /query/expand

This api uses GKG data to find key players (people and organizations) for a given country, set of seed people, and set of GDELT themes (topics)

Example input: { "country": "Peru", "players":["Nicolas Maduro"], "topics":["IMMIGRATION","PROTEST"], "candidateRankingMethods":["similarity","facetcount"] }

Response Class (Status 200)

successful operation

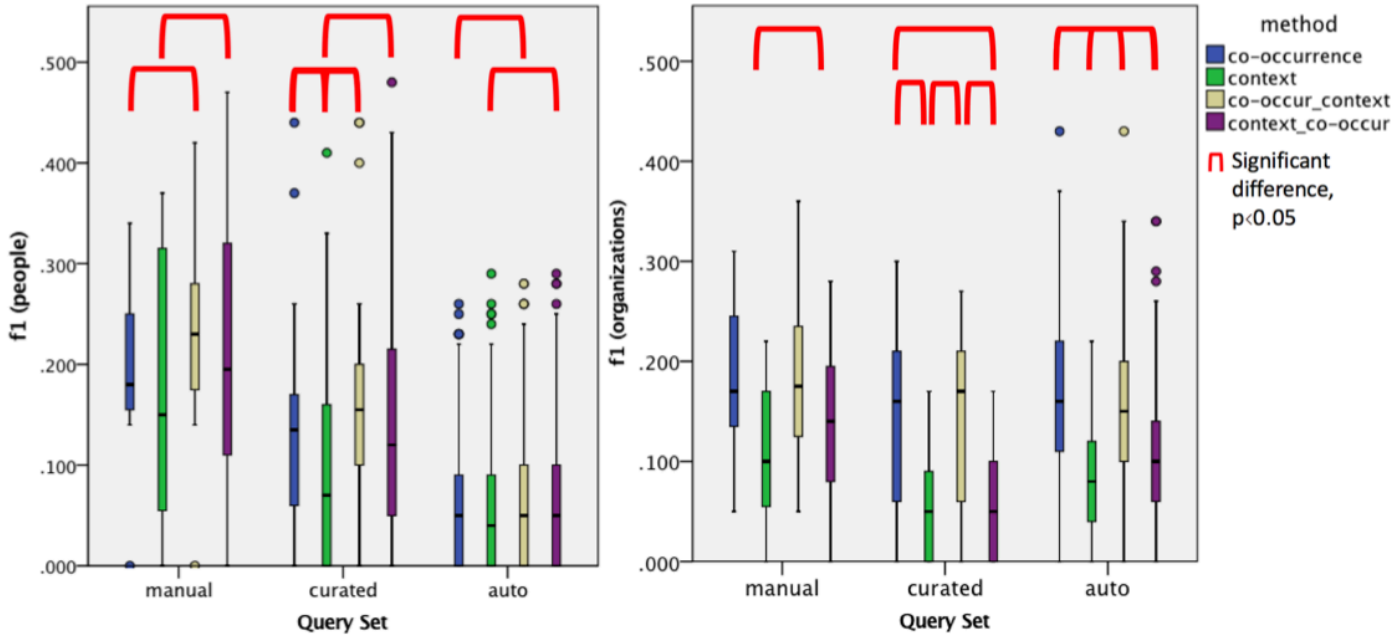
Model | Example Value

```
{
  "field": "string",
  "method": "string",
  "candidates": [
    "string"
```

Evaluation

- Benchmark created using reports written by human experts
  - Human Rights Watch organization & Wikipedia articles on events and people
- We measured the ability of different algorithms to find the concepts mentioned in the original reports
- A combination approach works best in most cases

	person				organization			
	co-occur	context	co-occur context	context co-occur	co-occur	context	co-occur context	context co-occur
MAP	0.233	0.199	0.251	0.233	0.179	0.143	0.184	0.189
F1	0.192	0.174	0.228	0.213	0.178	0.107	0.183	0.141
Pr.	0.133	0.121	0.158	0.149	0.117	0.066	0.119	0.089
Re.	0.372	0.328	0.437	0.388	0.436	0.304	0.459	0.374



Conclusion & Future Work

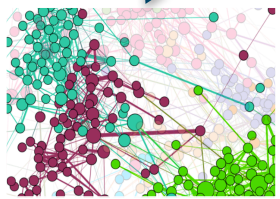
700+ million records
24+ billion values

Structured Event Data

Event Extraction Engine

Concept Index

Semantic Embeddings

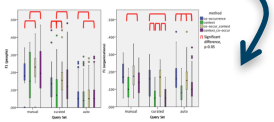


Semantic vectors for every field/value using various (temporal) models

Concept Extraction

Concept Discovery

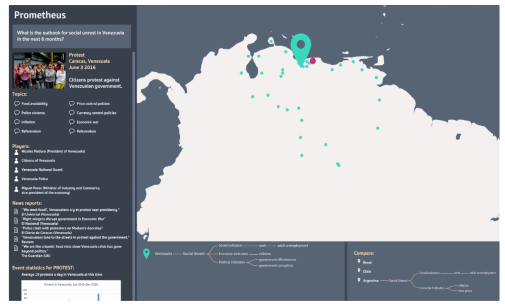
Concept Ranking

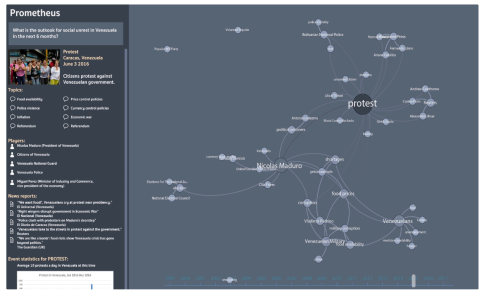


Developed a benchmark & thoroughly evaluated - showing very promising results.
Larger size and better quality benchmark is needed.

Analyst Input
(Natural Language Query, Document, Image, Video)

Visualization, Hypothesis Generation, Causal Reasoning, Forecast Engine



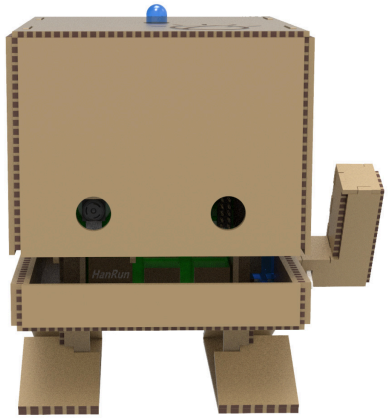


Exploration of key players, locations, events, and scenarios

We are hiring!

and we very much welcome academic collaborations

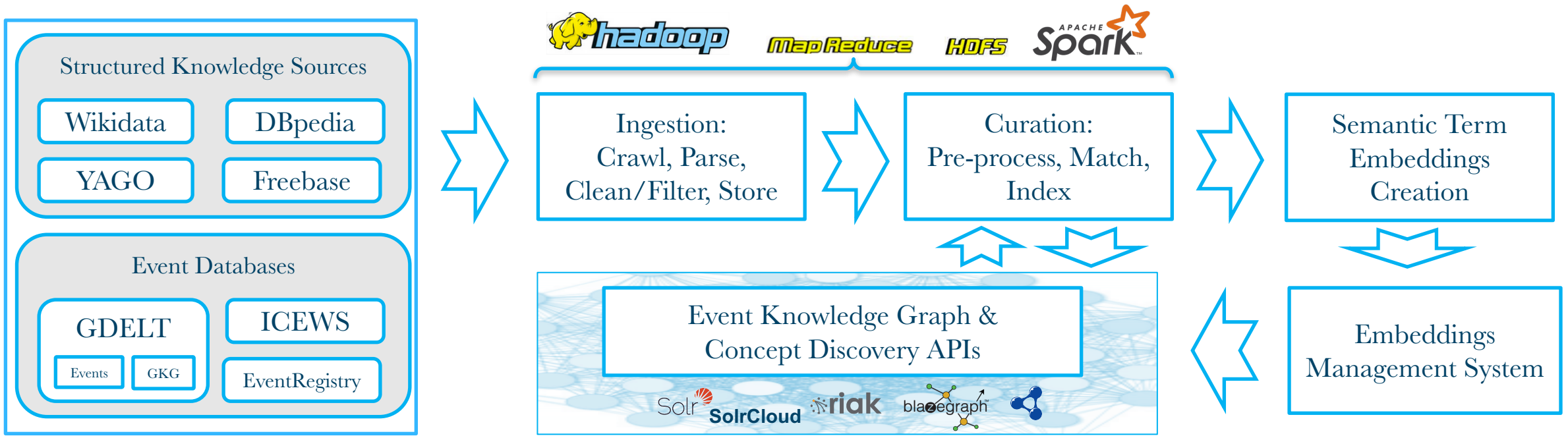
Come to our booth & get in touch.



(You may take one of ☝ these home!)

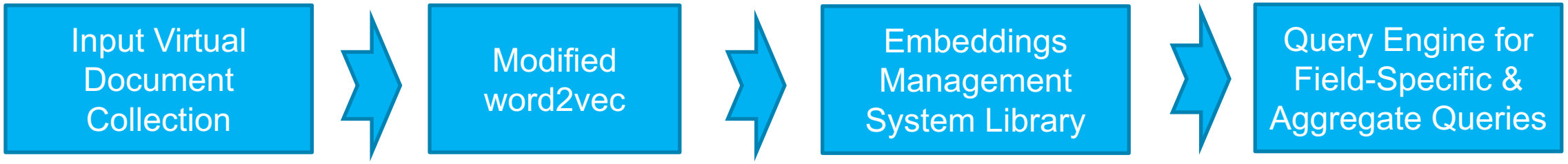
BACKUP

System Architecture



DeepSim (context) Analysis Details – Model Construction & Query Engine

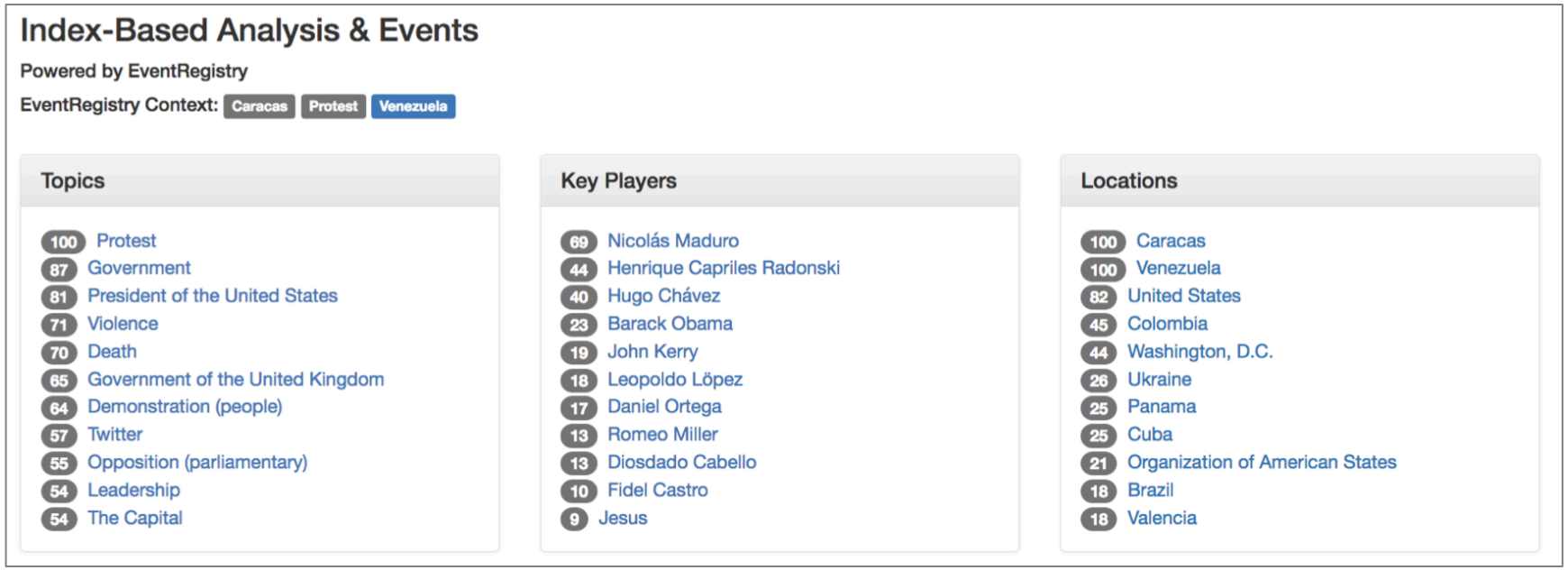
- Step 2: Model Construction & Query Engine



- Word2vec modifications
  - Fixing context window size, rotating window so column order does not affect the outcome
- Embeddings Management System
  - Super fast in-memory approximate nearest neighbor library & API
- Query Engine for DeepSim
  - Prefix-based query for field-specific retrieval (e.g., retrieve ”person”s similar to “location”)
  - AND query over input terms (e.g., retrieve “person” similar to “location X”, “persons Y & Z”

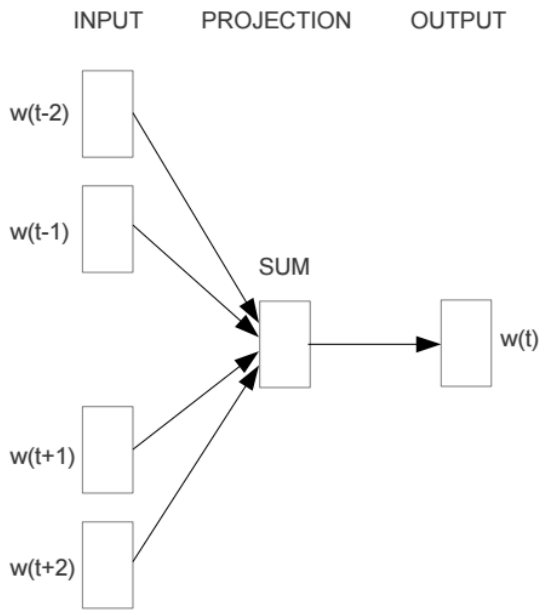
Concept Ranking: Index-Based Method (co-occurrence)

- Index-based Method: Measuring Co-occurrence
  - Formulate a search query using the concepts extracted from the input question
  - Count the concept annotations for every record in the input
  - Return the most frequent annotations of various types (persons, organizations, themes)
  - Use percentage of co-occurrence as the relevance score

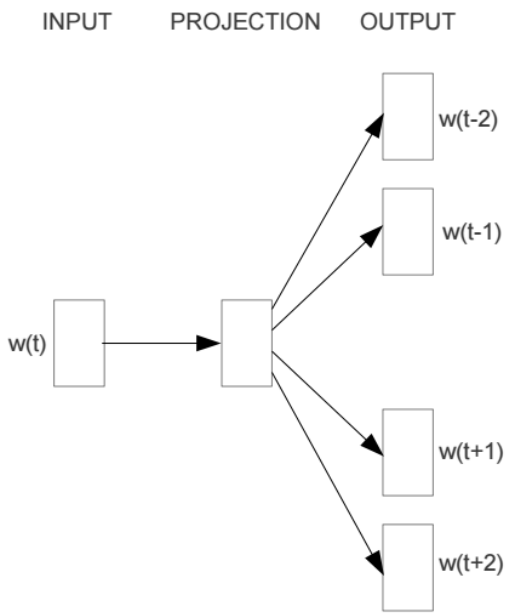


Concept Ranking: Deep Similarity Method (context)

- Using word embeddings to capture the semantic similarity of terms (field values)
- Embeddings in NLP: vectors representing the semantic context of each word
  - Similar terms have similar vectors (as per e.g. Cosine similarity between the vectors)
- Method used: modified **Skip-gram word2vec model** [Mikolov et al., 2013]
  - An efficient, shallow Neural Network model
  - CBOW architecture predicts the current word based on the context
  - Skip-gram predicts surrounding words given the current word



CBOW



Skip-gram

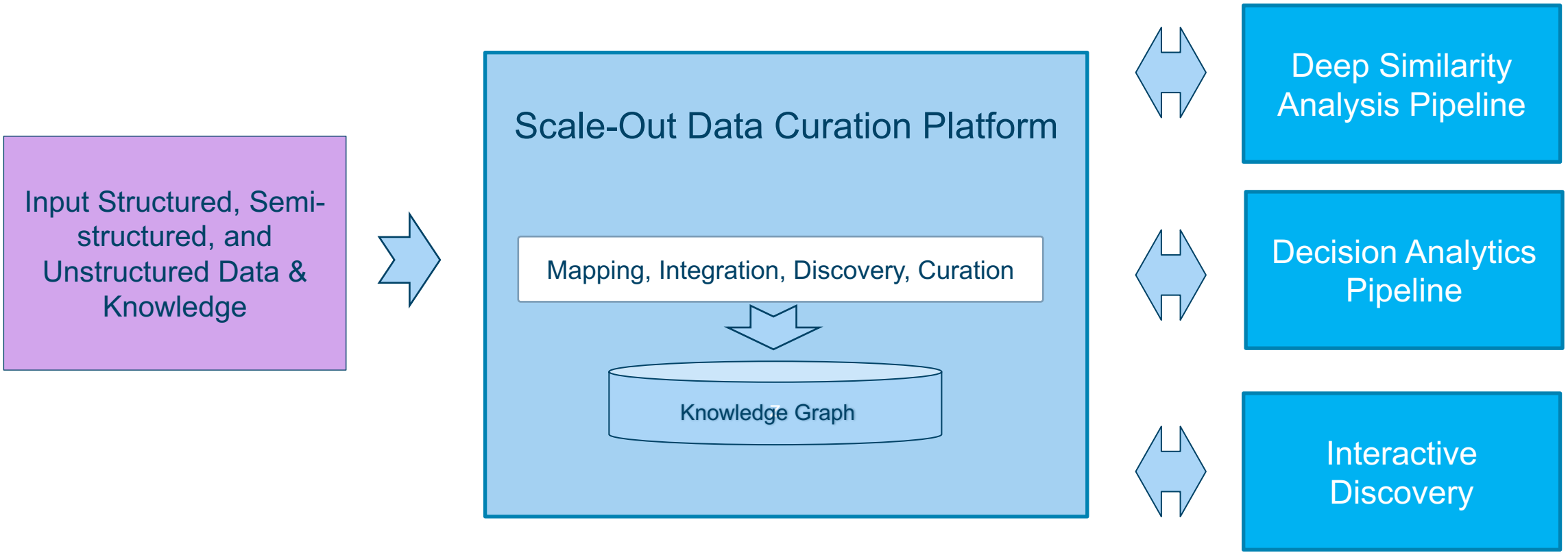
DeepSim (context) Analysis Details – Virtual Document Generation

- Step 1: Virtual Document (Context) Generation
  - A term is a concatenation of a field name (column header) and a field value (cell content)
  - Each record in the input data turns into a context
  - We may need transformations such as data binning / bucketing for numerical fields

date	Location_countryCode	Location_fullName	persons	names	themes	…
2017-01-01T08#00#00Z	AS	Goulburn, New South Wales, Australia	Barnaby Joyce, Alastair Starritt, …	Water Minister Barnaby, Southern Valley, …	GENERAL_GOVERNMENT, EPU_POLICY_GOVERNMENT, …	

```
date#2017-01-01T08#00#00Z tone_selfGroupReferenceDensity#0 tone_tone#-2 tone_activity#2
tone_positiveScore#0 tone_negativeScore#0 tone_polarity#0 person#Barnaby_Joyce
person#Alastair_Starritt person#Alastair_Starritt sourceCommonName#stockandland_com_au
organization#Deniboota_Landholders_Association organization#Deniboota_Landholders_Association
name#Deniboota_Landholders_Association name#Alastair_Starritt
name#Deniboota_Landholders_Association name#Murray-Darling_Basin name#Northern_Basin
name#Basin_Plan name#Southern_Basin name#Goulburn_Murray name#Goulburn_Valley name#Basin_Plan
name#Latrobe_Valley name#Water_Minister_Barnaby_Joyce sourceCollectionIdentifier#WEB
theme#GENERAL_GOVERNMENT theme#GENERAL_GOVERNMENT theme#GENERAL_GOVERNMENT
theme#EPU_POLICY_GOVERNMENT … location_latitudeLongitude#149_721--34_7515 location_countryCode#AS
location_fullName#Goulburn _New_South_Wales _Australia location_type#WORLDCITY
location_featureID#-1576139 location_ADM1Code#AS02 location_ADM2Code#154641
location_latitudeLongitude#149_721--34_7515 location_countryCode#AS location_fullName#Goulburn
_New_South_Wales _Australia location_type#WORLDCITY location_featureID#-1576139 …
```

Towards a Generic Pipeline for Semantic Similarity Analysis



The Need for an Event Extraction Engine

- **Shortcomings of existing event data**
  - No associated text articles. Only URLs are available.
  - No associated meta-data (GKG or GKG-like data) over **historic** articles -> we only have GKG data limited to the past two years.
  - Limited definition of "event" in each source. E.g., GDELT & ICEWS are CAMEO coded (and so only cover a specific type of political events) and EventRegistry defines an event as a collection of articles so we do not know the kind of actions and actors in each event.
  - Noise → both random noise and *systematic noise* (a result of rule-based extraction)
    - EventRegistry
      - encoding issues (tofu characters in labels)
      - inaccurate concept annotations
    - GDELT & ICEWS
      - wrong annotations, missing annotations, basically all the problems a rule-based system could have
- **What we need**: a comprehensive, accurate, and up-to-date event database with annotations similar to EventRegistry and GDELT GKG, historic coverage similar to ICEWS, event coding similar to GSR