MINING DEEP INSIGHTS FROM LARGE POLAR DATA REPOSITORIES WITH APACHE OPEN-SOURCE

Nithin Krishna Ottilingam[1], Karanjeet Singh[2], Chris a. Mattman[2], Siri Jodha S. Khalsa[3], Ruth Duerr[4], Madhav Sharan[1]

University of Southern California, NASA Jet Propulsion Laboratory, University of Colorado, The Ronin Institute
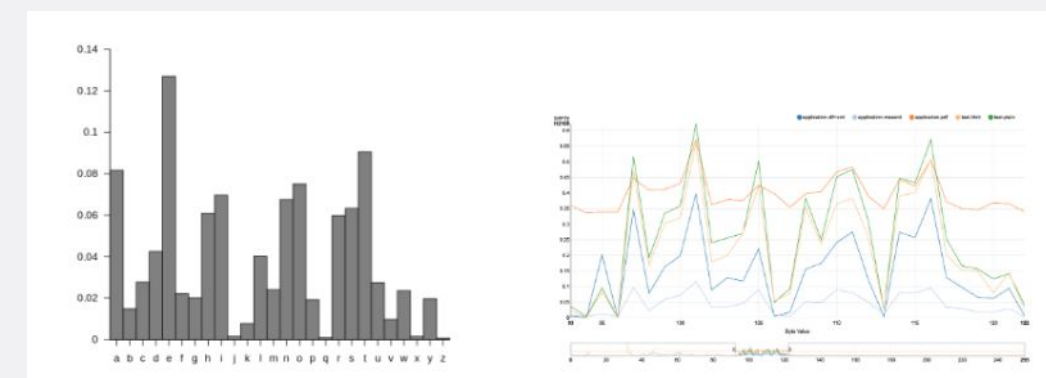
Motivation

Byte frequency analysis of a sub-sample of documents from polar repositories (corpus) indicates that nearly every filetype has some text stored in it, as part of the file contents or metadata

Documents contain rich text:
The mean ratio of size of extracted text from documents by the size of the document is 0.47. 48MB of the 13.51GB corpus is metadata, On average 12% of every document is textual metadata.
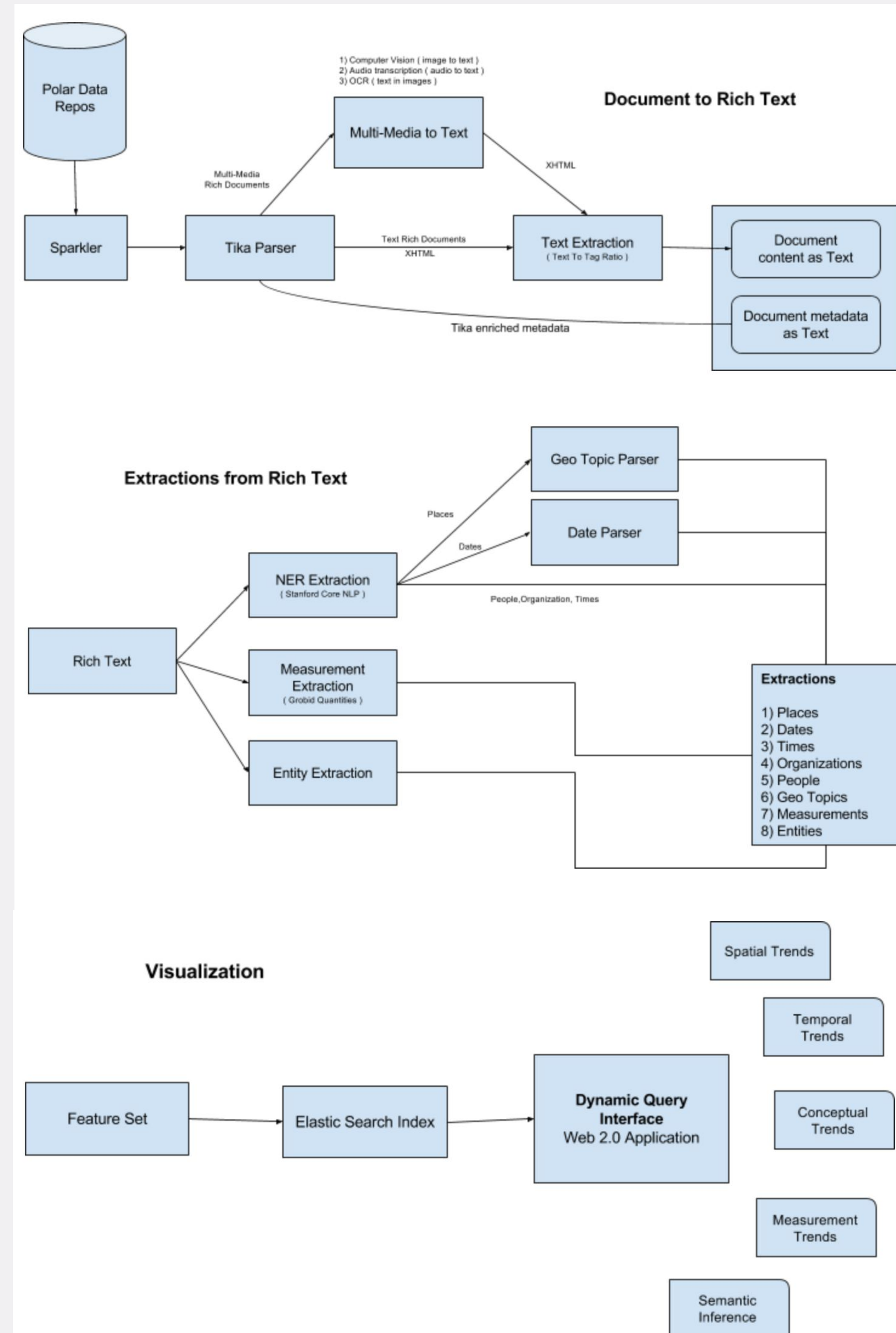
Documents contain rich spatial data:
400K locations extracted from the corpus. 10 locations extracted on average per document, 4 unique locations per document.

Documents contain rich temporal data:
450K dates extracted from the corpus. 11 dates per document.

Conceptual Richness:
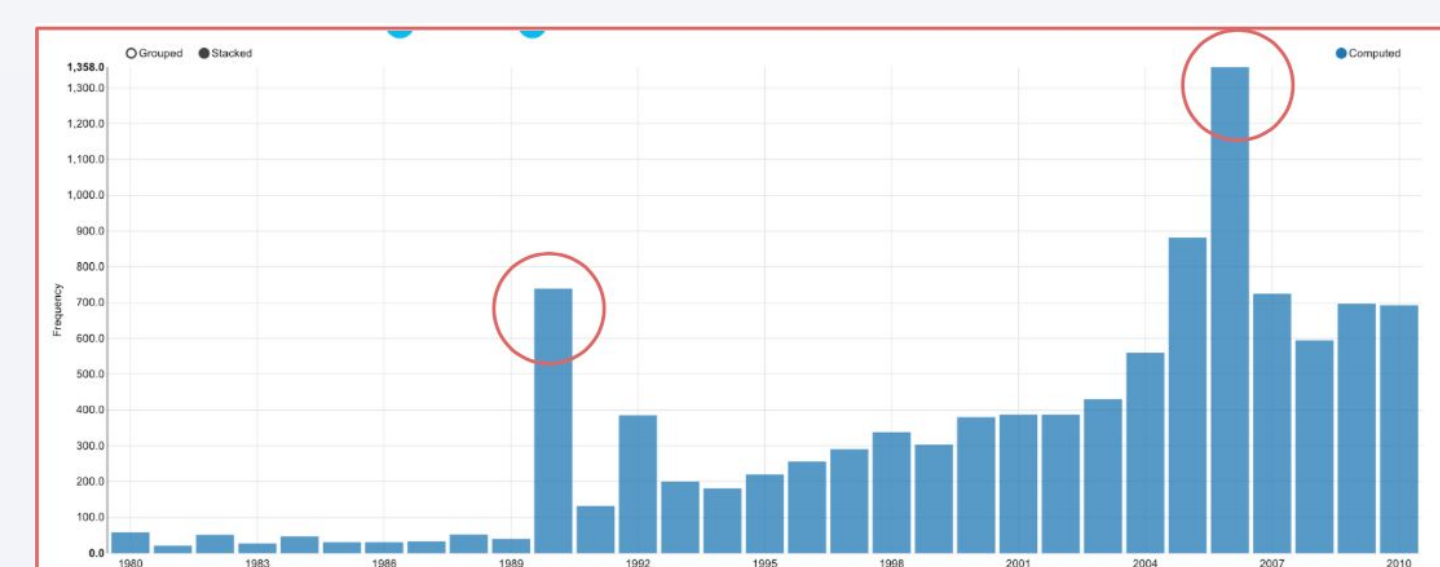833k entities extracted from the corpus. 21 entities extracted per document.

Measurement Richness:
Over unique 150 unit of measures extracted from the corpus.

Architecture - Extraction Pipeline
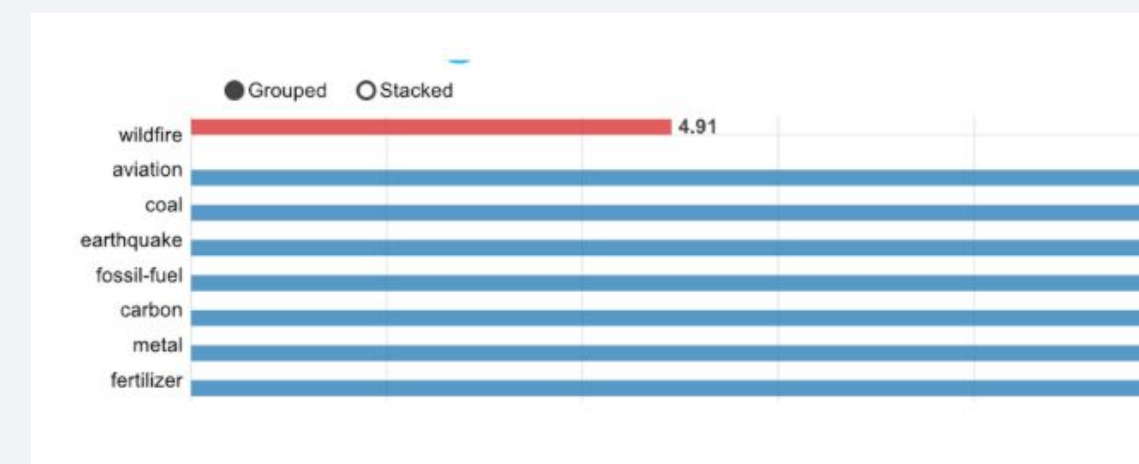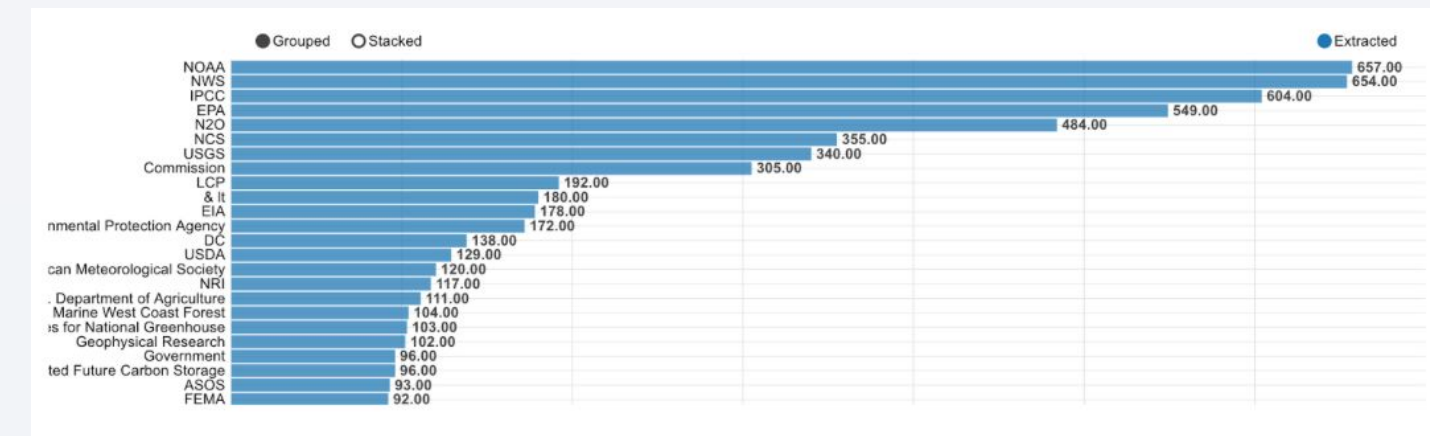

Polar Deep Insights

Wildfire

Temporal trends of Wildfires:
[NBC News] Record number of wildfires in 2006.
[WILDFIRE ACTIVITY STATISTICS] First detailed study about wildfires came out in 1990.

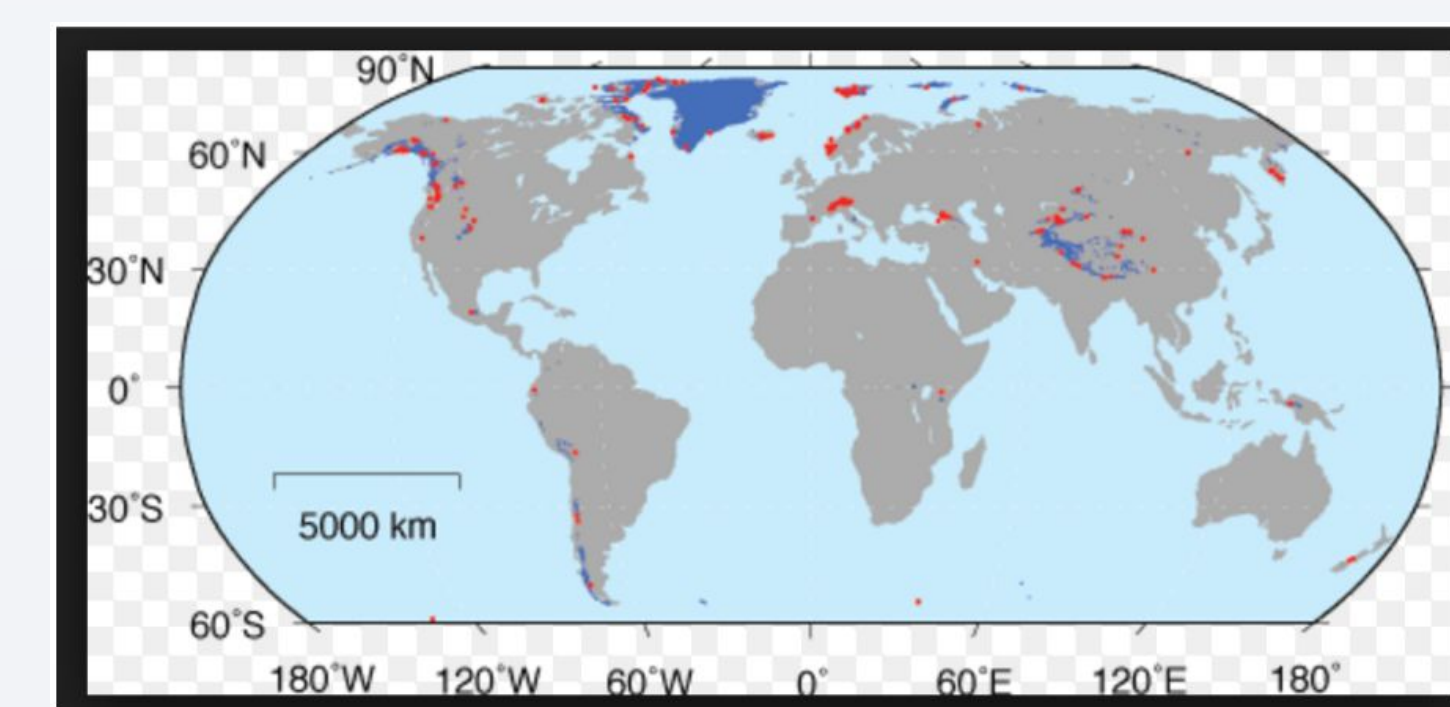**What are the major organizations involved with wildfire study?**

NOAA, NWS, IPCC, EPA

What are the concepts which co-occur with wildfire?
Aviation: use of aircraft and other aerial resources to combat wildfires
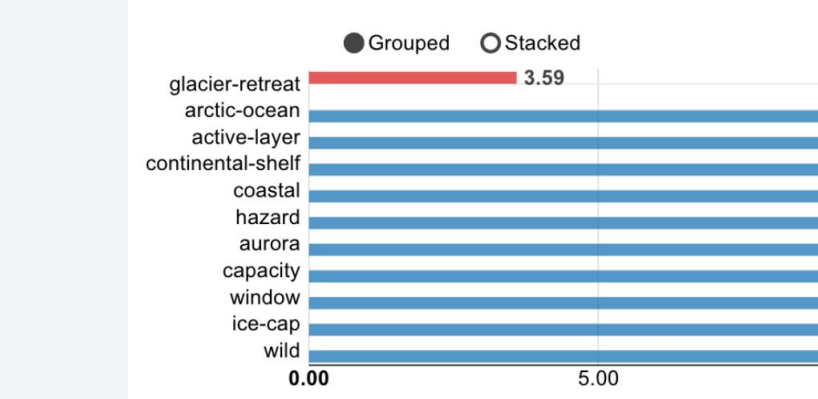**Fertilizer:** Ash from wildfires become fertilizer for plants

Sea level Rise

Documents talking about sea level rise, talk about the following broad concepts:
(Semantic Inference)

1. biological-phenomena
2. human-activity
3. decision-activity
4. sediment
5. ecological-dynamics
6. system-state-change
7. landform
8. realm-configuration
9. biological-role
10. land-use

What are the related concepts which co-occur with sea level rise?
Sedimentation, Renewable Energy Source

Crude oil Prices

Steep price fluctuations in crude oils prices in 1980, 1985, 1990 and 2008 marked by distinct increase in document discussions

An Oil Price Primer

Glacier Retreat

Where is glacier retreat prominent?
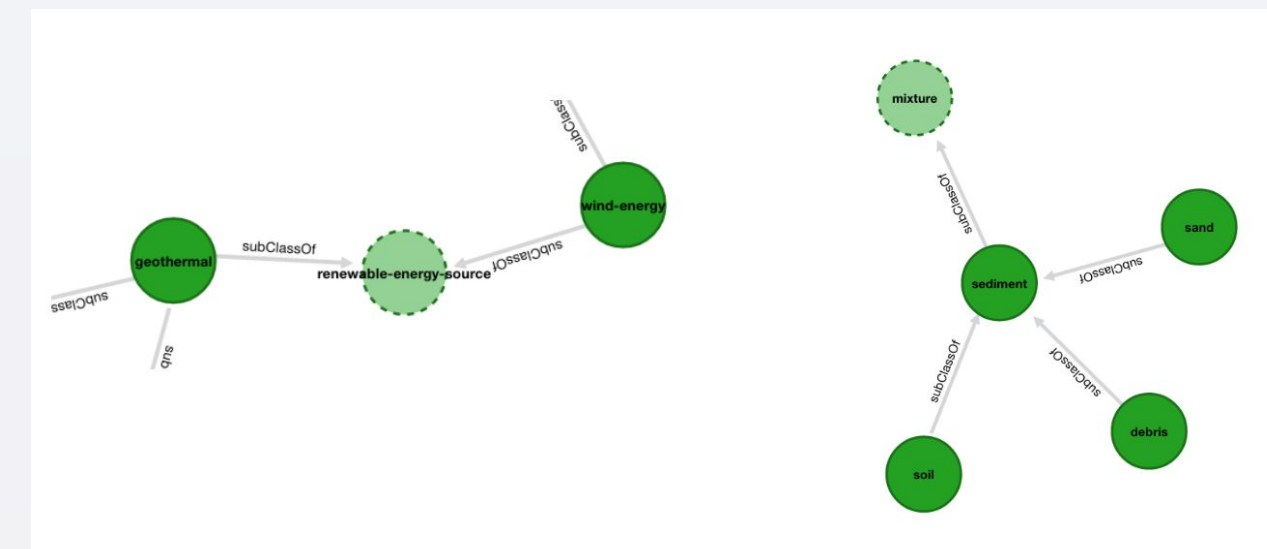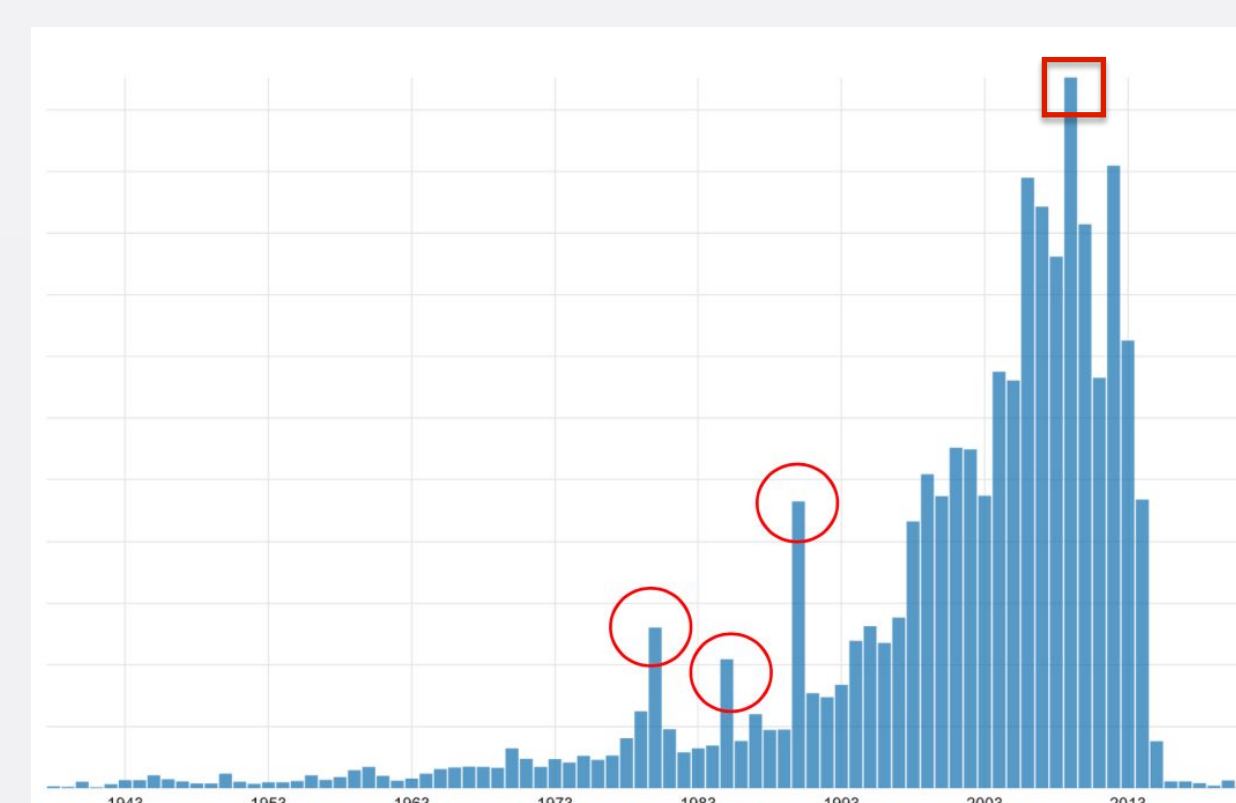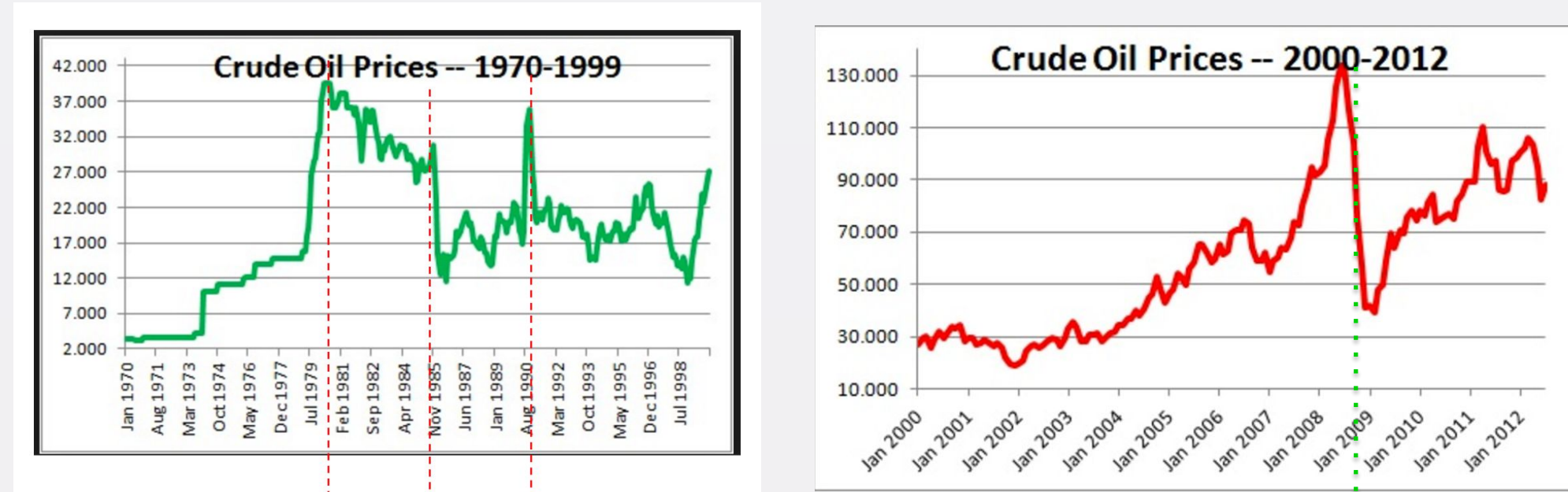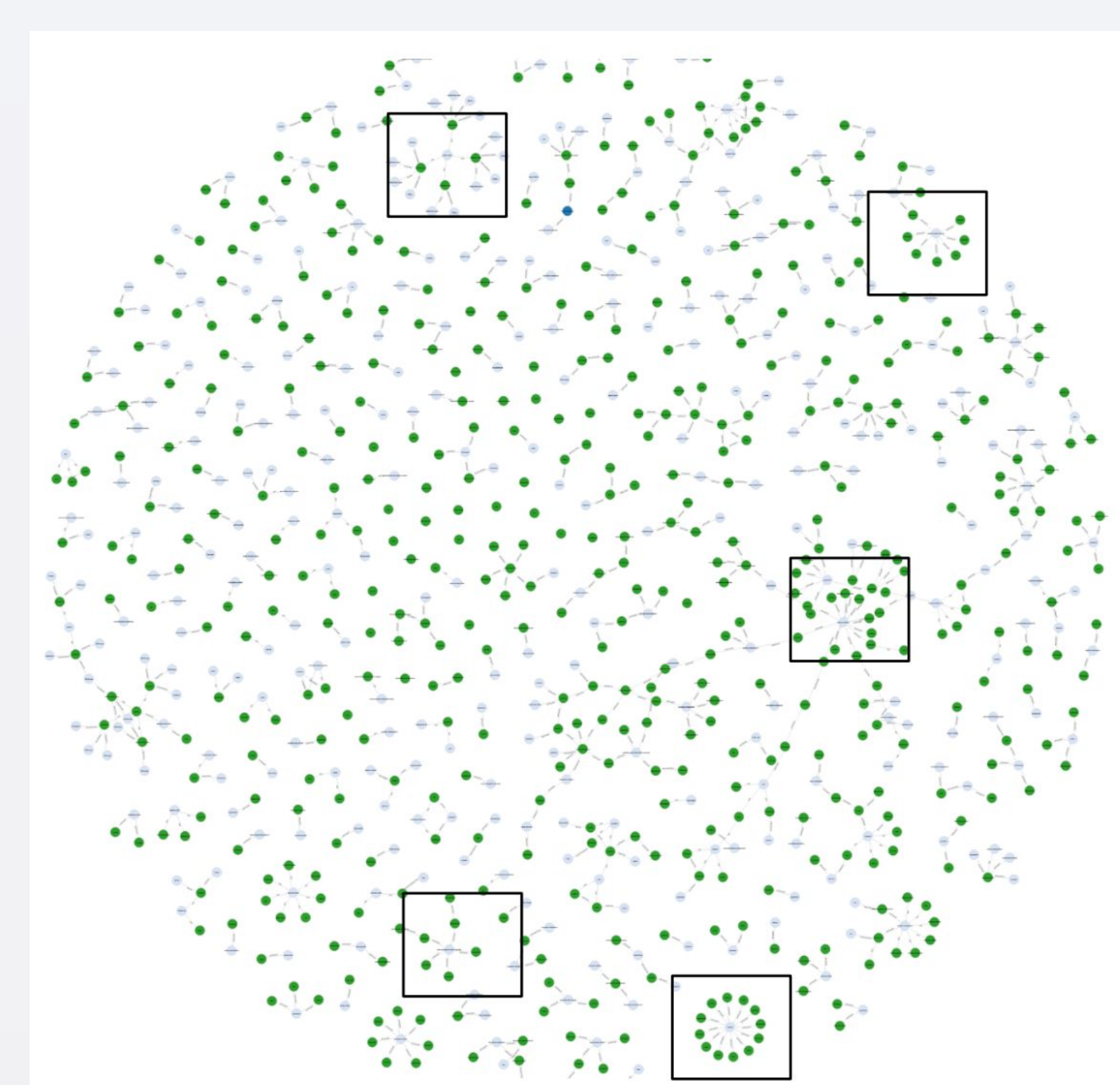Coastal regions of alaska, greenland, Northern Europe

GLIMS: Global Glacier Recession

What are the concepts which co-occur with glacier retreat?
Arctic Ocean, Continental shelf, Coastal, Hazard, Aurora ice-cap

Climate change indicators

Which are some of the indicators of climate change?
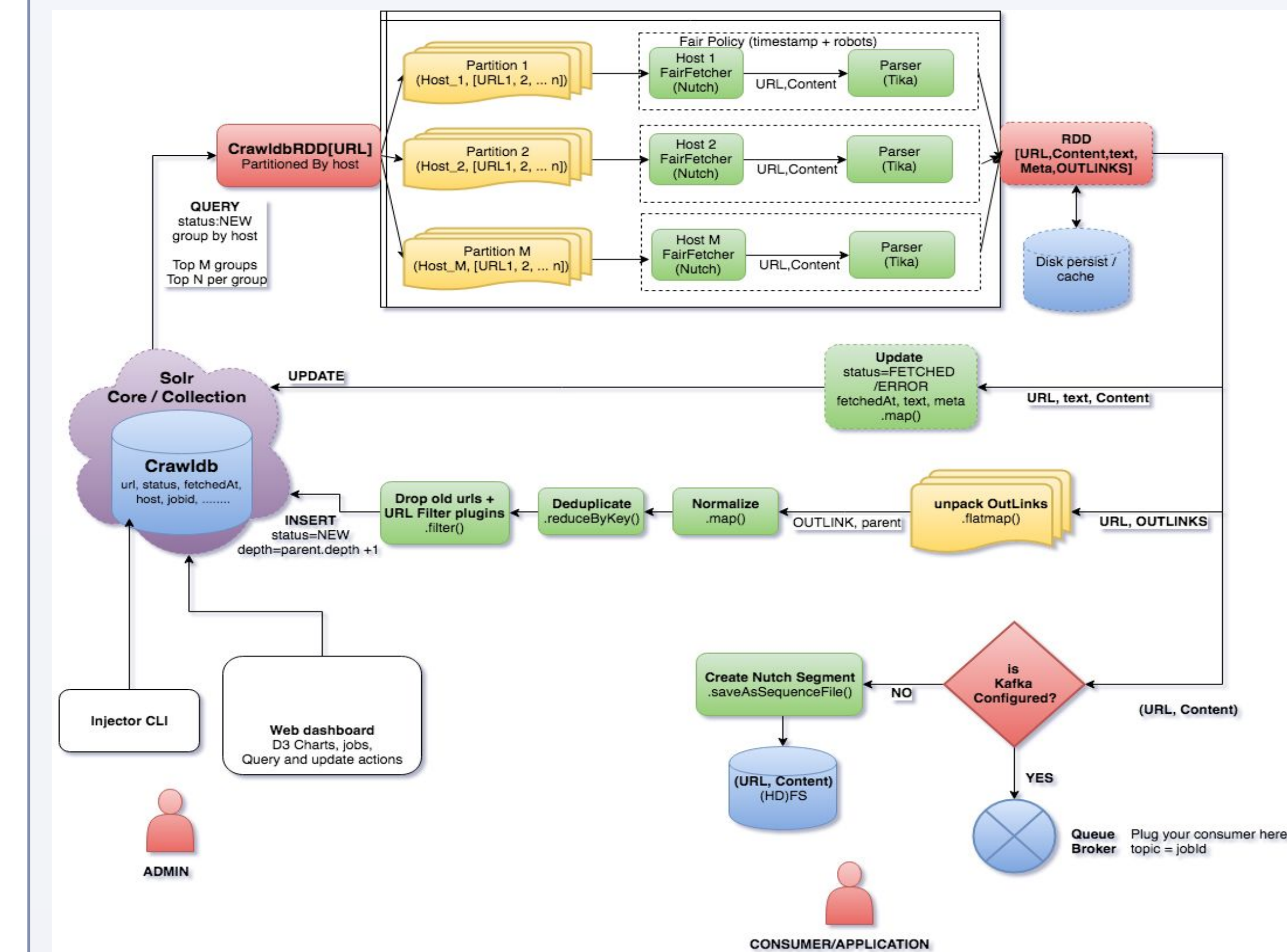(Semantic Inference)

- animal
- marine-animal
- organic-compound
- earth-oscillation
- system-state-change
- element
- Vegetation
- Human-activity
- environmental-control-strategy

Semantic Inference:

In all the documents which talk about climate change, we compute a subset of the knowledge graph with concepts within these documents. We then locate the concepts in the subgraph with the highest in-degree indicating causal inference.

Sparkler

Crawl Cycle

Banana Dashboard

Future Work

- We look to scale up our infrastructure to process larger data repositories faster.
- The SWEET ontology is generic earth sciences. We look to Investigate polar specific ontologies to gain specific insights about the cryosphere.
- We look study and derive insights about current polar research trends.
  ★ Spatial distribution of various types of ice/snow around the world
  ★ Temporal trends impacting snow/ice
  ★ Impact of Drought, population, temperature increase, natural hazards, public affairs, urban heat wave, dumping of chemicals etc. on the cryosphere.
  ★ Rate of glacier recession
  ★ Trends pertaining to Rotten Ice
  ★ Quantify sea water intrusion, Sea level rise and Inundation
  ★ Role of Heat Transport on Arctic Ice
  ★ Investigate Seasonal Changes to the cryosphere
  ★ Global Snow Estimation

References

Polar deep insights: http://polar.usc.edu/polar-deep-insights