Visualization Interface

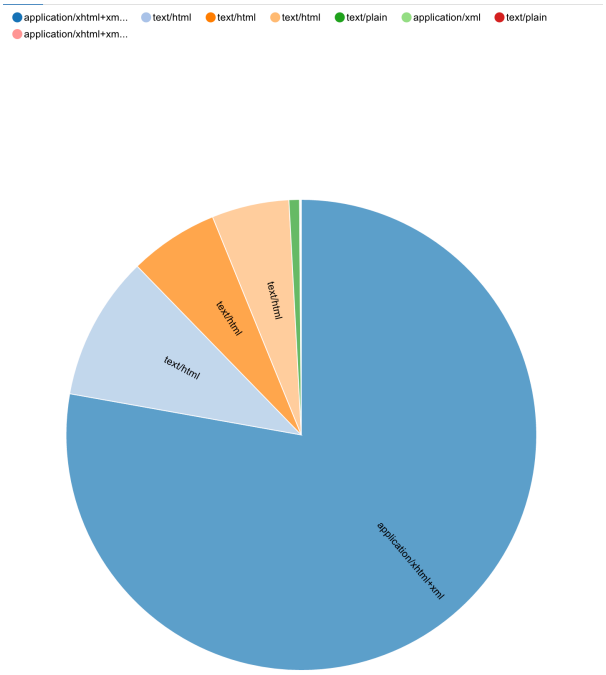
Walkthrough video:

All the described visualizations are built on a subset of documents returned as a result of a user query. If the user chooses not to enter a query; it displays statistics for all the documents in the index.

(e.g.) Query => All documents with the term 'sea-ice' and within 'greenland'.

Stats:

1) **Tika content-type distribution:** The pie chart describes the number of documents of a each content type as recognized by tika.



2) **Document count / size statistics:** This table describes statistics about the amount of information extracted from each document.

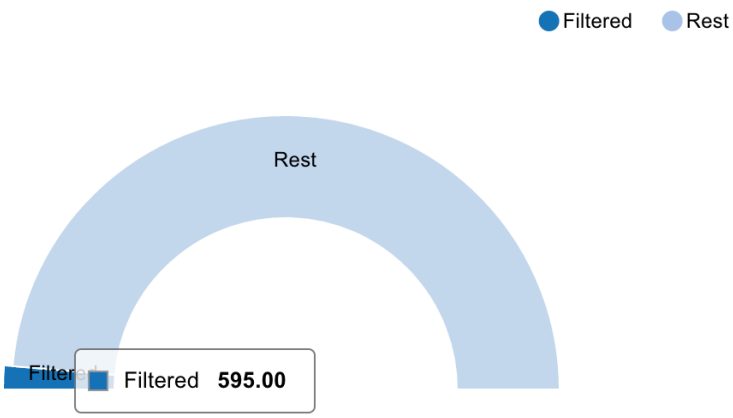
	Max	Average	Sum (Bytes)
Number of Documents		21,967	
Extracted information size	1,826,112	13,251.63	291,098,568
File size	462,306	46,268.40	1,016,378,049
Information Extracted (Extracted size / File size)	3.950	0.29	0.286
Metadata size	1,048	1,040.90	22,865,512
Metadata ratio (Metadata size / File size)	0.002	0.02	0.022

3) **Entity statistics:** This table describes statistics on the number of extracted entities/dates/locations per document.

entities

Type	Max	Average	Total
Extracted terms	172	1.92	42,116
Distinct extracted terms	19	0.72	15,797

4) **Query ratio:** This pie chart describes the ratio of documents returned as result of a given query to all the documents in the index.



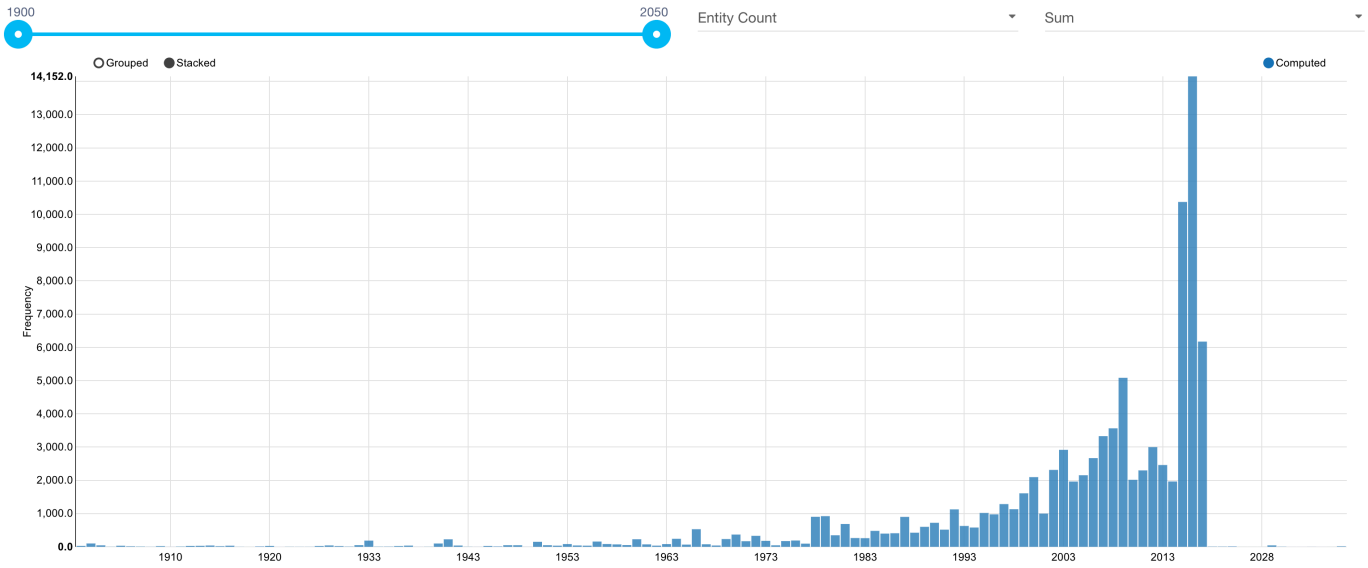
Temporal Distribution:

Purpose: It can be used to mine temporal trends of entity occurrences over the given set of documents.

Most relavent parameter: [Entity count] + [Sum]

(e.g.) How has the occurrence of the term 'sea-ice' varied with occurrence of time?

Check out the "crude-oil prices" use case in our ESIP poster.



Spatial Distribution

Purpose: It can be used to mine spatial trends of entity occurrences over the given set of documents based on references to places in these documents.

(e.g.) What are the most talked about places in documents talking about 'sea-ice'?

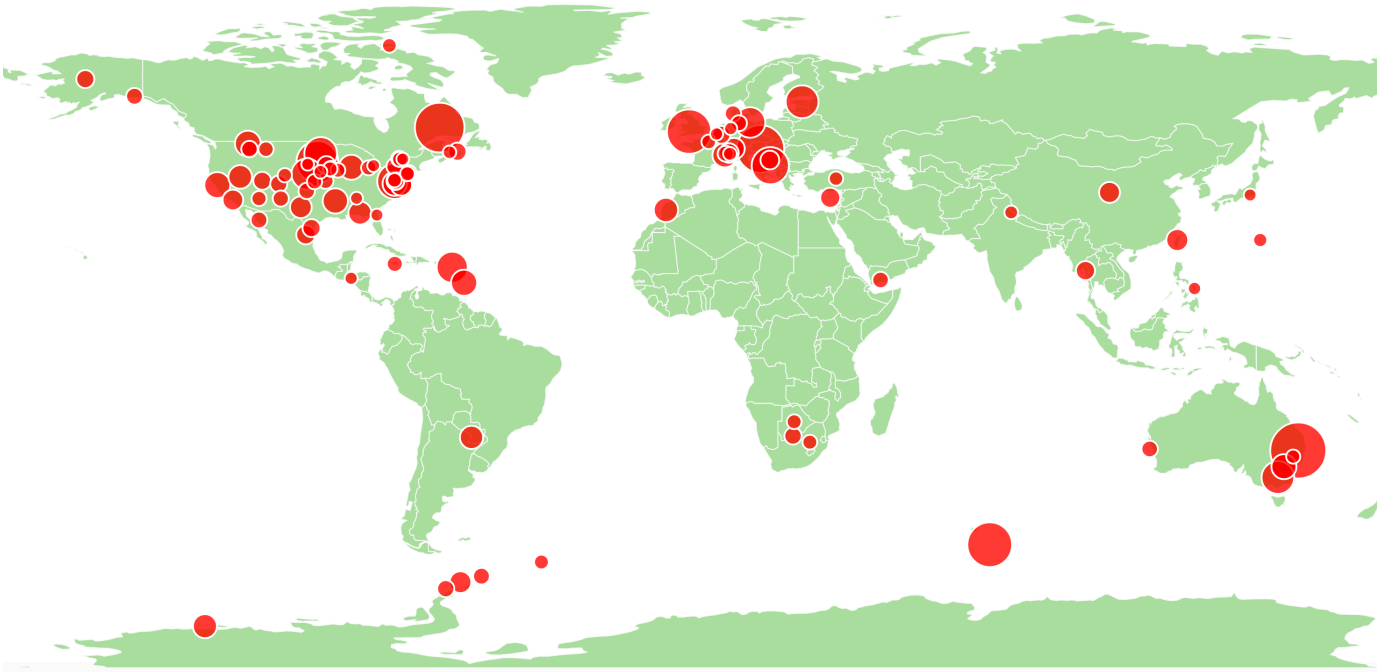
Parameter: [Entity count] + [Sum]

(e.g.) What are the 'significant' places in documents talking about 'sea-ice'?

Parameter: [TF-IDF] + [mean]

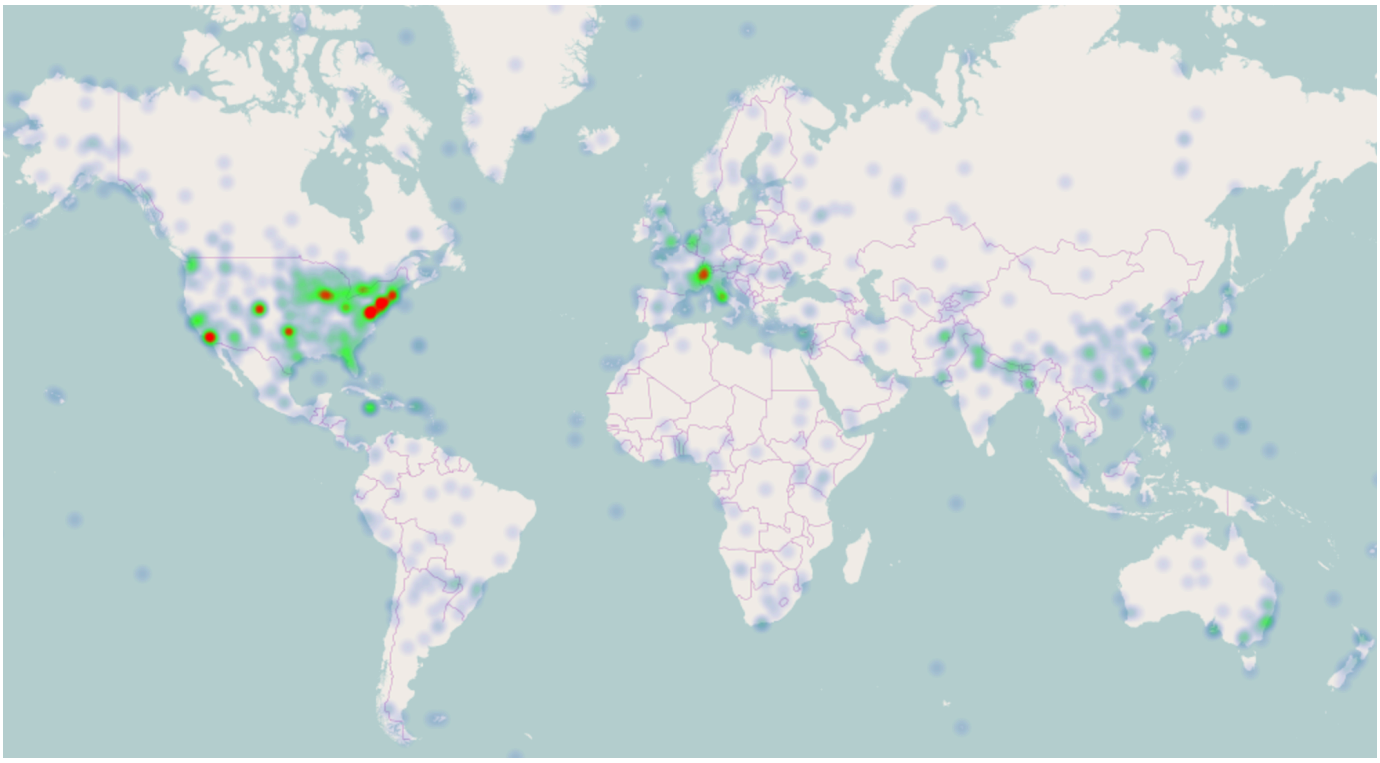
'significant' : Locations which are the most frequently occurring in set of documents might not be the most significant. (e.g.) Most authors of reports might be affiliated to 'University of Southern California'; hence Los Angles might not be significant in the context of this search. We overcome this issue by weighting each 'location' by a TF-IDF score (Term frequency - Inverse document frequency). It is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpous.

Check out the "glacier-retreat" use case in our ESIP poster.



Heat Map

Purpose: This visualization is exactly similar to the spatial distribution visualization in function; but varies in presentation. The heat map visualization allows the user to zoom-in / zoom-out into the map to view each location in a fine grained manner.



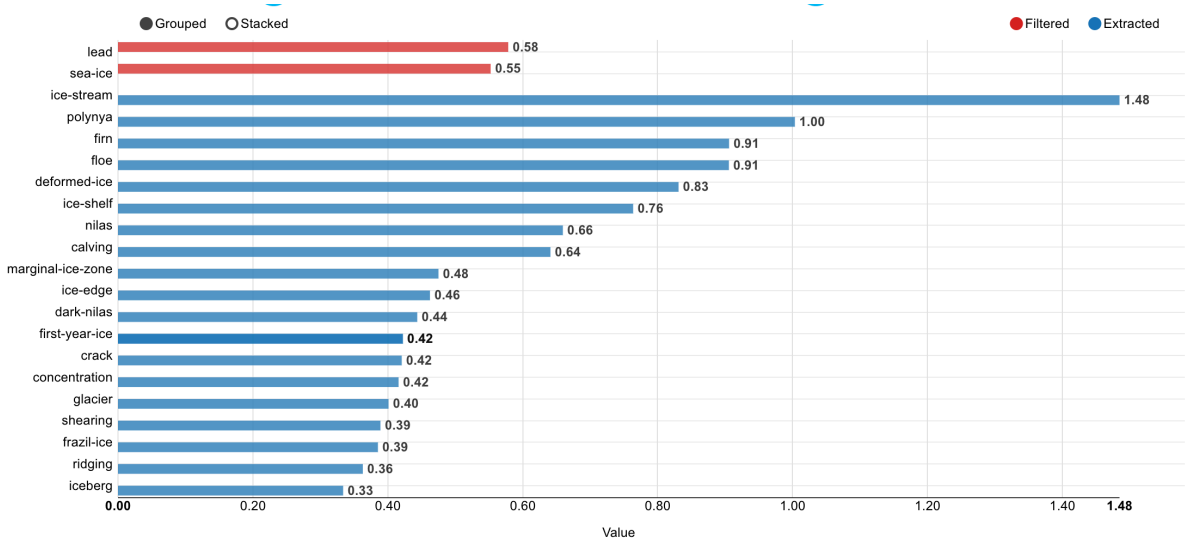
Concept Correlation

Purpose: This visualization tries to capture the correlation between concepts based on co-occurrence.

(e.g.) In documents which talk about 'lead' and 'sea-ice' what are the other concepts 'significance'?

Parameter: [IF-IDF] + [Mean]

Check out the "wildfire" use case in our ESIP poster.



Semantic inference

Inferred
glacier-ice
calved-ice-of-land-origin
ice-of-land-origin
openings-in-the-ice
forms-of-floating-ice
arrangement

Semantic Inference

Purpose: This visualization tries to capture the higher order concepts described in the queried document set.

(e.g.) In documents which talk about 'lead' and 'sea-ice' what are the 'higher order concepts' of significance?

Parameter: [IF-IDF] + [Mean]

'higher order concepts' : Parent concepts described in the concept graph. (e.g.) 'river-ice', 'lake-ice', 'drift-ice', 'fast-ice' .. (etc) are forms of 'floating-ice'. Documents might not necessarily contain the term 'floating-ice' but we can infer that they are talking about the 'floating-ice' if they contain one or many of the various 'floating-ice' types.

Such inference of higher order concepts from the concept graph is what we describe as semantic inference.

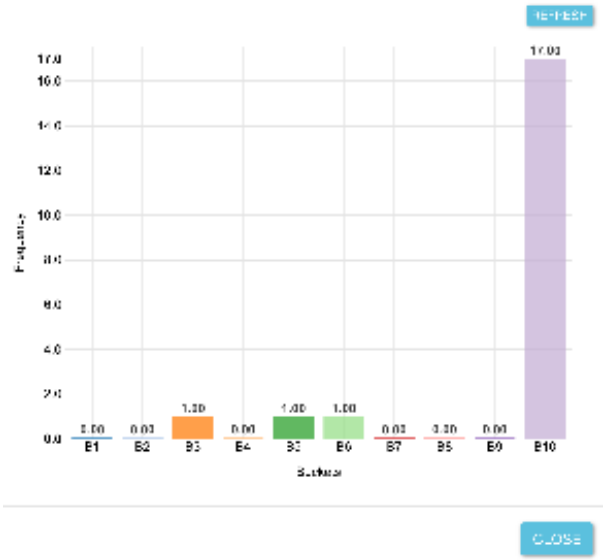
Check out the "climate-change indicators" use case in our ESIP poster.

Measurement Distribution

Purpose: This visualization displays the extracted measurements from the various documents. It also displays the range of the extracted measurements and a histogram of their values.

Unit	Type	DocCount	Min	Max	Average
GHz	misc	3431	-100.00	340.00	34.84
µm	misc	2526	0.00	1,105.00	29.27
square kilometers	misc	623	0.00	1,000,000.00	141,421.35
million square kilometers	misc	617	0.00	361.00	10.56
°C	misc	308	-89.00	1,000.00	67.36
°S	misc	244	0.00	90.00	83.27
km2	misc	104	0.00	500,000.00	59,558.27
micrometers	misc	104	0.00	1,100.00	21.06
gigatons	misc	68	-57.00	192.00	70.12
kg/m2	misc	57	0.00	24.00	5.63
ppb	misc	42	2.00	275.00	131.93
kHz	misc	35	1.00	420.00	57.43
milliradians	misc	26	0.00	8.00	7.15
square kilometers per day	misc	25	0.00	89,500.00	45,984.00
million km2	misc	24	6.00	18.00	12.38
m/year	misc	21	20.00	100.00	90.50

Measurement histogram



Known Issues:

1) **Preserving context** : The major problem with our extraction process is that it is context insensitive. It does not understand the difference between 'Southern California' in the reference section of a report as supposed to 'The Bearing Sea' which might be of importance to the topic of discussion.

2) **Measurement extraction :** The measurement extraction library which we use to extract measurements from the crawled documents has the following issues.

- Parse errors: Grobid quantities is prone to parse errors. (e.g.) (negative) '-100m' extracted measurement might in actually be a hyphen in the actual text.

- Normalization: It normalizes all occurrences of measurements into 'SI' scale. (i.e.) 'Light years' and 'Km' are both measures of distance and hence are converted to meters. 'Light years' are specific to space. However if it's converted to meters we are loosing out on the context reference.

Solutions:

1) Using the intuition that extracted entities in the top / centre of the document might be of more importance than the ones in the end; we use document offsets to weigh entities. This is almost built; in our next extraction iteration we will test out some of the results from this approach.

2) There are known limits to the current measurement extraction library. Open to discussion on how this can be improved.