MINING DEEP INSIGHTS FROM LARGE POLAR DATA REPOSITORIES WITH APACHE OPEN-SOURCE [1]

Submitted by karanjeetsingh on Tue, 2016-12-20 03:53  Event: Winter Meeting 2017 [2]

Abstract:
We look to generate deep insights about trends influencing the polar domain by mining deep insights from documents in arctic data repositories like ACADIS, AMD and NSIDC. Most polar geosciences data centers have a difficult time answering analytic questions about their data repositories because of their bias toward high volume scientific data over textual data; due to the popular belief that textual data is less voluminous and hence is less meaningful. Our work is motivated by a desire to understand trends and draw insights from the neglected textual data collected in the polar geosciences community. Our studies indicate significant amounts information can be gleaned from the text in documents. We find that the literature in this text is spatially and temporally rich. It describes a wide array of issues pertaining to the arctic region like oil spills, glacier retreat, sea level rise (etc). It contains a diverse set of scientific measurements and observations captured in by field, airborne instruments and spaceborne sensors (eg ice core and weather station data, high resolution LIDAR snow depth measurements and ICESat elevations and waveform returns). We have built a scalable end-to-end analytics pipeline with Open Source technologies out of the Apache Software Foundation. This pipeline crawls data from various polar data repositories, extracts textual information and builds an index of rich features for each given document. Our visualization module then uses these enriched features to mine insights about arctic ecosystem with context from predefined semantic knowledge about the polar domain (JPL SWEET). We use Sparkler (an evolution of Apache Nutch) to crawl the polar data repositories. Sparkler is a new distributed web crawler which is developed in-house by USC Data Science team. This crawler is heavily inspired from Apache Nutch and handshakes with many other Apache projects thereby supporting the open source community and pushing the limits for further development. We use Apache Tika to extract text and metadata from any document. We convert non-textual multimedia files like audio and images into text through open source transcription and computer vision libraries integrated into Tika. We then build enriched feature sets from the Tika refined text; with locations, dates, measurements and terms-of-interest extracted out using Tika’s GeoParser, GrobidQuantitiesParser and NERParser. With Wrangler, an NSF funded supercomputer, managed by TACC (Texas Advanced Computing Center) we are able to scale our pipeline to crawl and extract insights from hundreds of thousands of documents. For multiple analytic use cases, we find that reasonable correlation between generated trends and real world data.

Reference: ACADIS
AMD
NSIDC
Apache Software Foundation
Sparkler
Apache Nutch
Apache Tika
GeoParser
GrobidQuantitiesParser
NERParser
JPL SWEET
Wrangler, TACC


Author(s):
Name: Chris Mattmann [4]
Organization(s): Jet Propulsion Lab [5]
Email: chris.a.mattmann@jpl.nasa.gov [6]

Name: Nithin Krishna Ottilingam [7]
Organization(s): University of Southern California [8]
Email: ottiling@usc.edu [9]
MINING DEEP INSIGHTS FROM LARGE POLAR DATA REPOSITORIES WITH APACHE OPEN-SOURCE

Published on Commons (http://commons.esipfed.org)

Name: Karanjeet Singh [10]
Organization(s): University of Southern California [8]
Email: karanjes@usc.edu [11]

Name: Madhav Sharan [12]
Organization(s): University of Southern California [8]
Email: msharan@usc.edu [13]

Name: Ruth Duerr [14]
Organization(s): The Ronin Institute [15]

Name: Siri Jodha S. Khalsa [16]
Organization(s): University of Colorado, CIRES, NSIDC [17]

Keywords: Big Data [20]
analytics [21]
Text mining [22]
Apache [23]
Open Source [24]
Crawl [25]
Polar [26]

Source URL: http://commons.esipfed.org/node/9663

Links:
[6] mailto:chris.a.mattmann@jpl.nasa.gov
[9] mailto:ottiling@usc.edu
[11] mailto:karanjes@usc.edu
[13] mailto:msharan@usc.edu
[22] http://commons.esipfed.org/taxonomy/term/2578