NETL's SmartSearch Deep Learning Tool



Scalable Data Search and Parse for Carbon Storage Data Discovery



Scaling the Data Pyramid

artificial intelligence, & machine learning



Challenges scientists face in order to effectively use data resources:

Data Interoperability:

Large variety of data makes it difficult to create, exchange, and use



Inform

Data Analytics & Visualization:

Require advanced computational capabilities and large data stores

Data **Discovery**:

Studies estimate avg data project only accesses 20% of relevant data

> Scaling the data pyramid is not simple

> > 4

Data Access:

~80% loss of published data after 20 yrs.

U.S. DEPARTMENT OF

Instrumentation, lab reports, sensors, external (Expert-driven Web Search data, user generated data



A machine learning, big data tool for rapid, online, .zip, & FTP spatial & non-spatial data mining with Hadoop + Bing + ESRI

The "Big Data" Problem

The value of data



SmartSearch N

NATIONAL

TECHNOLOGY ABORATORY

Data Discovery Challenges

Collect

Data is often unstructured, mixed:

- Spatial, contextual
- FTP, WWW, local filesystems, storage area networks, etc.

Convoluted ways to search for and identify data:

• Hard to identify all the data, i.e., see the whole "Elephant", without falling down the "rabbit hole"



SmartSearch



NATIONAL

Conquering the Data Avalanche

- How do you currently search?
 - Type in a few keywords
 - Skim the top few results

EPARTMENT O

• Type in more keywords and try again

- How do you find something relevant?
 - Open a file / web page
 - Read it (skim it)
 - Decide if it's relevant





What is SmartSearch?



Problem: You like these files.

You want to find more data relevant to the content of these files



Solution:

SmartSearch automates data discovery by ...

1) Analyzing content you like



2) Finding new content via www, local, enterprise data stores



3) Telling you how relevant the new data is to what you like





- Infinitely Scalable (Automated) Data Discovery
 - Analyze millions+ of files and generate comparison metrics
 - Generate topic models, categorization
 - Desktop, cluster, cloud
- Treat geospatial data like a document
 - Automatically extract text from geospatial data (shapefiles, gec
 - Compare textual vs geospatial data to identify relevancy
- Search for meta tags within HTML body of discovered web sites
 - i.e., find map tags
- Analyze archive files even archives within archives (zips within zips, etc.)
 - Process every file docs, spatial, etc.











Seed Entry



Users input urls to web resources, along with any relevant API keys

SmartSearch Enter Seeds Sessions	s Seeds & Discoveries		admin 🔻
Setup Seeds			
Custom Tag: EDX			
Seed URLs: https://edx.netl.doe.gov			💼 Add Additional Link 🕂
Add He Add Headers for Seed Links			×
URL: https://edx.netl.doe.gov	Header Key:	Header Value:	Add:
			Cancel 🗙 Save 🔀







Initial seed data provided to SmartSearch – document here is parsed and ready to be processed via NLP – key terms, locations

SmartSearch Enter Seeds Sessions Seeds & Discoveries	jo	bradovich 👻
Seed Parsing Status - Tag: paige_test_data		
Source_uri	File	Status
https://edx.netl.doe.gov/dataset/aa360bdf-6d44-428b-bd64-fa702d9399a1/resource_download/363b8502-57c6-4d3d-839b-303e191476ab	web.pdf	Finished
API 🏟 Edit Terms 😰 Add Seeds 🕂 Pause Sessi	on 💵 🛛 Canc	el Session 🔲

© 2021 Mid-Atlantic Technology, Research and Innovation Center, Inc. With acknowledgment of DOE/NETL sponsorship under contract DE-DT0013924.



SmartSearch

NATIONAL ENERGY

TECHNOLOGY LABORATORY

Term Editor - Tag: EDX

- Seed document keywords being extracted.
- Produces a list of terms which can be approved by user
- User can also provide approved, supplemental terms to aid search

U.S. DEPARTMENT OF



Select	: a Model/Embe	edding			□~
ner_o	dl/glove_100d 🛛 r	er_dl_bert/bert_bas	se_cased onto_10	0/glove_100d	Download CSV 🗎
Show	10 🗸 entries			Search:	
	Term 🔶	Count 🗸	Model 🔶	Embeddings \$	Labels 🔶
	EDX	4	ner_dl	glove_100d	ORG
	Department of EnergyDOE Office of Fossil EnergyDOE	3	ner_dl	glove_100d	ORG
	SubmissionsA Multi-criteri	1	ner_dl	glove_100d	ORG
	Gulf	1	ner_dl	glove_100d	LOC
	Welcome - EDXSkip	1	ner_dl	glove_100d	ORG
	Google Chrome.Fin	1	ner_dl	glove_100d	ORG
	upLoginRegist ered Users: 2,764Resource Count: 198,913Resour ce Downloads	1	ner_dl	glove_100d	ORG
	Energy Data eXchange?You	1	ner_dl	glove_100d	ORG
	Mexico	1	ner_dl	glove_100d	LOC

- Seed parsed terms are then used to search the WWW for other relevant results
- Results are returned of webpages that likely contain relevant results

U.S. DEPARTMENT OF

-28	8-2021 14:09:32	2				~
ckli:	st Selected Domain:	s &			Search:	
l	Search Provider	¢ Domains	Count per 🛛 🖨 Domain	АРІ Кеу	API Value	View All Links
i.	Bing	www.michigan.gov	8			
	Bing	irp-cdn.multiscreensite.com	7			
	Bing	en.wikipedia.org	6			
	Bing	wmich.edu	6			
	Bing	pubs.usgs.gov	5			
	Bing	www.searchanddiscovery.com	5			
	Bing	www.osti.gov	4			
	Bing	ohiodnr.gov	3			
	Bing	www.researchgate.net	3			2
	Bing	www.uky.edu	3			



NATIONAL ENERGY

TECHNOLOGY LABORATORY

SmartSearch

9/21/2021



Latent Dirichlet Allocation – a type of NLP topic modeling – can be calculated and visualized



SmartSearch



Results of relevant resources discoveries are returned with a Similarity score (0.1 < x < 1) (Min threshold customizable)

	ېر	≥2 N≣TL SmartSear	ch ^{Enter See}	ds Sessions Se	eeds & Discoveri	es				admin	•	
Relevance Ar	nalysis											
Toggle column: Se Show 10 🗸 entr	ed Session ID Seed	Tag Seed Token	Seed Source URI S	eed File Seed Body	Discovered Session IC	Discovered Tag	Discovered Token	Discovered Source UR	Discovered File	Discovered Body Sea	imilarity arch:	
Seed Session 👌	Seed Tag 🎈	Seed Token 单	Seed Source URI	Seed File 🔶	Seed Body 🔶	Discovered Session ID	Discovered Tag	Discovered Token	Discovered Source URI	Discovered File	Discovered Body	Similarity 🔻
sida4c26dfd7	paige_test_data	f4e62d4c31a	https://edx.n	web.pdf	Volume VI.IMi	sida4c26dfd7	paige_test_data	af7578627c3	https://www	web.html	Leveraging R	0.28418
sida4c26dfd7	paige_test_data	<u>f4e62d4c31a</u>	https://edx.n	web.pdf	Volume VI.IMi	sida4c26dfd7	Paig Leveraging P United State	Regional Exploration s (Technical Report)	to Develop Geologic OSTI.GOVskip to ma	Framework for CO2 ain contentSign InCr	Storage in Deep For eate AccountU.S. De	mations in Midweste partment of Energy
sida4c26dfd7	paige_test_data	<u>f4e62d4c31a</u>	https://edx.n	web.pdf	Volume VI.IMi	sida4c26dfd7	Office of Scie	entific and Technical anced Search queries	Information Search use a traditional Te	terms: Advanced sea rm Search. For more	rch options Advance info, s	ed Search
sida4c26dfd7	paige_test_data	<u>f4e62d4c31a</u>	https://edx.n	web.pdf	Volume VI.IMi	sida4c26dfd7	paige_test_data	edceba41eea	https://www	web.html	State Charlto	0.28102
sida4c26dfd7	paige_test_data	f4e62d4c31a	https://edx.n	web.pdf	Volume VI.IMi	sida4c26dfd7	paige_test_data	<u>864054cc560</u>	https://www	web.html	State Charlto	0.28095
sida4c26dfd7	paige_test_data	<u>f4e62d4c31a</u>	https://edx.n	web.pdf	Volume VI.IMi	sida4c26dfd7	paige_test_data	af7578627c3	https://www	web.html	Leveraging R	0.27846
sida4c26dfd7	paige_test_data	f4e62d4c31a	https://edx.n	web.pdf	Volume VI.IMi	sida4c26dfd7	paige_test_data	af7578627c3	https://www	web.html	Leveraging R	0.27846
sida4c26dfd7	paige_test_data	<u>f4e62d4c31a</u>	https://edx.n	web.pdf	Volume VI.IMi	sida4c26dfd7	paige_test_data	<u>af7578627c3</u>	https://www	web.html	Leveraging R	0.27846
sida4c26dfd7	paige_test_data	f4e62d4c31a	https://edx.n	web.pdf	Volume VI.IMi	sida4c26dfd7	paige_test_data	<u>af7578627c3</u>	https://www	web.html	Leveraging R	0.27846
sida4c26dfd7	paige_test_data	f4e62d4c31a	https://edx.n	web.pdf	Volume VI.IMi	sida4c26dfd7	paige_test_data	e2d800b77fc	https://www	web.html	Leveraging R	0.27843
Showing 1 to 10 of	241 entries	-							Previous	1 2 3	4 5	25 Next

Download



Visualizations

SmartSearch offers rich visualization capabilities

spark-recommendations-strip-graph

wiki_544			•			seed_tag • wiki_544
wiki	• •		8	•	• 11	• wiki • wiki2
wiki2	· · ·	an a	•	•	•	• edx • bbq
xbe Ltag		•			•	 joetest3 paige_test_
pdd Seed		seed_tag=paige_test_data sim_tfidf=0.2841789 in=12213	•			 con_test
joetest3		seed_session_id=sida4c26dfd767c272ad9ff41ceeb122dab seed_token=t4e62d4c31ad7ec0eecbfad7c3e06f97				9
paige_test_data	80 8	seed_source_uri=https://edx.netl.doe.gov/dataset/aa360bdt-6d44-428b-bd64-fa702d9399a1/resource_download/363b8502-57c6-4d3d-839b-303e191476ab seed_file=web.pdf				
con_test		discovered_session_id=sida4c26dfd767c272ad9ff41ceeb122dab discovered_tag=paige_test_data				
	0.2	discovered_token=af7576627c3ed2a19d30f1d1b3240152 discovered_source_url=https://www.ostl.gov/biblio/979447-leveraging-regional-exploration-develop-geologic-framework-co2-storage-deep-formations-midwestern-united-states discovered_file=web_html	.9			

o Q+0000241== 🖩







Ultimately, launch SmartSearch in EDX as a **Deep Analysis** Recommendation Engine



Related Resources

Environmental benefits of advanced oil Has a 23% match
Has a 23% match
014 08:51 AM Eastern
0

Download Total: 230



Anticipate releasing SmartSearch v1 in next few months via EDX

- Perform deep contextual analysis
- Machine learning, natural language processing
- Generates correlation matches of contextually similar content
- Expanded to include spatial and webcrawl assets
- Implemented using Spark, Scala, Python, Kubernetes
- Ideal for cluster RAM, CPU, and bandwidth intensive



Building A **Big Data Ecosystem** for CS R&D Data Discovery

CS data driven research requires:

• Lots of data

J.S. DEPARTMENT OF

- Incorporating different data types & formats,
- Integrating data from multiple locations (web, local, databases)

Traditional Search methods impede our efforts:

• Search engine limits context to a few terms

- Labor intensive to conduct data searching
- Even more difficult to find relevant data



SmartSearch in Press:

- Baker, D.V., Rose, K., Bauer, J., and Rager, D., 2016, **Computational Advances and Data Analytics to Reduce Subsurface Uncertainty**, ARMA 16-493, June 26-29, 2016, 16 pgs.
- Baker, D.V., Bauer, J., and Rose, K., 2018. Developing a Smarter Way to Search Parsing the online "forest" to find data for your research needs via EDX. U.S. DOE Mastering the Subsurface through technology innovation, partnership and collaboration: Carbon Storage and Oil and Natural Gas Technologies review meeting
- Rose, K.; Bauer, J.; Baker, V.; et al., 2018, Development of an Open Global Oil and Gas Infrastructure Inventory and Geodatabase; NETL-TRS-6-2018; NETL Technical Report Series; U.S. Department of Energy, National Energy Technology Laboratory: Albany, OR, 2018; p 594; DOI: 10.18141/1427573.

https://edx.netl.doe.gov

NETL Resources

VISIT US AT: www.NETL.DOE.gov







@National Energy Technology Laboratory

vic.baker@matricinnovates.com jennifer.bauer@netl.doe.gov

https://edx.netl.doe.gov

