



Fast & Effective: Natural Language Understanding

Mike Conover, Ph.D.
Principal Data Scientist

SkipFlag

- Smart Knowledge Base
- Instant Answers
- Expert Identification
- Intelligent Bot



Smart Knowledge Base

- Entity Graph
- Projects & Jargon
- Relevant Articles
- Documentation
- Source Code

Filter By

TYPE

- All
- Notes
- Conversations
- Articles
- Events
- Apps

TOPICS

- Docker
- Salt updates (internal)
- docker configuration (internal)
- email generation (internal)
- Airflow
- docker bugs (internal)
- Amazon Web Services
- Docker images (internal)
- virtualization
- Ansible
- Docker Compose
- + Add topic

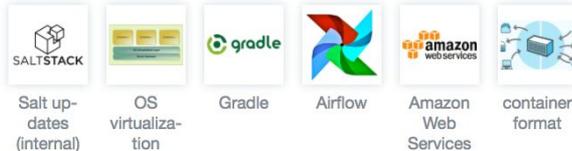
Search > Docker

Topic: Docker



Docker provides an additional layer of abstraction and automation of operating-system-level virtualization on Linux.

Related Topics



Notes

Recreating Indices for One Ad Campaign ✎ ...

  Emily Barbosa and 1 other

Aug 15, 2017 - Below is an example workflow for running a job to recreate data for a single campaign....[see more](#)

[Docker](#) [login](#) [SSH](#) [+11 more](#)

Deep Learning Prototyping Resources ✎ ...

 John Phelps

Jul 10, 2017 - A collection of resources & tutorials outlining approaches to deep learning experimentation....[see more](#)

[Amazon EC2](#) [tutorial](#) [GPU](#) [+35 more](#)

ElasticSearch Cheat Sheet ✎ ...

 Mike Chao

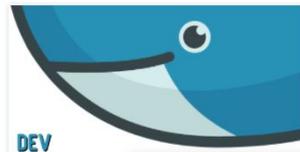
Related People

-  **Ravi Patel**
ravi
-  **Steven Tate**
Design
state
-  **Shirley Nelson**
Software Engineer
shirley

Articles



Jupyter + Tensorflow + Nvidia GPU + Docker + Google Compute Engine
[medium.com](#)



[dev.to](#)

[twitter.com](#)



Prototype Rapidly:

Or how to solve open research problems in a production environment on deadline.

Reflections



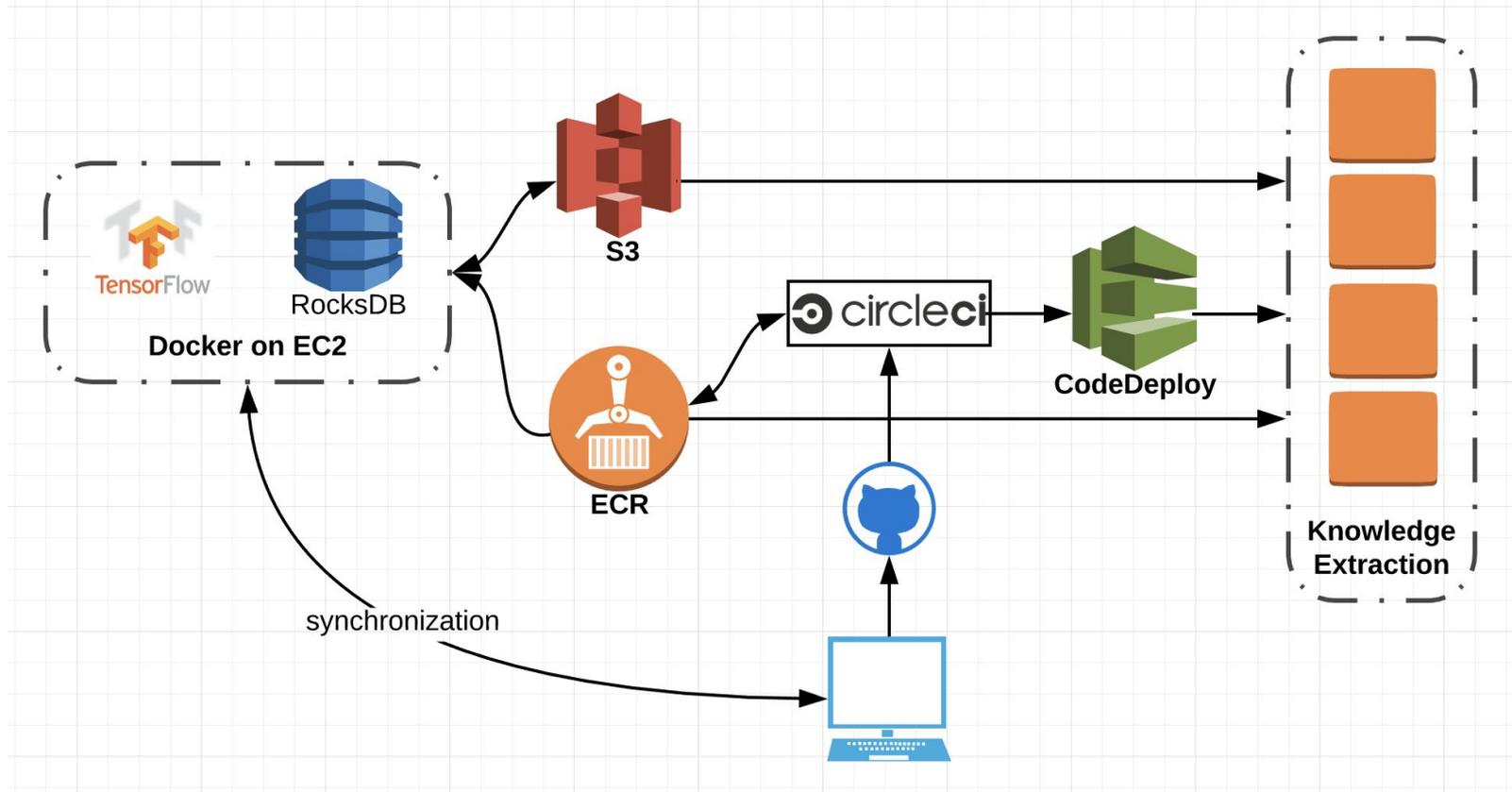
Exercise is good for you.

Reflections

Start with the model the state of the art claims
to beat and implement that.



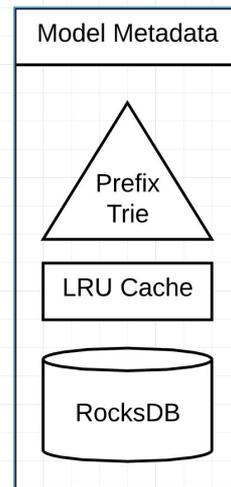
Containers & Model Deployment



Tiered Metadata Architecture



- Compute local data access
- Memory constrained environments
- Fast bulk write



Language in the Wild

Common Crawl

BIG PICTURE - THE DATA - ABOUT - BLOG - CONNECT - Donate

Us

We build and maintain an open repository of web crawl data that can be accessed and analyzed by anyone.

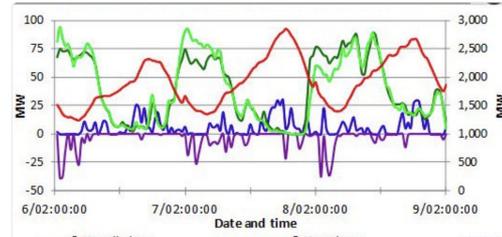
You

Need years of free web page data to help change the world.



RenewEconomy @renew_economy · Feb 22

#Tesla big battery results suggest local storage better than "monster" projects
ow.ly/ggY130iz9H1 #auspol



Tesla big battery results suggest local storage better than "monster" ...

New analysis says performance of Tesla big battery shows advantages of distributed storage rather than a single "monster" project like Snowy 2.0.

reneweconomy.com.au

Yosemite National Park

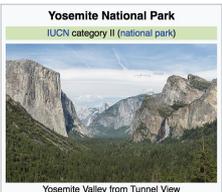
From Wikipedia, the free encyclopedia

Coordinates: 37°51'N 119°33'W

"Yosemite" redirects here. For other uses, see Yosemite (disambiguation).

Yosemite National Park (/ˈjoʊsemiti/ *yoh-SEM-i-tee*^[*]) is a United States national park lying in the western Sierra Nevada^[a] of Northern California.^[a] The park, which is managed by the U.S. National Park Service, covers an area of 747,956 acres (1,168,681 sq mi; 302,687 ha; 3,026.87 km²).^[a] Designated a World Heritage Site in 1984, Yosemite is internationally recognized for its granite cliffs, waterfalls, clear streams, giant sequoia groves, lakes, mountains, glaciers, and biological diversity.^[7] Almost 95% of the park is designated wilderness.^[a]

On average, about 4 million people visit Yosemite each year,^[8] and most spend the majority of their time in the 5.9 square miles (15 km²) of Yosemite Valley.^[7] The park set a visitation record in 2016, surpassing 5 million visitors for the first time in its history.^[9]



Common Crawl

- Petabyte Scale Web Crawl
- Available for Free on S3

Twitter

Cornucopia of Malformed Text

Wikipedia

- Linked Structured
- Taxonomic

Word Embeddings

Twitter	Wikipedia	Common Crawl
protest	occupying	protesters
taiji	occupied	ows
ows	reside	protest
boycott	places	protestors
burma	within	protests
protests	surrounded	occupying
occupygezi	adjacent	occupied
lebanon	enter	activists
activists	occupies	demonstrators
protesters	houses	protesting

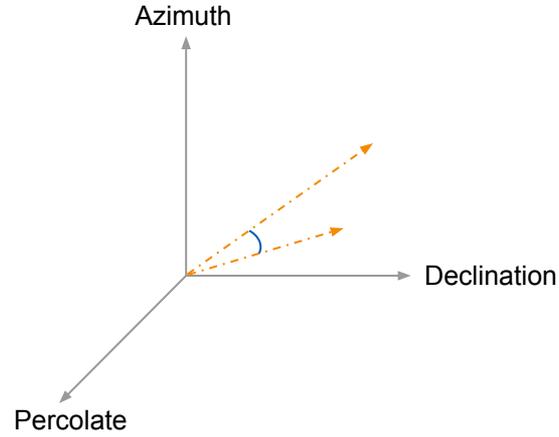
occupy



“All models are wrong, but some are useful.”

George Box

Who Needs Grammar, Anyway?



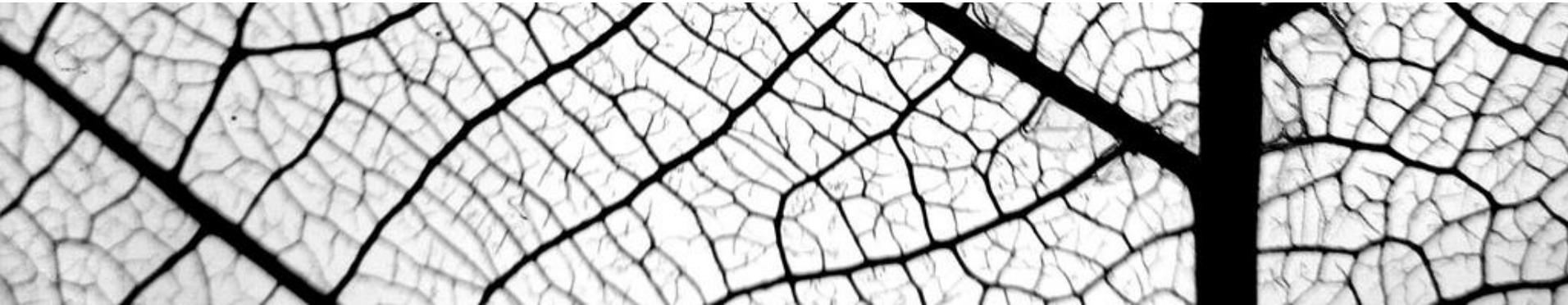
Azimuth	Declination	Percolate
.5	.9	.01

.. M's of Dimensions

↓
LSA / LDA,
etc.

Orienteering	Physics
.9	0.1

.. 100's of Dimensions

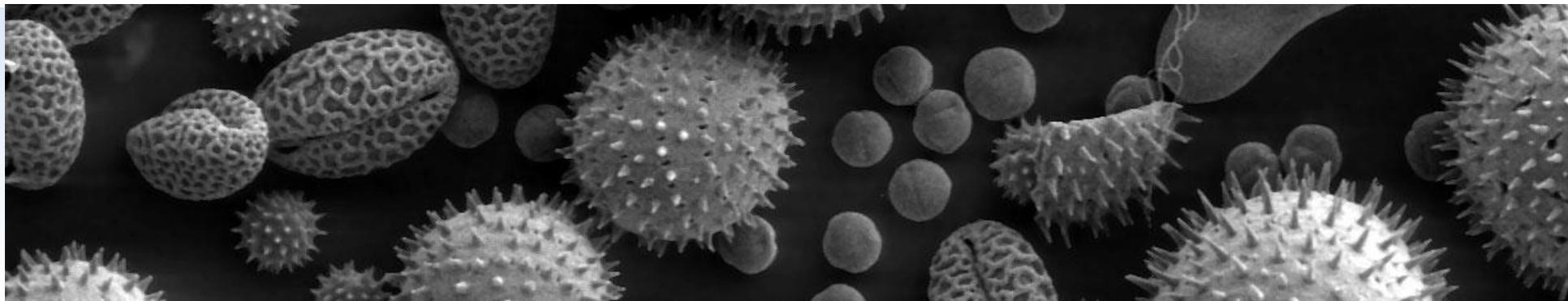
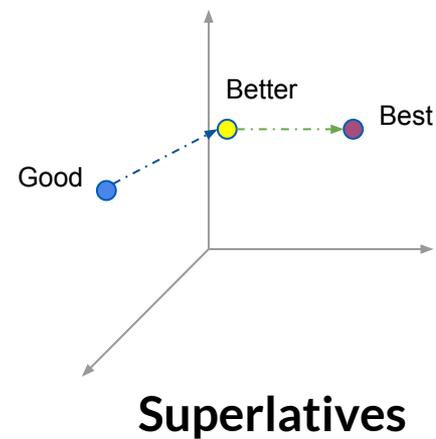
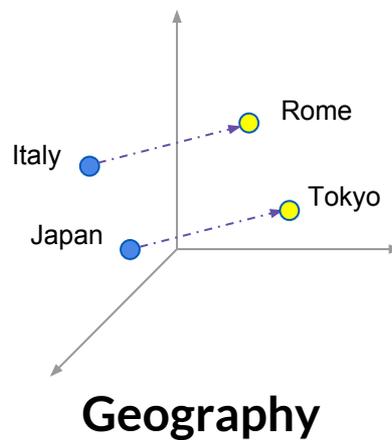
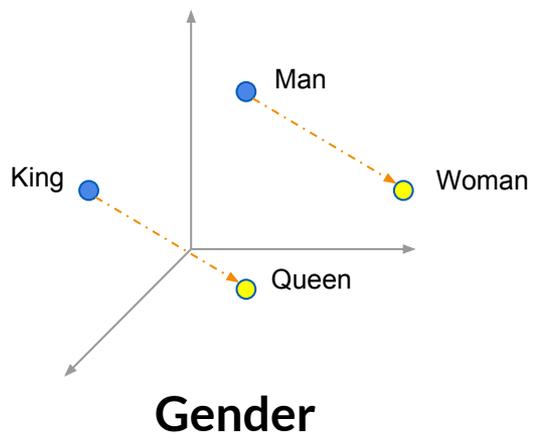


Targets of Interest

$$A = \begin{matrix} & \text{Document} \\ \begin{matrix} \text{Feature} \\ a_{11} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{m1} \end{matrix} & \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & a_{i3} & \dots & a_{in} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix} \end{matrix}$$



Semantic Structure



Embedding Vectors

The sky above the port was the color of television, tuned to a dead channel.

$$A = \begin{matrix} & \begin{matrix} \text{the} & \text{sky} & \text{above} & \dots & \text{channel} \end{matrix} \\ \begin{matrix} \text{Embedding} \\ \text{Dimension} \end{matrix} & \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & a_{i3} & \dots & a_{in} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix} \end{matrix}$$



Document Vector



Word Embeddings

Glove Vectors

Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors, 822 MB):

Common Crawl (42B tokens, 1.9M vocab, uncased, 300d vectors, 1.75 GB)

Common Crawl (840B tokens, 2.2M vocab, cased, 300d vectors, 2.03 GB)

Twitter (2B tweets, 27B tokens, 1.2M vocab, uncased, 25d, 50d, 100d, & 200d vectors, 1.42 GB)

Word2Vec

Google News (100B tokens, 3M vocab, 300d)

Freebase (100B words, 1.4M vocab, 300d)

Corpus Casing
Dimensionality Size

Build Your Own Embeddings

Out of the Box



Word2Vec

Doc2Vec

Poincare Embeddings

LDA / LSA



Tensorflow Embedding Projector

Embedding Projector

DATA | Points: 9978 | Dimension: 1024 | Selected 101 points

5 tensors found
GNMT Interlingua

Label by
label

Color by
No color map

Sphereize data

Load data | Publish

Checkpoint: Demo datasets
Metadata: oss_data/bnmt_enjako_interlingua_10k_1024_metadata.tsv

T-SNE | PCA | CUSTOM

Dimension 2D 3D

Perplexity 25

Learning rate 10

Re-run | Stop

Iteration: 164

[How to use t-SNE effectively.](#)

[17] With the advent of the Internet , the number of home work has increased dramatically .

label	[17] With the advent of the Internet , the number of home work has increased dramatically .
index	39
srclang	ja
tarlang	en
translation	With the advent of the Internet , the number of home work has increased dramatically .
source	インターネットの登場により、在宅の仕事の数が飛躍的に増大しました。
srctar	jaen
bleu	0.315355420113
bleu_summary	metric=bleu 1g=9/16 2g=6/15 3g=4/14 4g=2/13 prec=0.315355 r=15 c=16 BP=1.000000 Bleu=0.315355
gleu	0.36206895113
gleu_summary	num_sentences:1 metric:gleu 1g=9/16 2g=6/15 3g=4/14 4g=2/13 Gleu=0.36206895 AverageSentenceScore=0.36206895

[15] インターネットの普及により、ホームビジネスが大幅に拡大しました。
[18] With the advent of the Internet , the number of home work has increased dramatically .
[14] インターネットの普及により、ホームビジネスが大幅に拡大しました。
[14] 인터넷의 등장으로 수많은 가정용 업무가 급격히 증가하였습니다.
[12] 인터넷의 등장으로 수많은 가정용 업무가 급격히 증가하였습니다.
[17] With the advent of the Internet , the number of home work has increased dramatically .
[16] With the advent of the Internet , the number of home work has increased dramatically .

Text
Images
Music
..
Get Crazy

Compositional Embeddings

Domain Specific Corpora

Initialize with Pre-trained Embeddings



WASHIP

Cut to the Chase

trichlorodifluorene

FastText

- Multiclass Classification
- Subword Embeddings

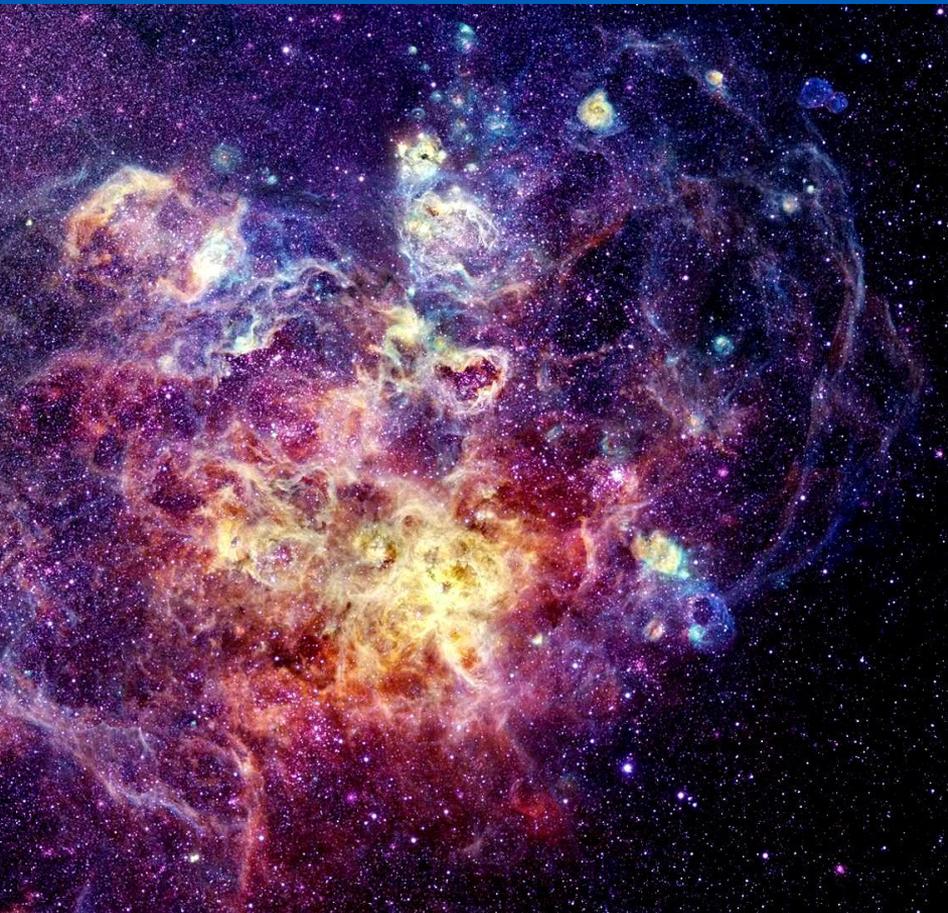


$$s(w, c) = \sum_{g \in \mathcal{G}_w} \mathbf{z}_g^\top \mathbf{v}_c.$$

<https://github.com/facebookresearch/fastText>
Bojanowski, Piotr, et al. "Enriching word vectors with subword information." *arXiv* (2016)



Embed All the Things!



StarSpace

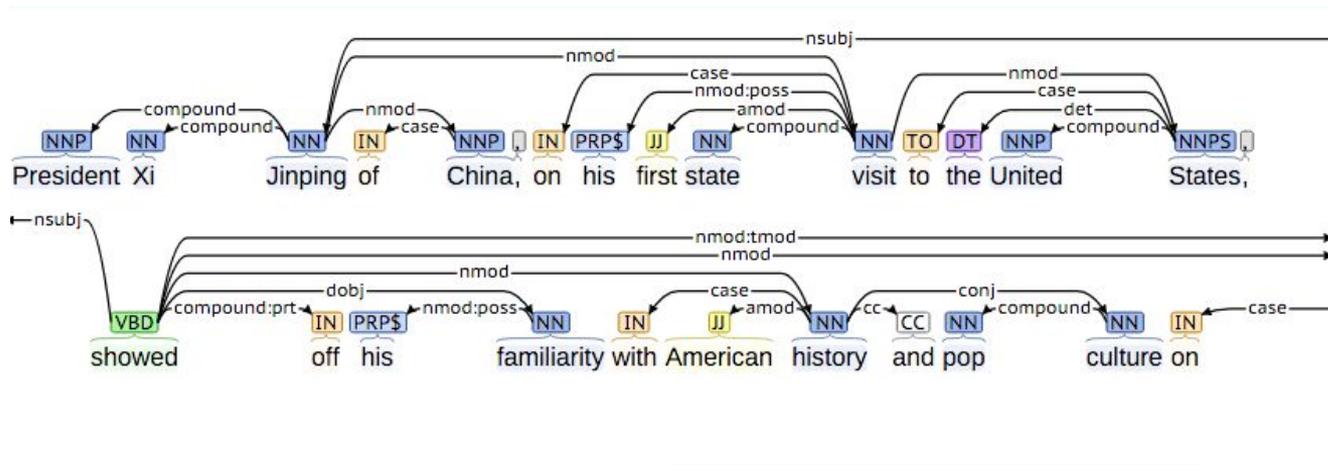
- Text Classification
- Graph Embeddings
- Similarity / Ranking
- Image Classification

Fine-Grained Structure

spaCy

Graham Askew PERSON, a biomechanics professor at the University of Leeds ORG in England GPE, leads research to understand better how the chambered nautilus moves.

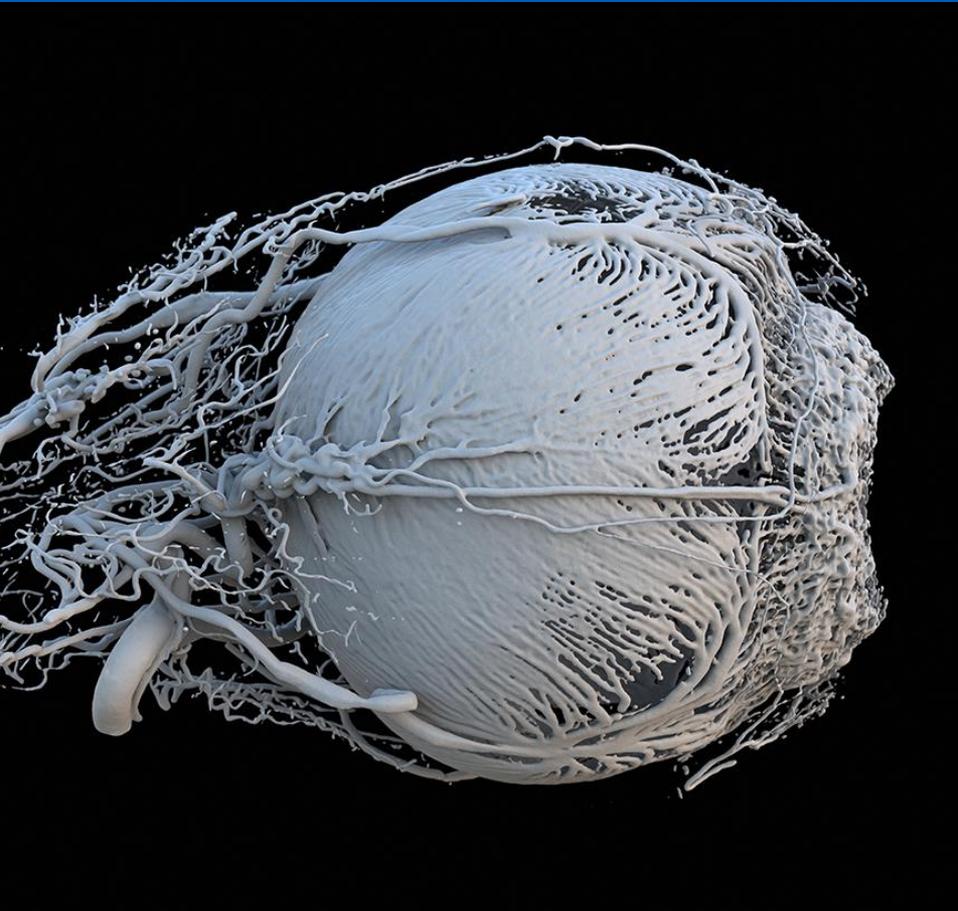
Breakdown



🏠 Stanford CoreNLP

graham askew, a biomechanics professor at the university of leeds in england, leads research to understand better how the chambered nautilus moves.

Piece by Piece



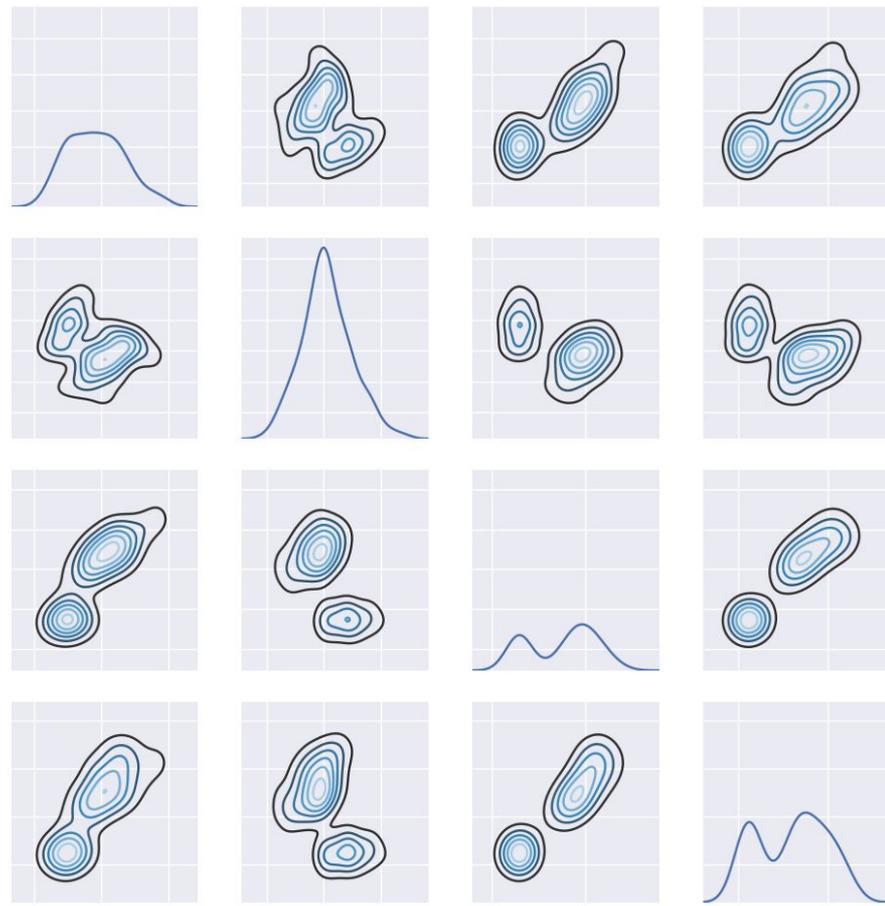
Keyphrase Extraction

- RAKE Algorithm
- Segphrase / Autophrase

graham_askew | a | biomechanics_professor | at
the | university_of_leeds | in | england | leads
research | to | understand | better | how | the |
chambered_nautilus | moves

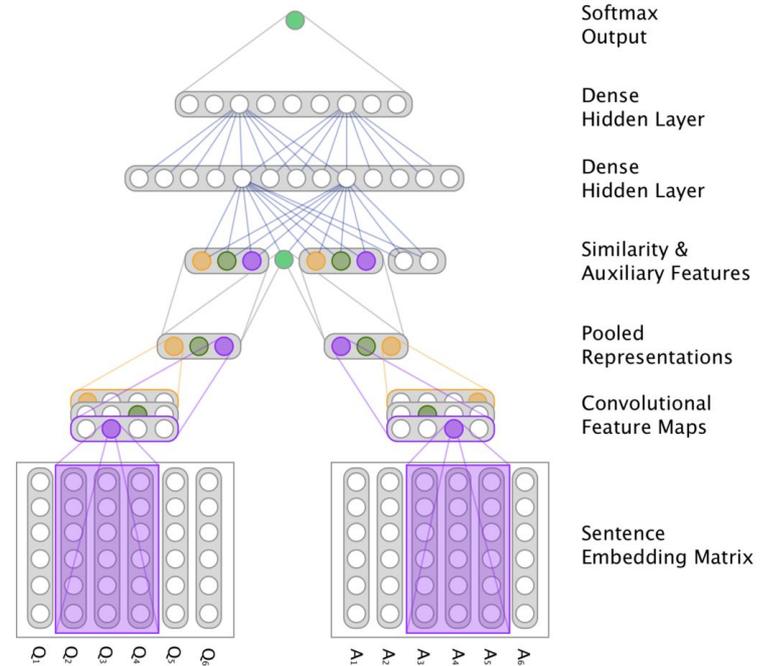
Taking Sentences Apart

Zeroth Law: This only works
in practice, never in theory.



Learning to Rank with Neural Nets

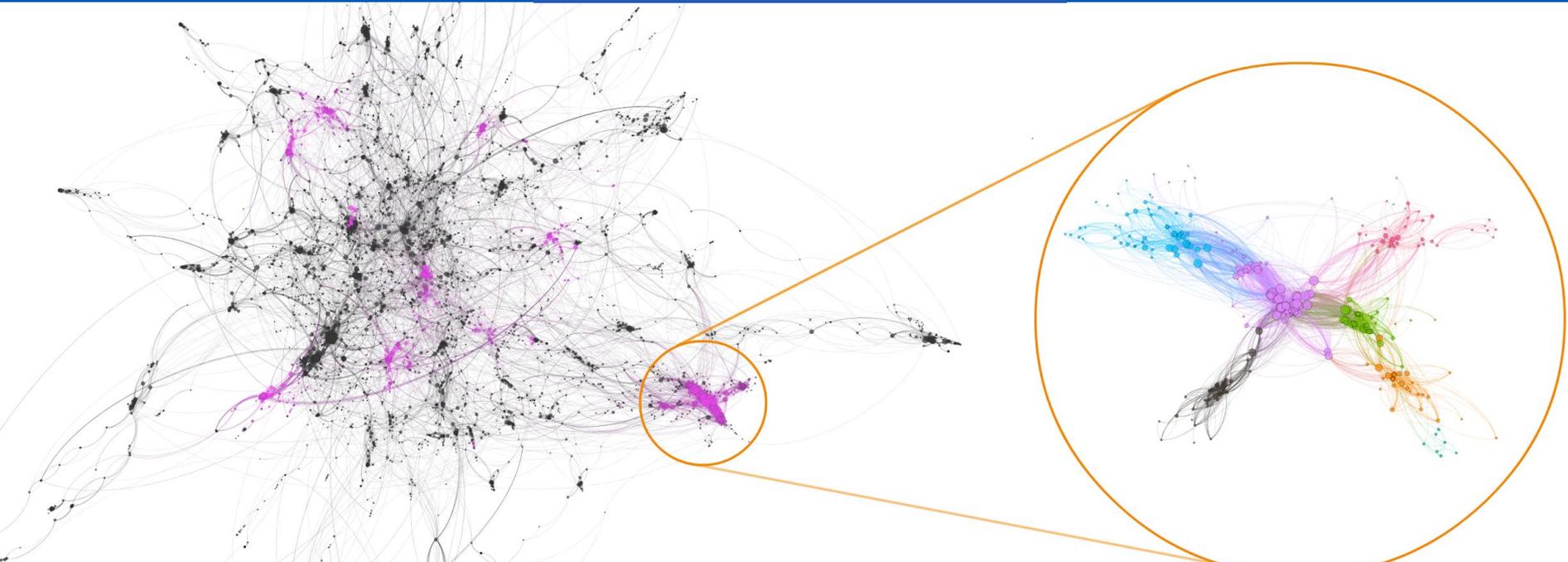
Sometimes Good Enough Isn't Good Enough



Severyn, Aliaksei, and Alessandro Moschitti. "Learning to rank short text pairs with convolutional deep neural networks." *SIGIR* 2015.

<http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/>

workday®

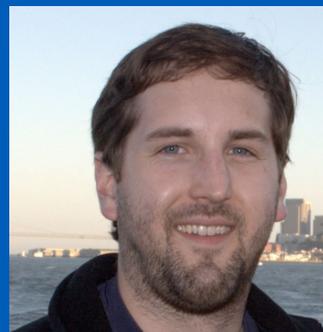




Pete
Skomoroch



Sam
Shah



Scott
Blackburn



Matt
Hayes

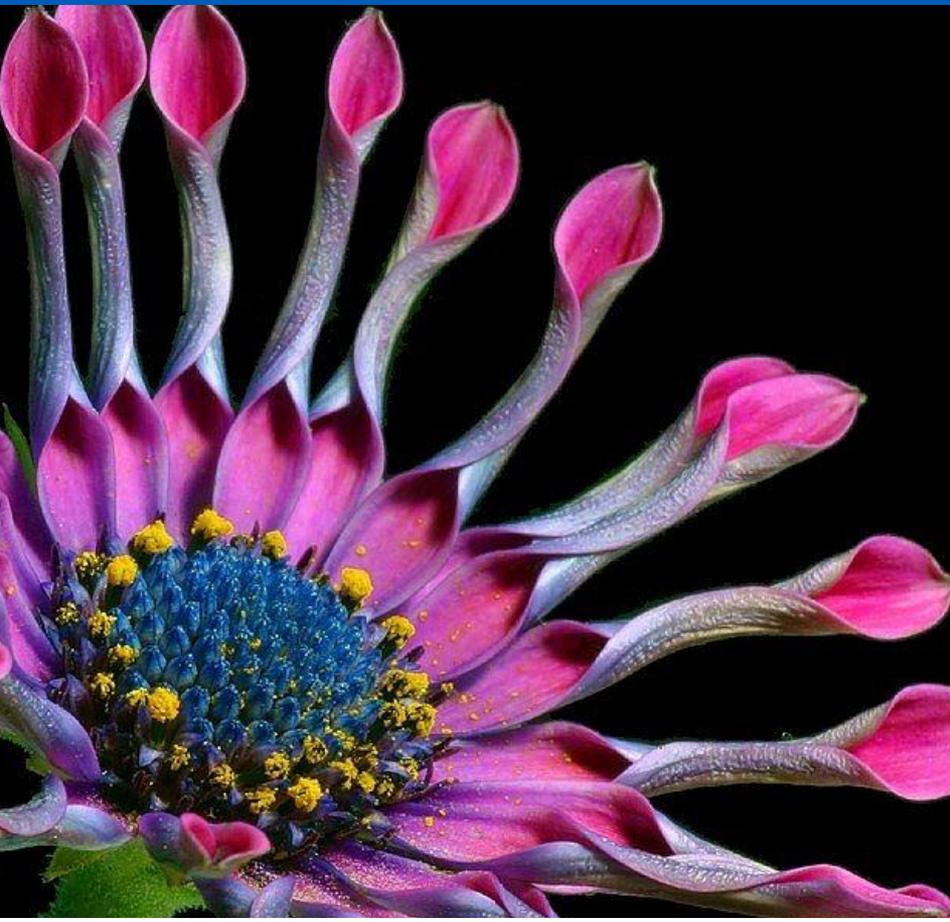




Fast & Effective: Natural Language Understanding

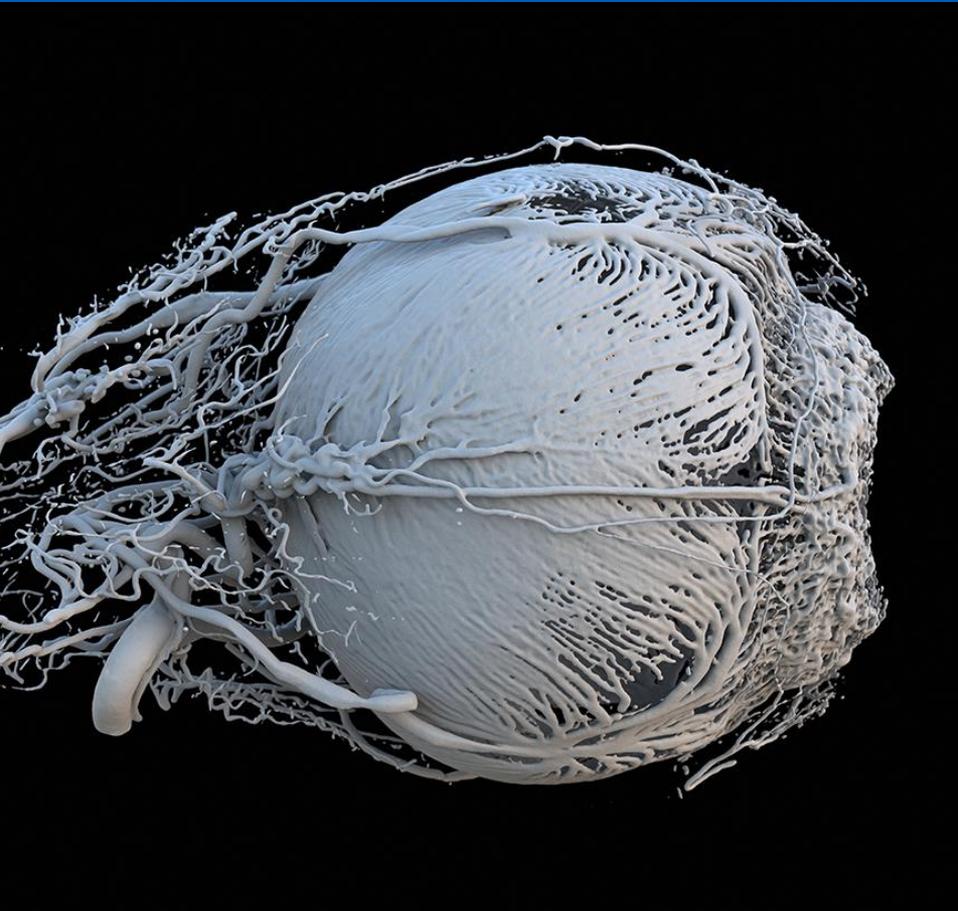
Mike Conover, Ph.D.
Principal Data Scientist

Cut to the Chase



Emoji Space

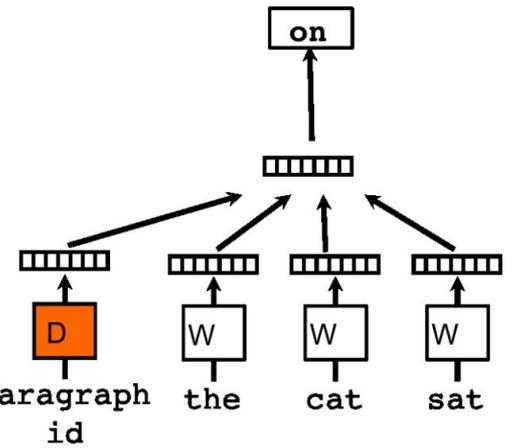
Build Your Own Embeddings



Classifier

Average/Concatenate

Paragraph Matrix



Paragraph Vectors (Doc2Vec)

Ship It!

1

Prototype (2 Weeks)

Literature Review
Operationalization
Strawman Models

2

Productionize (2 Weeks)

Schedule
Service
Profiling

3

Harden (Ongoing)

Kaggle Challenge
Compute Footprint



Filter By

TYPE

Q All

Notes

Conversations

Articles

Events

Apps

TOPICS

Amazon Web Services (328)

Data (91)

Amazon.com (88)

Slack (79)

web service (74)

machine (68)

Python (48)

application (37)

Search >

⚡ Instant Answer



Is this helpful?



Matthew Hayes and 2 others

Last Modified: Jun 15, 2017 2:50 PM

Adding a new EBS volume in AWS

How to add a new EBS volume in AWS.

1. Create the new EBS volume in AWS. After it's created attach it to the instance you want.
2. Log into the machine. You should see the new device when you run `cat /proc/partitions`. You can run `mount` to compare against what is already mounted.
3. Format the device with a filesystem using `sudo mkfs -t ext4`

[See More](#)

Topics

Amazon Web Services

machine

file system

volume

directory

+1 more