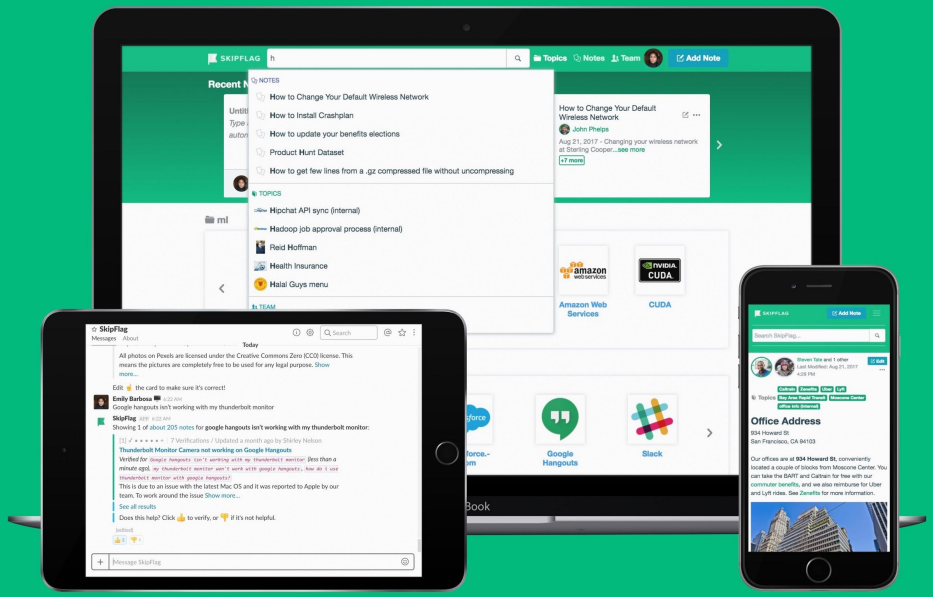
Fast & Effective: Natural Language Understanding

Mike Conover, Ph.D.
Principal Data Scientist

SkipFlag



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

workday.

slack Atlassian HipChat Outlook Google Drive JIRA zendesk

Confluence GitHub

Smart Knowledge Base

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

SKIPFLAG | Search SkipFlag... | Topics | Notes | Team | Add Note

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

Salt up-dates (internal) | OS virtualiza-tion | Gradle | Airflow | Amazon Web Services | container format

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

Mikes monster list of docker tips
dev.to

twitter.com

Prototype Rapidly:

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

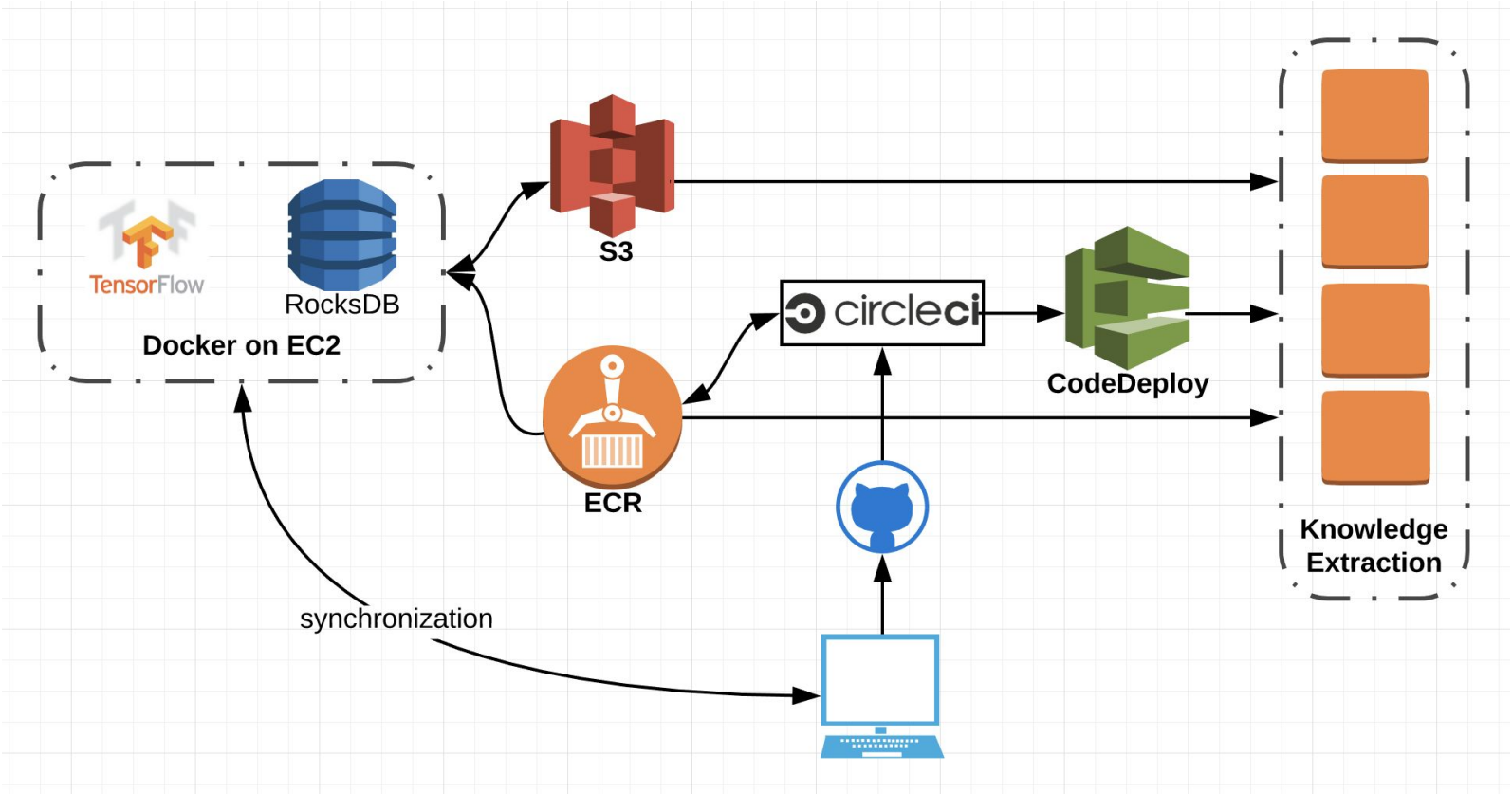
Reflections

Exercise is good for you.

Reflections

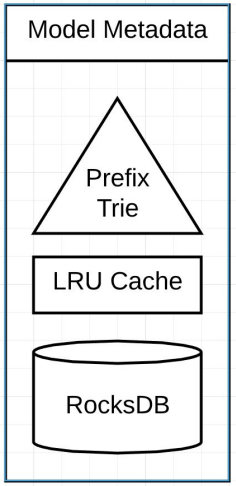
Start with the model the state of the art claims to beat and implement that.

Containers & Model Deployment

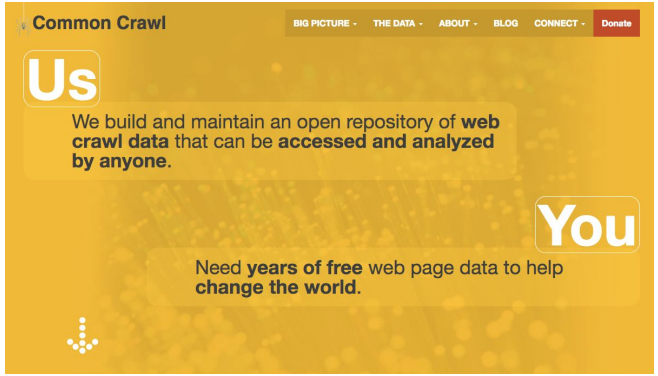


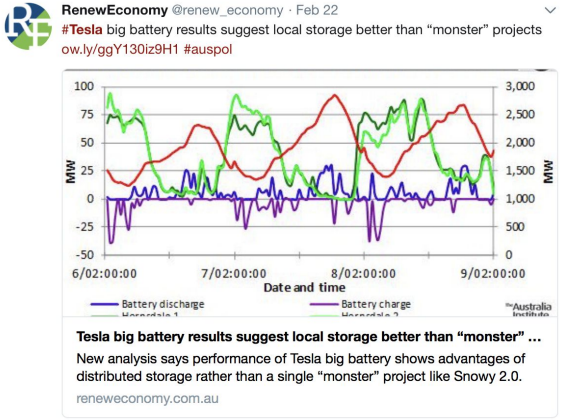
Tiered Metadata Architecture

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



Language in the Wild





Common Crawl

- Petabyte Scale Web Crawl
- Available for Free on S3

Twitter

Cornucopia of Malformed Text

Wikipedia

- Linked Structured
- Taxonomic

Word Embeddings

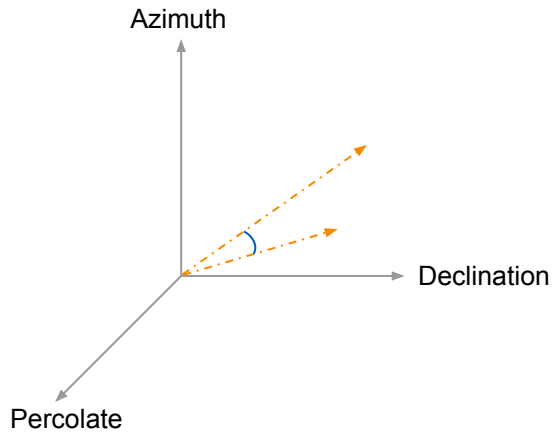
Twitter	Wikipedia	Common Crawl
protest	occupying	protesters
taiji	occupied	ows
ows	reside	protest
boycott	places	protestors
burma	within	protests
protests	surrounded	occupying
occupygezi	adjacent	occupied
lebanon	enter	activists
activists	occupies	demonstrators
protesters	houses	protesting

occupy

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

George Box

Who Needs Grammar, Anyway?



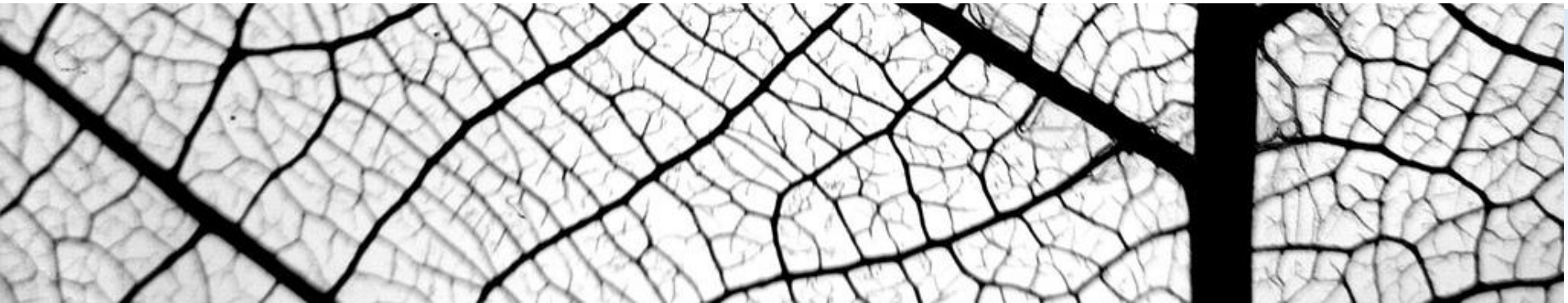
Azimuth	Declination	Percolate
.5	.9	.01

.. M's of Dimensions

LSA / LDA, etc.

Orienteering	Physics
.9	0.1

.. 100's of Dimensions

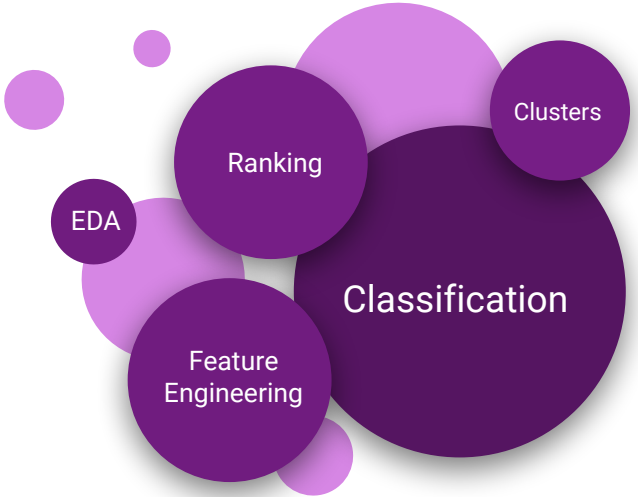


Targets of Interest

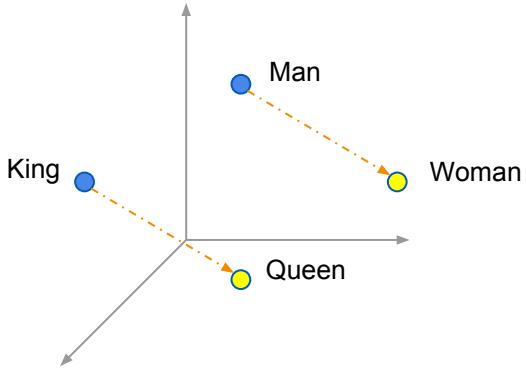
Document

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

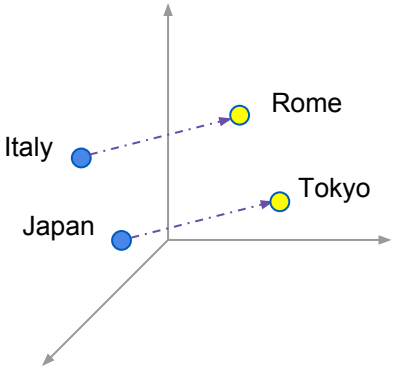
Feature



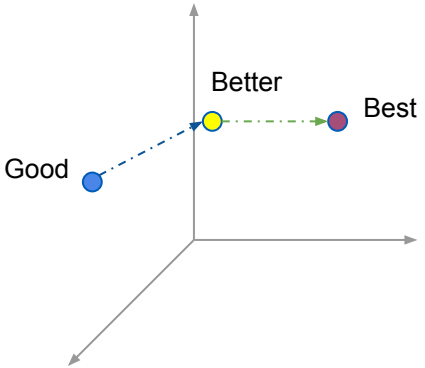
Semantic Structure



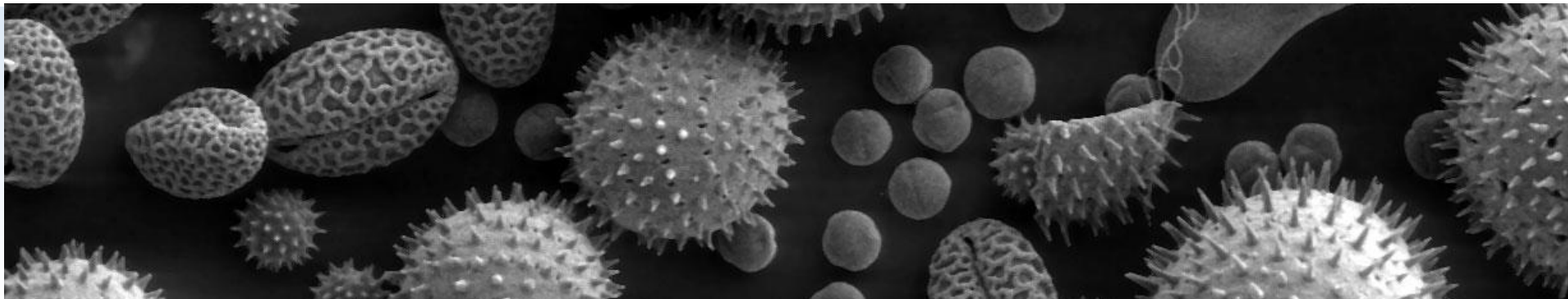
Gender



Geography

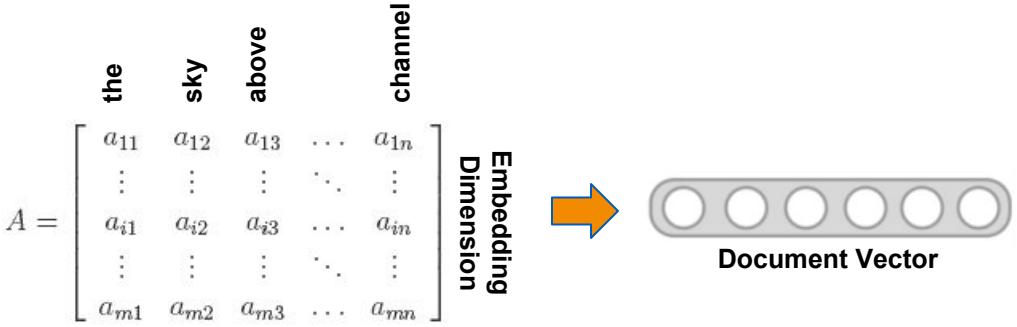


Superlatives



Embedding Vectors

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



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

https://nlp.stanford.edu/projects/glove/

Build Your Own Embeddings

Out of the Box

gensim
topic modelling for humans

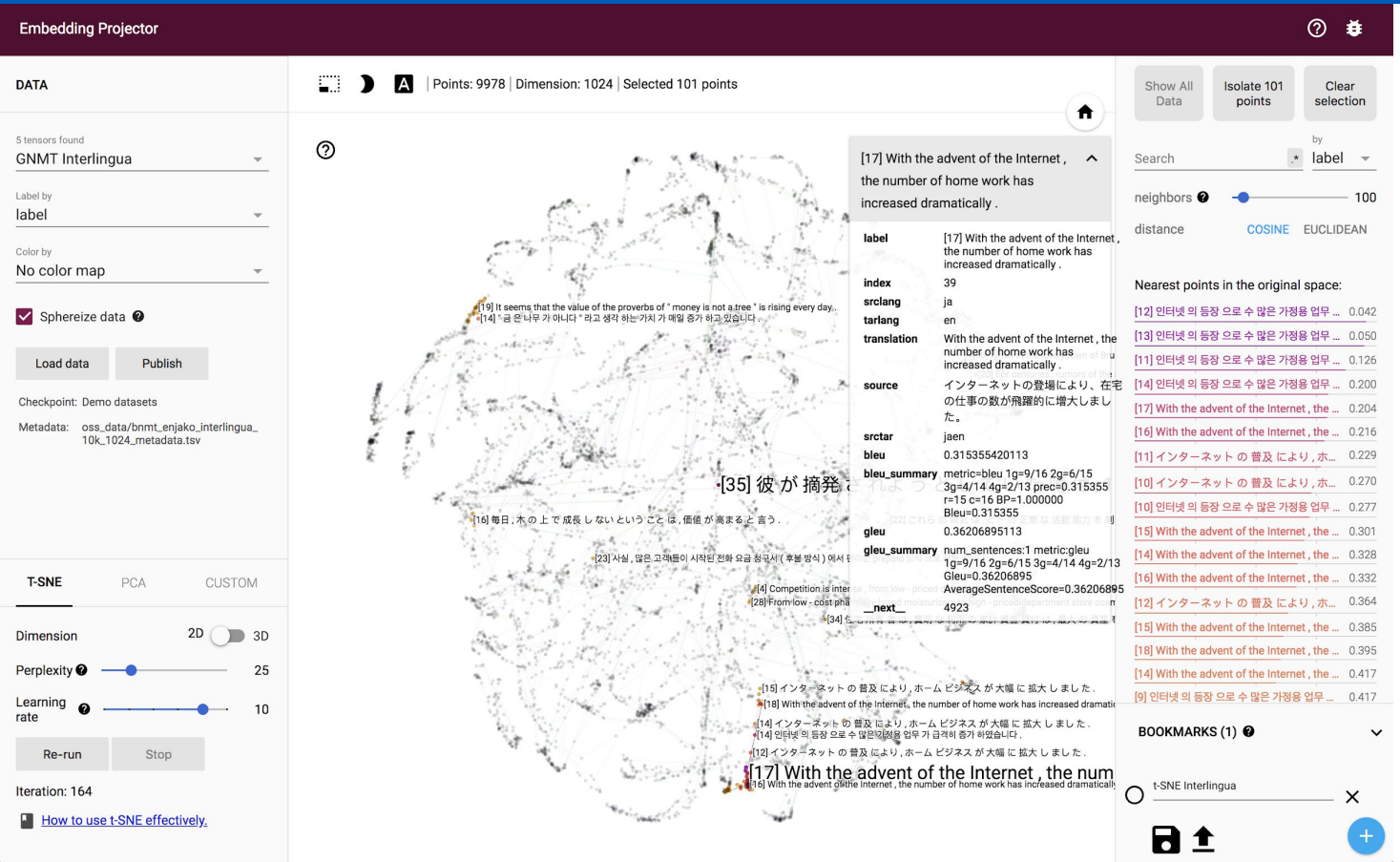
Word2Vec

Doc2Vec

Poincare Embeddings

LDA / LSA

Tensorflow Embedding Projector



Text

Images

Music

..

Get Crazy

Compositional Embeddings

Domain Specific Corpora

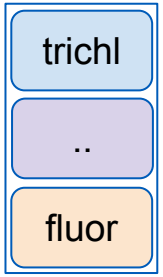
Initialize with Pre-trained Embeddings

Cut to the Chase

FastText

- Multiclass Classification
- Subword Embeddings

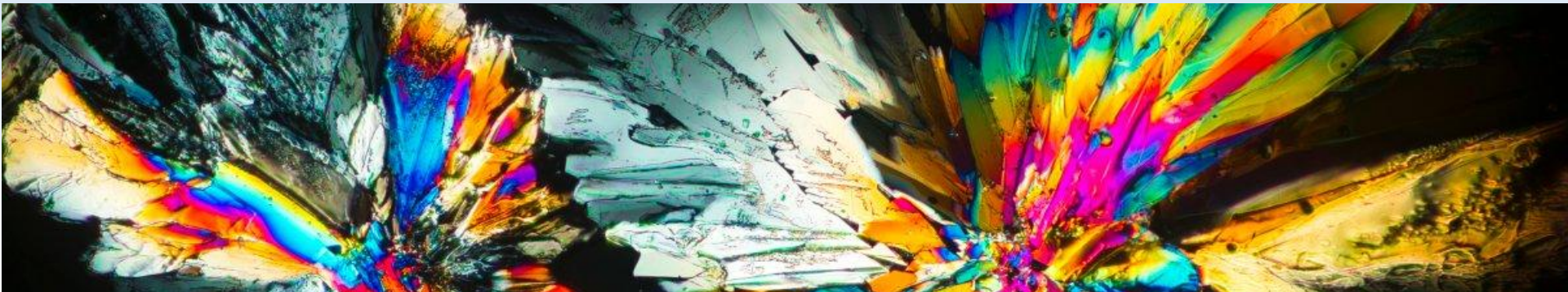
trichlorodifluorene



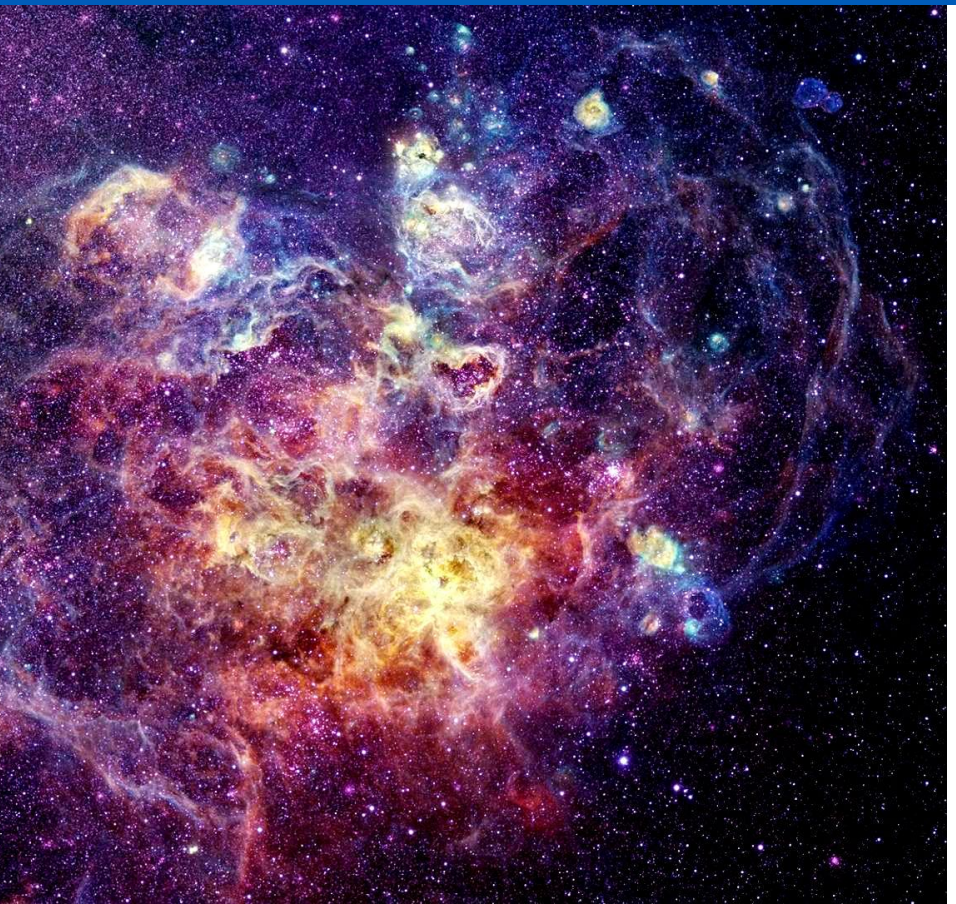
$$s(w, c) = \sum_{g \in \mathcal{G}_w} \mathbf{z}_g^\top \mathbf{v}_c.$$

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

https://github.com/facebookresearch/StarSpace

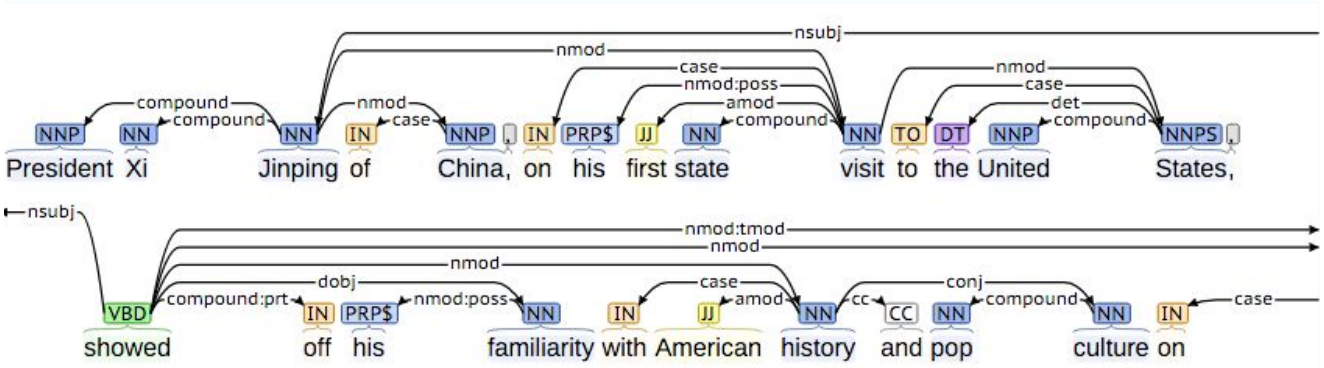
L. Wu "StarSpace: Embed All The Things!." *arXiv* (2017)

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

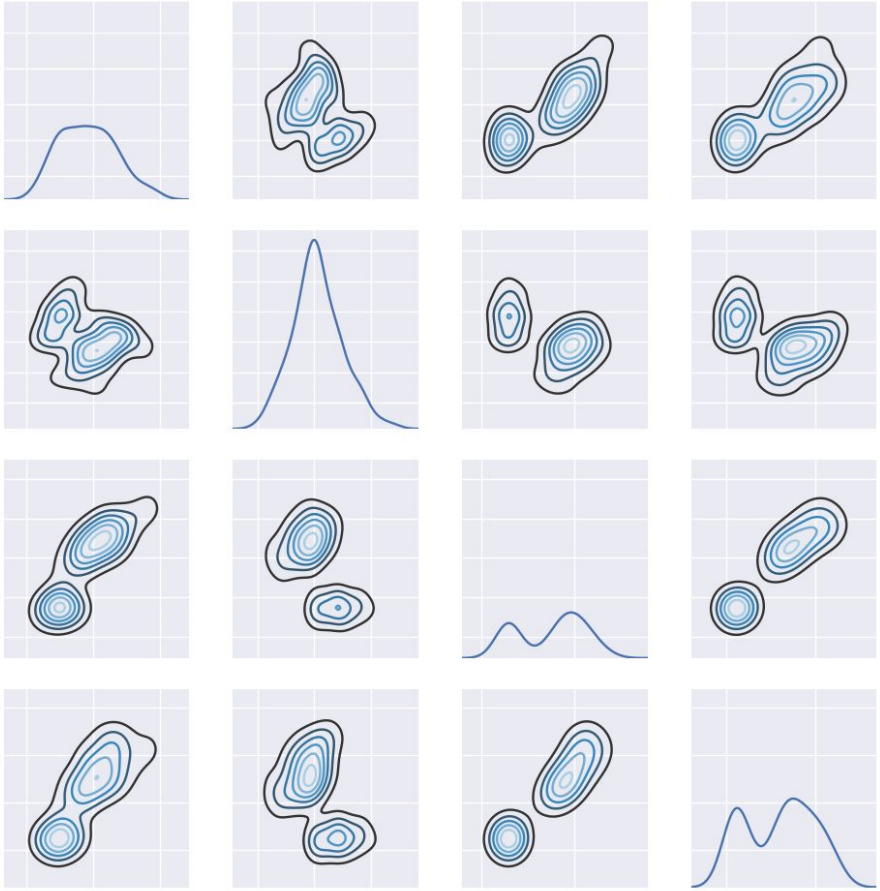
- RAKE Algorithm
- Segphrase / Autophrase

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

F. Diaz. "Query expansion with locally-trained word embeddings." *arXiv* (2016)

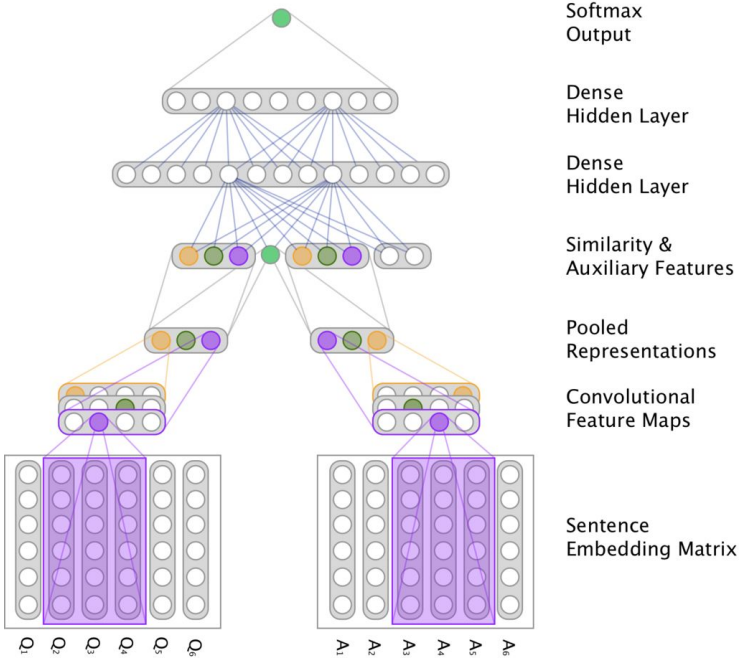
Taking Sentences Apart

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



Learning to Rank with Neural Nets

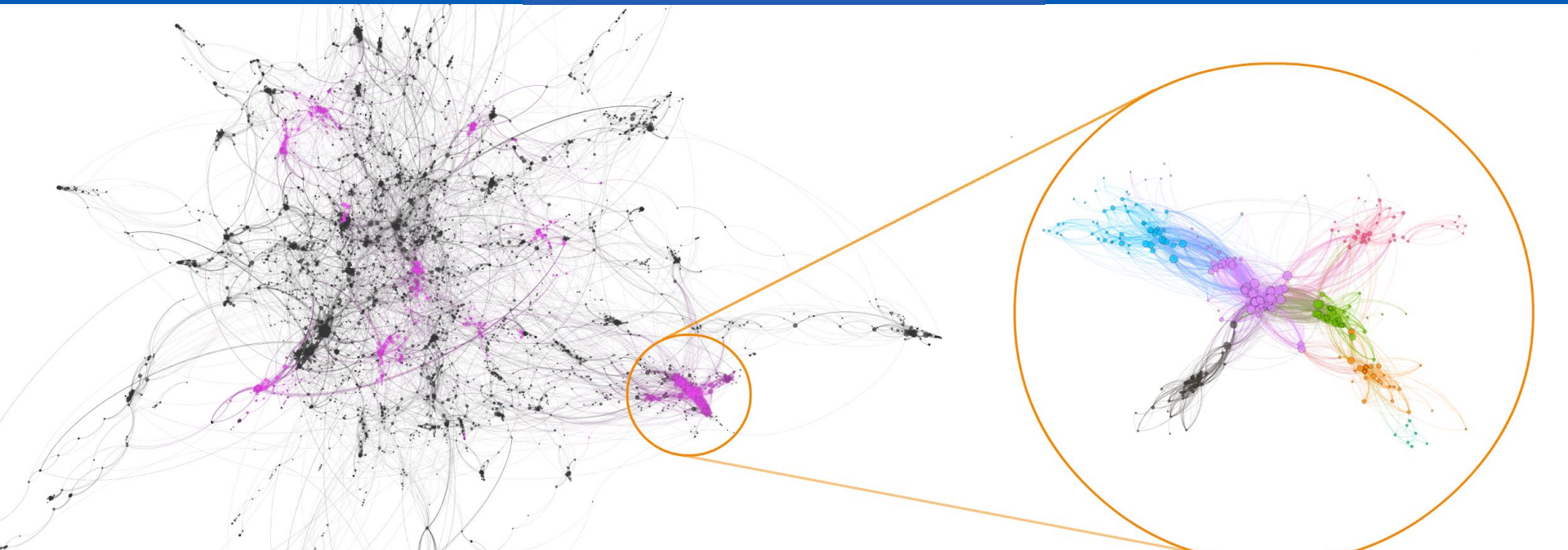
Sometimes Good Enough Isn't Good Enough



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

workday®

Fast & Effective: Natural Language Understanding

Mike Conover, Ph.D.
Principal Data Scientist

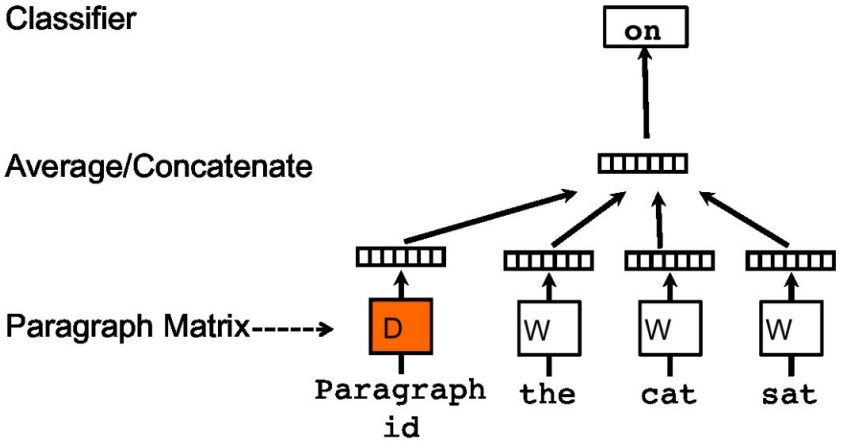
Cut to the Chase

Emoji Space

https://github.com/facebookresearch/fastText

P.Bojanowski, "Enriching word vectors with subword information." *arXiv* (2016)

Build Your Own Embeddings



Paragraph Vectors (Doc2Vec)

Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." *International Conference on Machine Learning*. 2014.

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

Instant Answers

SKIPFLAG | How do I mount a disk on an EC2 machine? | Topics

Filter By

TYPE

- 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