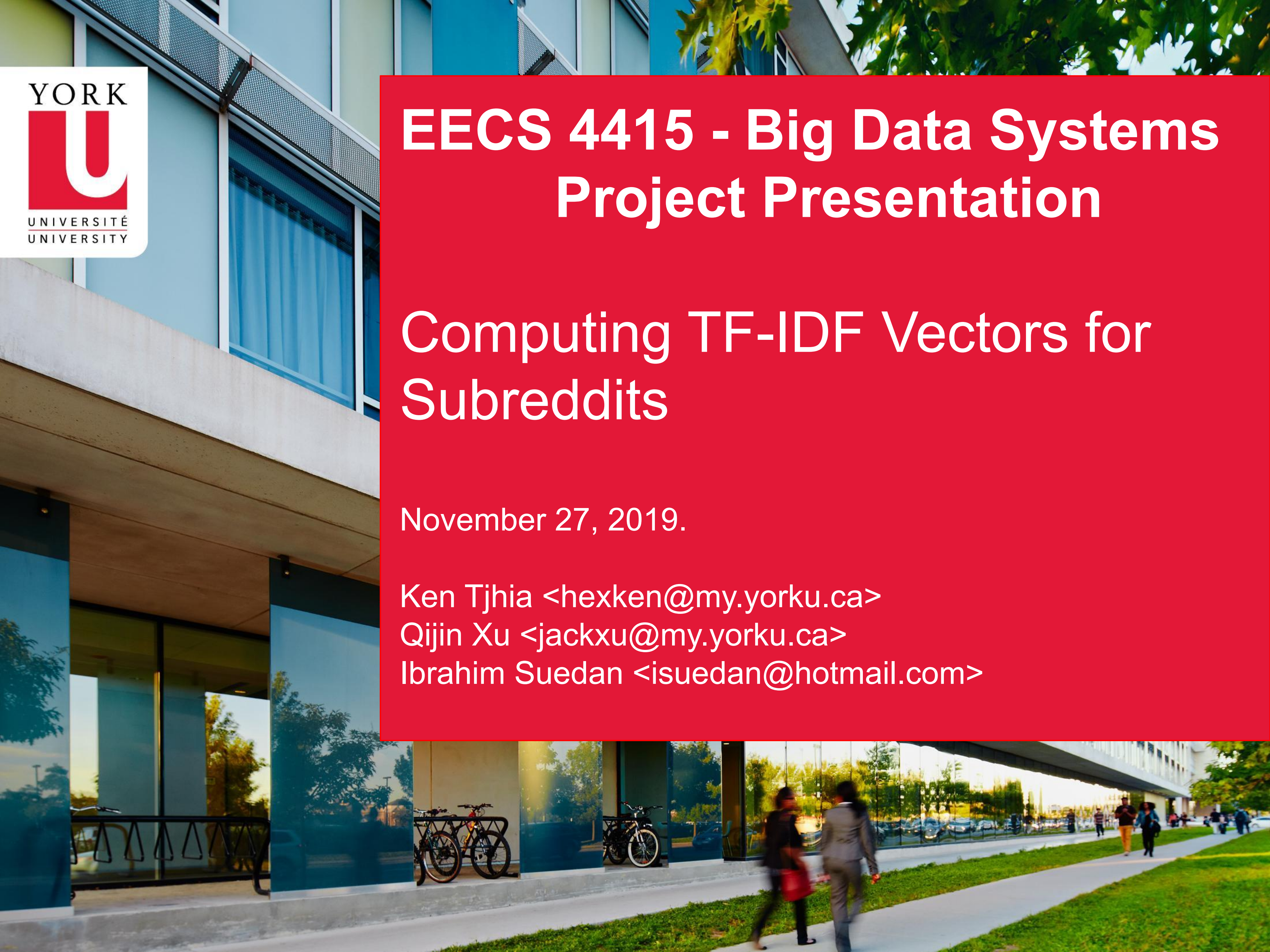


EEECS 4415 - Big Data Systems Project Presentation

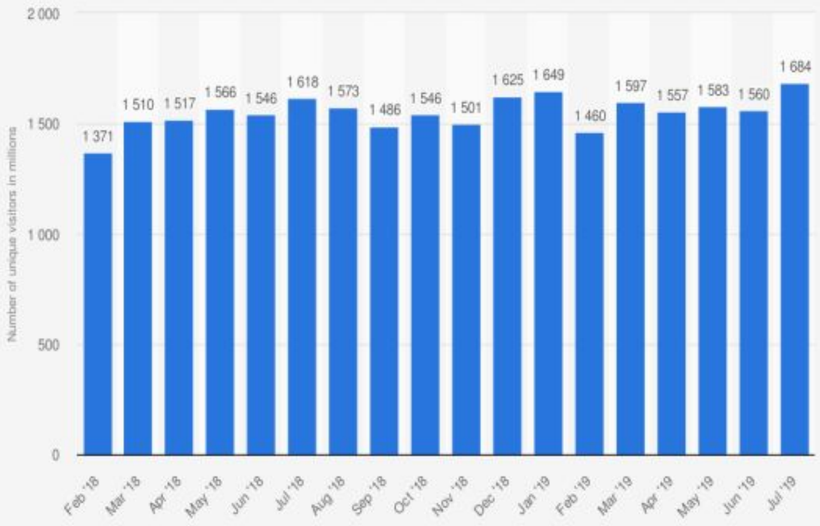
Computing TF-IDF Vectors for Subreddits

November 27, 2019.

Ken Tjhia <hexken@my.yorku.ca>
Qijin Xu <jackxu@my.yorku.ca>
Ibrahim Suedan <isuedan@hotmail.com>

Motivation

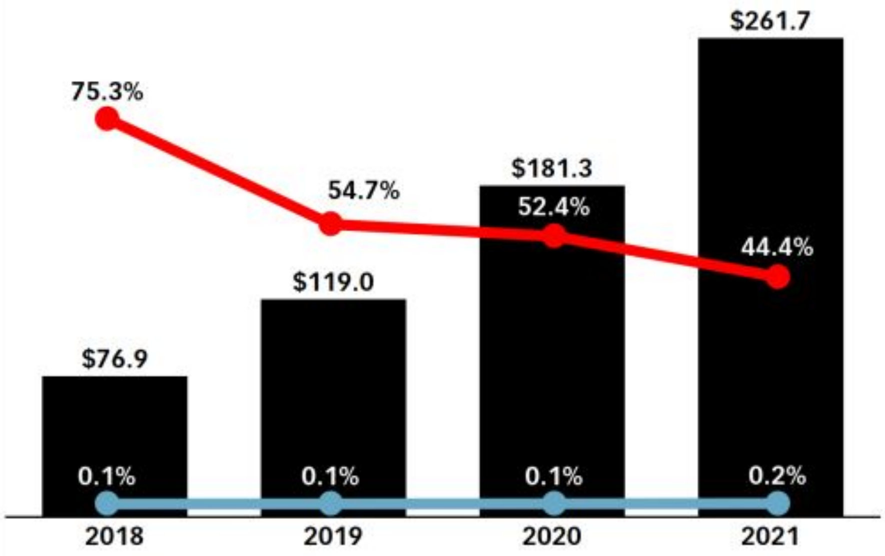
Combined desktop and mobile visits to Reddit.com from February 2018 to July 2019 (in millions)



Source
SimilarWeb
© Statista 2019

Additional Information:
Worldwide; SimilarWeb; February 2019 to July 2019; on desktop & mobile web, excluding app

Reddit Net Ad Revenues in the US, 2018-2021
millions, % change and % of total digital ad spending



■ reddit net ad revenues
🟥 % change ■ % of total digital ad spending

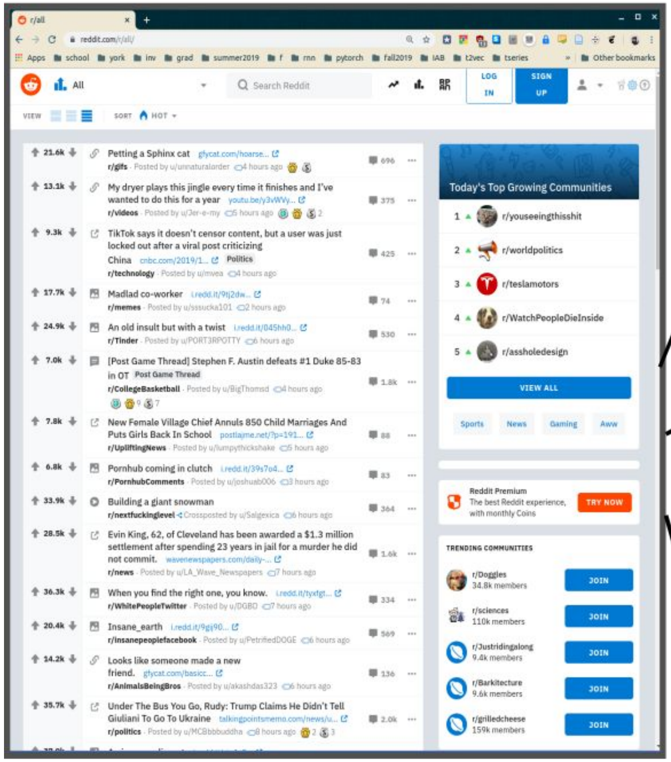
Note: includes advertising that appears on desktop and laptop computers as well as mobile phones, tablets and other internet-connected devices, and includes all the various formats of advertising on those platforms; net ad revenues after companies pay traffic acquisition costs (TAC) to partner sites; excludes nonadvertising revenues (e.g., Reddit Premium, Reddit Coins)
Source: eMarketer, February 2019

T10128 www.eMarketer.com

- Almost 1.7 Billion unique visitors and over 145 Million comments in July 2019!
- $261 Million projected spending on digital ads in 2021.

Organization of Reddit

Reddit Subreddits Posts Comments


r/all
reddit.com/r/all/
Apps school york inv grad summer2019 f rnn pytorch fall2019 IAB t2vec tseries Other bookmarks
All
Search Reddit
LOG IN
SIGN UP
VIEW
SORT HOT

21.6k Petting a Sphinx cat gfycat.com/hoarse...
r/gifs · Posted by u/unnaturalorder 4 hours ago · 696

13.1k My dryer plays this jingle every time it finishes and I've wanted to do this for a year youtu.be/y3vWVy...
r/videos · Posted by u/Jer-e-my 5 hours ago 2 · 375

9.3k TikTok says it doesn't censor content, but a user was just locked out after a viral post criticizing China cnbc.com/2019/1... Politics
r/technology · Posted by u/mvea 4 hours ago · 425

17.7k Madlad co-worker i.redd.it/9tj2dw...
r/memes · Posted by u/sssucka101 2 hours ago · 74

24.9k An old insult but with a twist i.redd.it/045hh0...
r/Tinder · Posted by u/PORT3RPOTTY 6 hours ago · 530

7.0k [Post Game Thread] Stephen F. Austin defeats #1 Duke 85-83 in OT Post Game Thread
r/CollegeBasketball · Posted by u/BigThomsd 4 hours ago 9 7 · 1.8k

7.8k New Female Village Chief Annuls 850 Child Marriages And Puts Girls Back In School postlajme.net/?p=191...
r/UpliftingNews · Posted by u/lumpythickshake 5 hours ago · 88

6.8k Pornhub coming in clutch i.redd.it/39s7o4...
r/PornhubComments · Posted by u/joshuab006 3 hours ago · 83

33.9k Building a giant snowman
r/nextfuckinglevel · Crossposted by u/Salgexica 6 hours ago · 364

28.5k Evin King, 62, of Cleveland has been awarded a $1.3 million settlement after spending 23 years in jail for a murder he did not commit. wavenewspapers.com/daily-...
r/news · Posted by u/LA_Wave_Newspapers 7 hours ago · 1.6k

36.3k When you find the right one, you know. i.redd.it/tyxfgt...
r/WhitePeopleTwitter · Posted by u/DGBD 7 hours ago · 334

20.4k Insane_earth i.redd.it/9gj90...
r/insanepeoplefacebook · Posted by u/PetrifiedDOGE 6 hours ago · 569

14.2k Looks like someone made a new friend. gfycat.com/basicc...
r/AnimalsBeingBros · Posted by u/akashdas323 6 hours ago · 136

35.7k Under The Bus You Go, Rudy: Trump Claims He Didn't Tell Giuliani To Go To Ukraine talkingpointsmemo.com/news/u...
r/politics · Posted by u/MCBbbbuddha 8 hours ago 2 3 · 2.0k

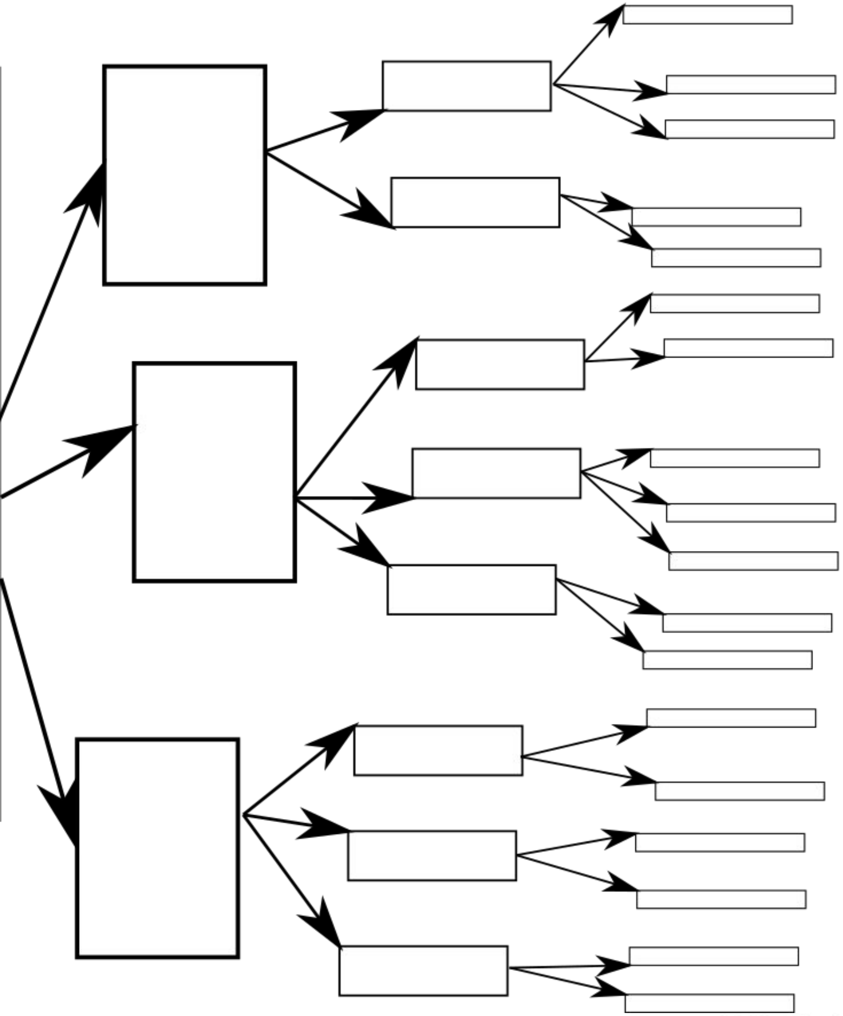
Today's Top Growing Communities

1 r/youseeingthisshit
2 r/worldpolitics
3 r/teslamotors
4 r/WatchPeopleDieInside
5 r/assholedesign

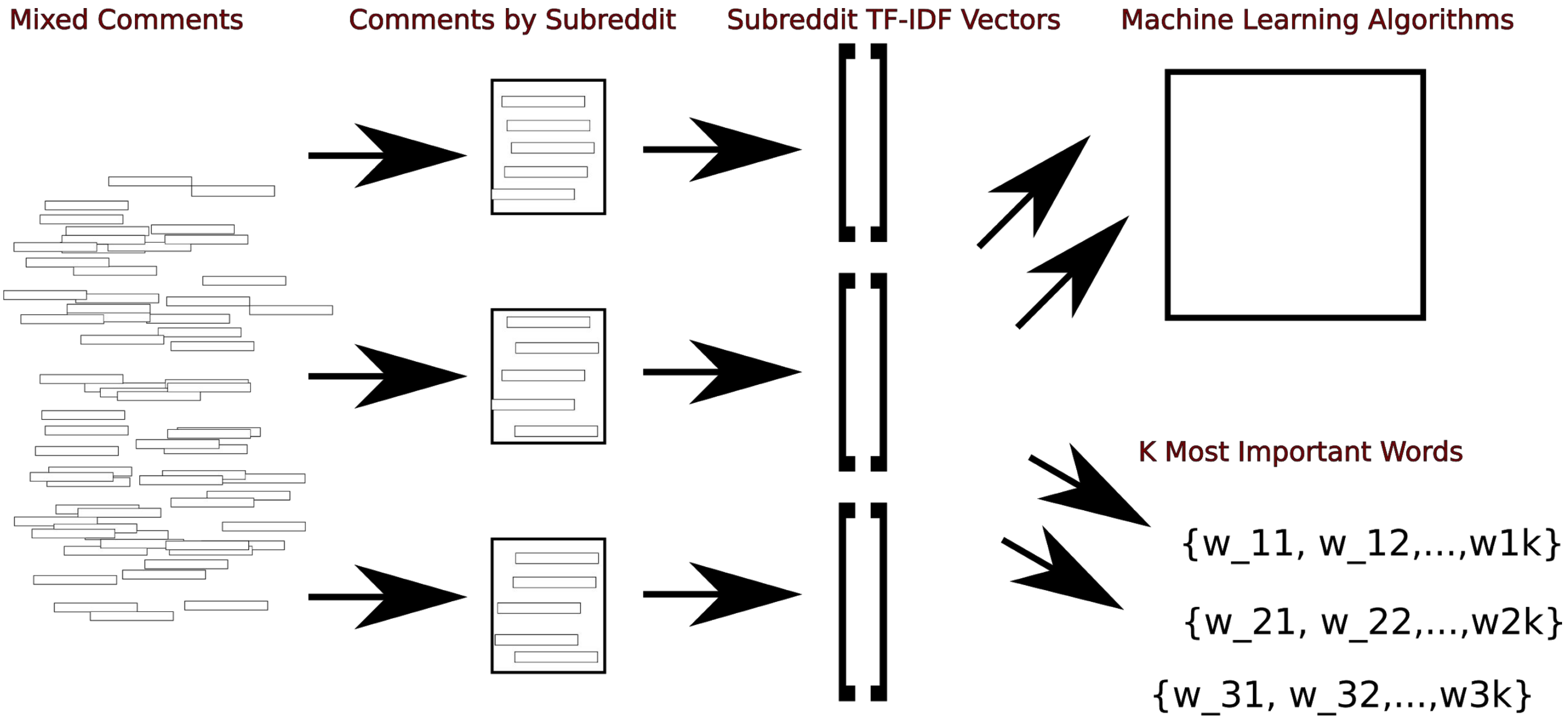
Reddit Premium
The best Reddit experience, with monthly Coins

TRENDING COMMUNITIES

r/Doggies 34.8k members JOIN
r/sciences 110k members JOIN
r/Justridingalong 9.4k members JOIN
r/Barkitecture 9.6k members JOIN
r/grilledcheese 159k members JOIN



Data and Processing Dimensions



Mixed Comments

Comments by Subreddit

Subreddit TF-IDF Vectors

Machine Learning Algorithms

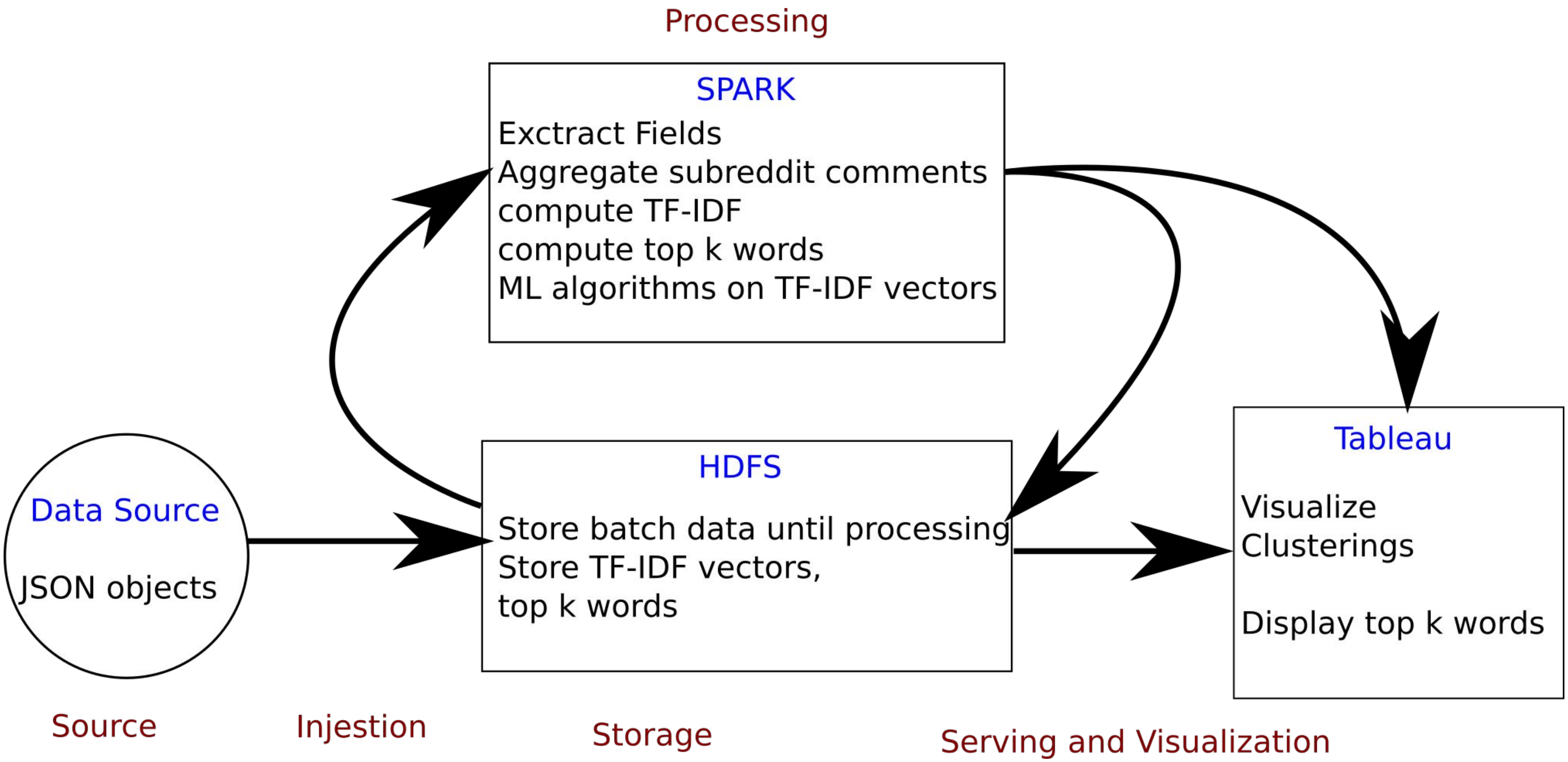
K Most Important Words

{w_11, w_12,...,w1k}

{w_21, w_22,...,w2k}

{w_31, w_32,...,w3k}

Architecture Overview



Processing

SPARK

Exctract Fields
Aggregate subreddit comments
compute TF-IDF
compute top k words
ML algorithms on TF-IDF vectors

Data Source

JSON objects

HDFS

Store batch data until processing
Store TF-IDF vectors,
top k words

Tableau

Visualize
Clusterings

Display top k words

Source

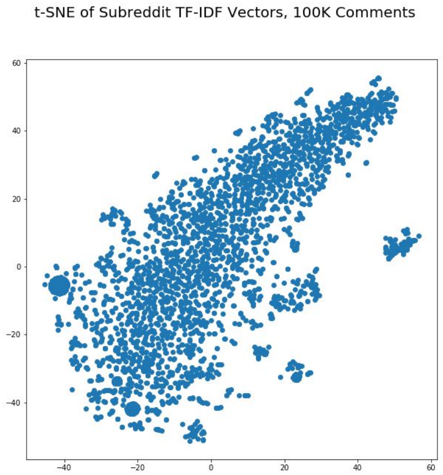
Injestion

Storage

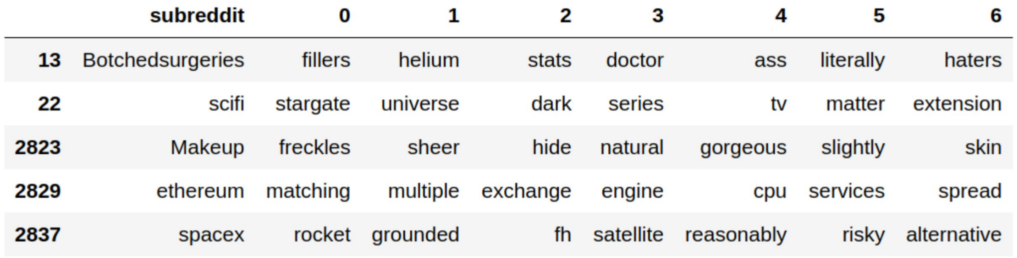
Serving and Visualization

Results

No PCA

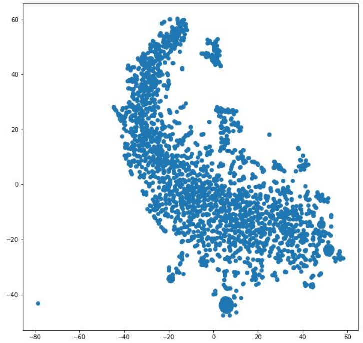

t-SNE of Subreddit TF-IDF Vectors, 100K Comments

Example of Words with Top 7 TF-IDF Scores

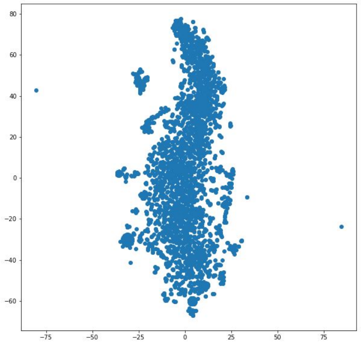


	subreddit	0	1	2	3	4	5	6
13	Botchedsurgeries	fillers	helium	stats	doctor	ass	literally	haters
22	scifi	stargate	universe	dark	series	tv	matter	extension
2823	Makeup	freckles	sheer	hide	natural	gorgeous	slightly	skin
2829	ethereum	matching	multiple	exchange	engine	cpu	services	spread
2837	spacex	rocket	grounded	fh	satellite	reasonably	risky	alternative

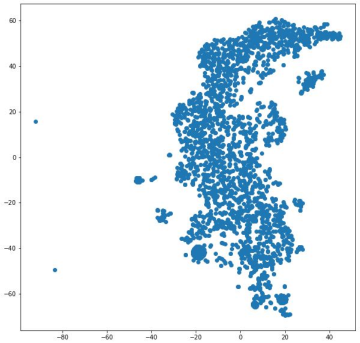
PCA(10)



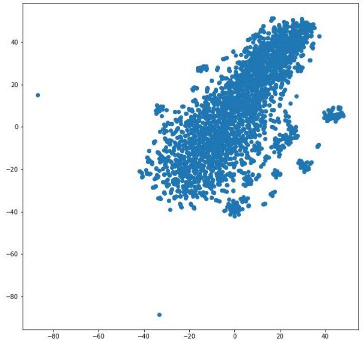
PCA(20)



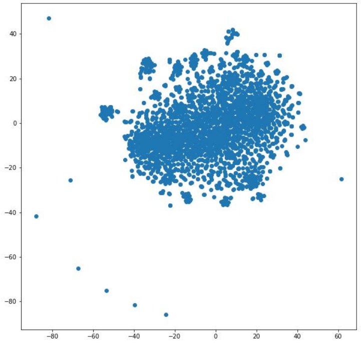
PCA(30)



PCA(40)

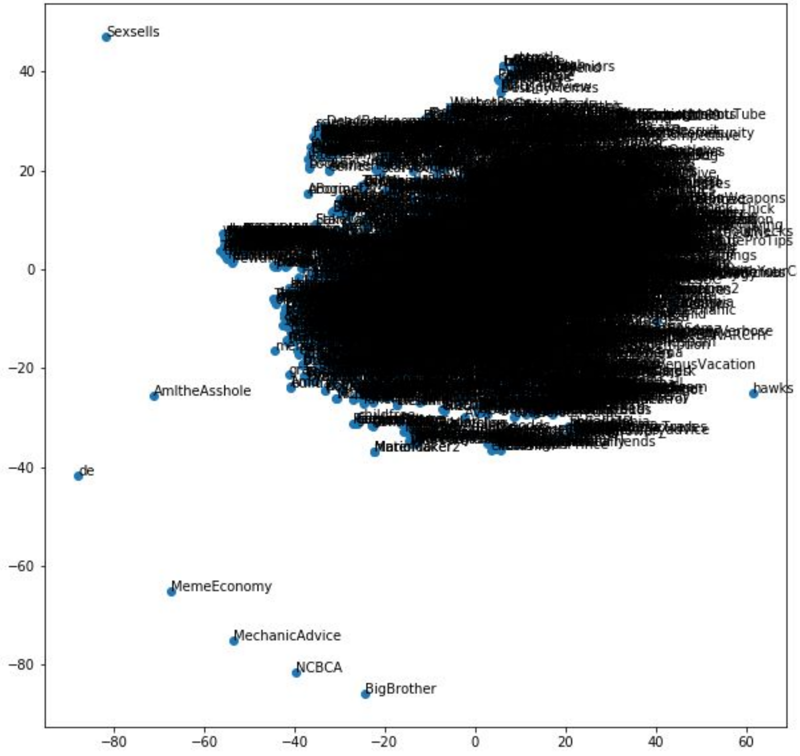


PCA(100)

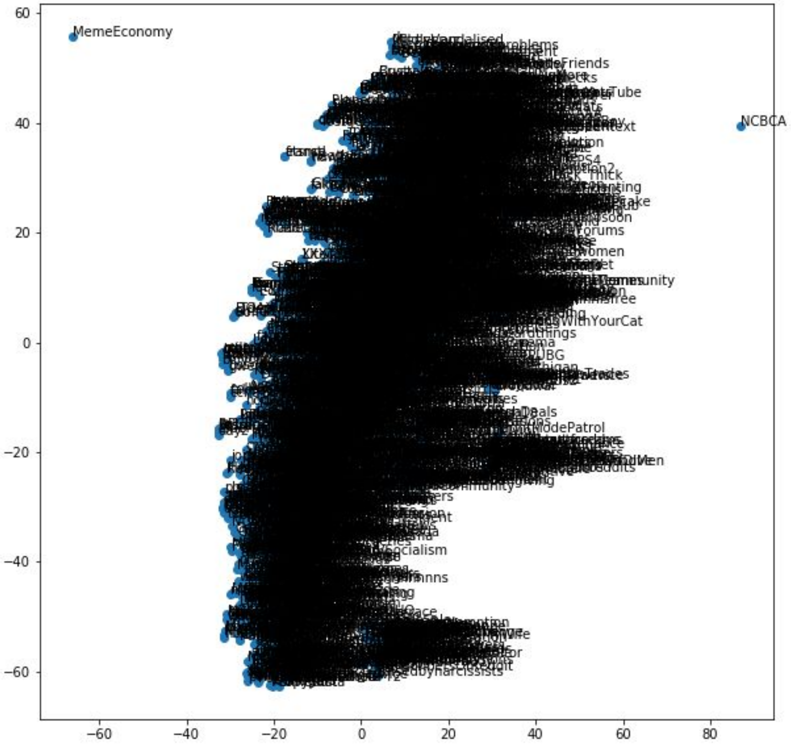


Results

PCA(100)



PCA(20)



- Manually choose subreddits we expect to be ‘similar’ and ‘dissimilar’, then compare their TF-IDF vectors.
- Cluster the TF-IDF vectors

Limitations

1. Currently cannot update TF-IDF vectors incrementally
2. Don’t have an actual cluster..
3. Current design requires two passes over the data

Scalability

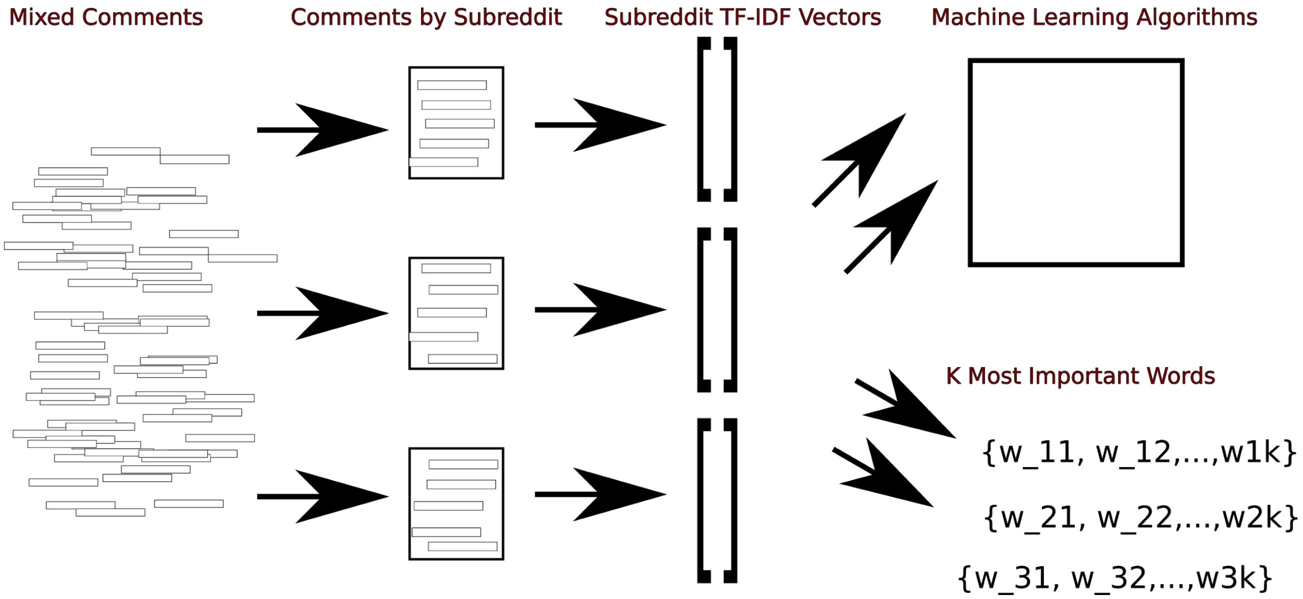
1. Horizontally scalable: Uses Spark and HDFS.
2. Fault tolerant, if running on a cluster.
3. Can handle more data.

Variations and Extensions

1. Add streaming Component, update TF-IDF vectors in real time
2. Consider other document sets, for example

   a document = all of a users text concatenated
1. Use the TF-IDF vectors as features for downstream tasks
2. Compute different kinds of TF-IDF scores, or learn different features in general

Conclusion



Mixed Comments

Comments by Subreddit

Subreddit TF-IDF Vectors

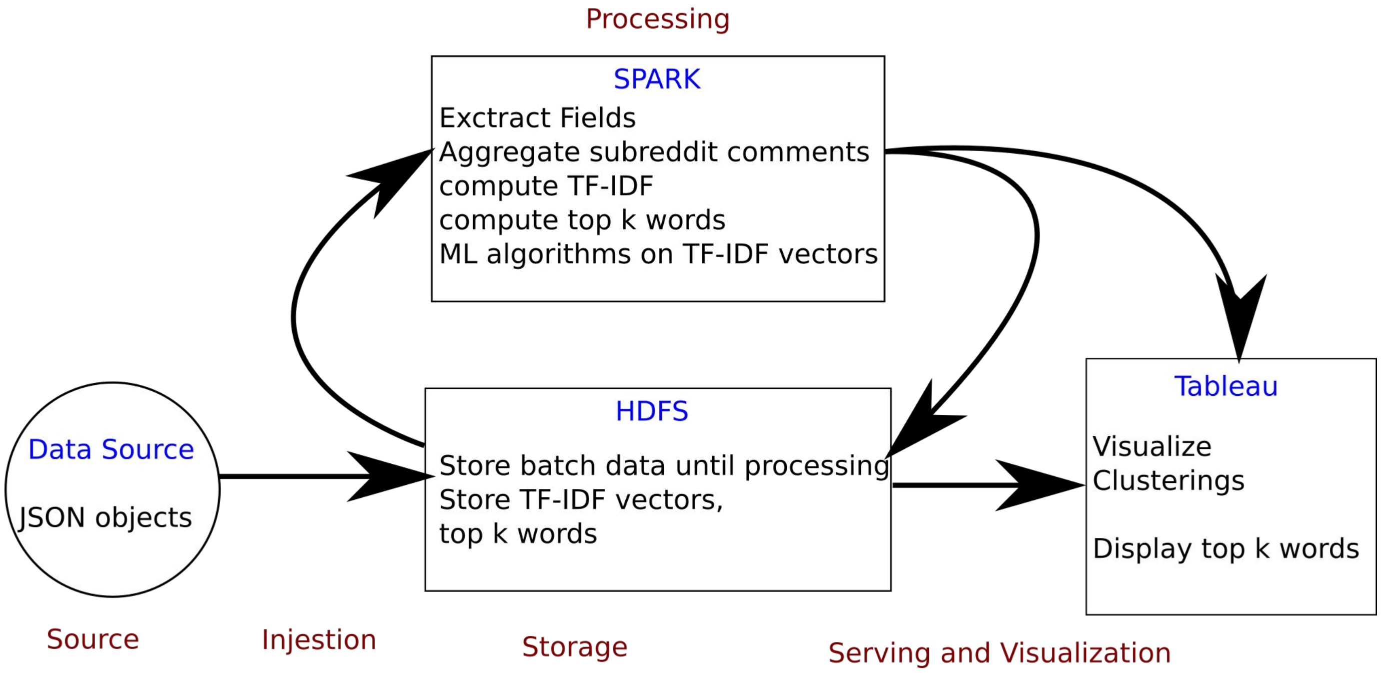
Machine Learning Algorithms

K Most Important Words

{w_11, w_12,...,w1k}

{w_21, w_22,...,w2k}

{w_31, w_32,...,w3k}



Processing

SPARK

Exctract Fields
Aggregate subreddit comments
compute TF-IDF
compute top k words
ML algorithms on TF-IDF vectors

Data Source

JSON objects

HDFS

Store batch data until processing
Store TF-IDF vectors,
top k words

Tableau

Visualize
Clusterings

Display top k words

Source

Injestion

Storage

Serving and Visualization

QUESTIONS?