



REDDIT



Home to microcultures:

- * Academics
- Movie Geeks
- Political Activists
- Music Lovers
- Deviants

- Gamers
- * Foodies
- Gym Rats
- Meme Aficionados
- & More

As of February 2018, Reddit boasted:

- · 6th most visited website in the world
 - * 4th most visited in the US
- ❖ 234 Million unique users
 - * 542 Million monthly visitors
- ♣ 82+ Billion pageviews
 - * 73+ Million submissions/ posts
 - * 725+ Million comments
- 16 Minutes on Site Daily per user
 - * Marketers seek your attention

car (aquels cresificial affine beautiful property along a for the formal property and the following a grant from the following a

grante of me gate pane as



and the said of the said of the control of some street of the said for the said of the said of the said of the



Google BigQuery





❖ 321 GB of Reddit posts available on BigQuery

case (a quali crestrent offer a broad along as

predering thems because office the

רומס מפנה של היים ביותר ביותר

- * Extracted a subset of self posts from June of 2016 with filters for good data
- Nearly 150k posts to explore in a cloudtop powered by 14 vCPU's and 128 GBs RAM

PROCESS

A single not proble come lapare the inne elect bring excelle excellence of the end of a land of the end of the

property of the state of the st

The order of the state of the s

with the control of the control of a facility of the control of th

most weapon another actions with appeared the store trappage of mire is arrived and

- Setup
 - * Google Cloud CL configurations (port, account, access, etc.)
 - * BigQuery API

of idea and again to a company on the same where against a same where against a same where

- Exploratory Data Analysis
 - * Build intuition about text sample (subreddits, lengths, etc.)
 - * Looked for unusual spikes to rid sample of bot posts
 - NULL, de (Denmark), test, tr (Turkey), A6XHE (?), es (Spain), removalbot, tldr (Too Long; Didnt Read)
 - * Could spend many more hours for refined results

PROCESS

Records an propose of one I aport file and chert plant conseque of

car (aquali carriprints office against come of the property of the against carrie of the property of the carries of the carries of the carries of actual to a carries of a carries of

A company of the property of t

- Preprocessing presented room for creativity from query to data cleaning
 - Scrub data of non-alphanumeric + links
 - * Remove stopwords
 - * Tokenize with NLTK
 - Count + Term Frequency-Inverse Document Frequency Vectorizers
 - * Bot activity and other noise required a couple passes of data collection and preprocessing

PROCESS

called the state of the second second from the second of t

the count open and of mentals

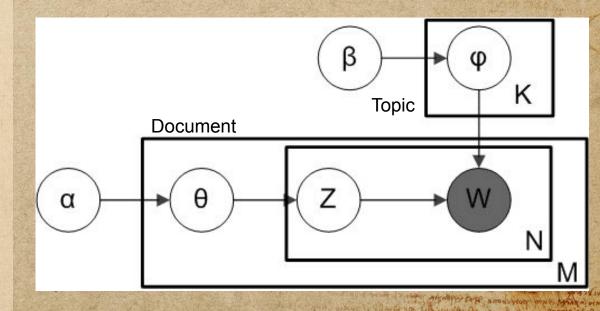
CHARLES THE CHARLES A DOCE

- Singular Value Decomposition / Latent Semantic Analysis (SVD / LSA)
 - * Inexplicable formula for topics (+ / -)
 - * Poor results with TF-IDF, not enough info after SVD
 - K-Means Clustering
 - K-Means could not find meaningful clusters
 - Silhouette scores suggested more and more clusters would result in higher average score
 - * Latent Dirichlet Allocation (LSA with sparse non-uniform Dirichlet dist.) for soft clustering and spatial reduction



- * Assumes document
 - generated in a particular fashion
- distribution in a document
- β plate: prior for word
 distribution in a topic

to the control of the property of the state of the control of



car (aquale confront office bear losses who for

grading of the spane on

walling to widays or one

LATENT DIRICHLET ALLOCATION (LDA)

Vicings not brokers officers found following forth

purpola (Amolhiceria ma Africameri quifi) in mana compo grant alliche superchific i friss en acquificament aliche con in

as orange of the party of the orange

CHARLES CALLEDON OUT

- LDA gave distributions on 15 topics for each document
- ♣ Final model trained in just over 5 hours with 100 epochs
- ❖ Map of documents → topics
- * Collection of topics per user

and were also over also among the among thanking again the . Andrew and also recent up

A nearly not propose expense the laser every private execute excepts a few many and private expenses and private e

man compo grate she'te sapo ofe the e form.

En factor for she'te cos to man all of a same a contract comments for a same a contract comments and a contract co

Live adapted alpos to the second

the same of the same

Pogrado 3

Jodnes 3

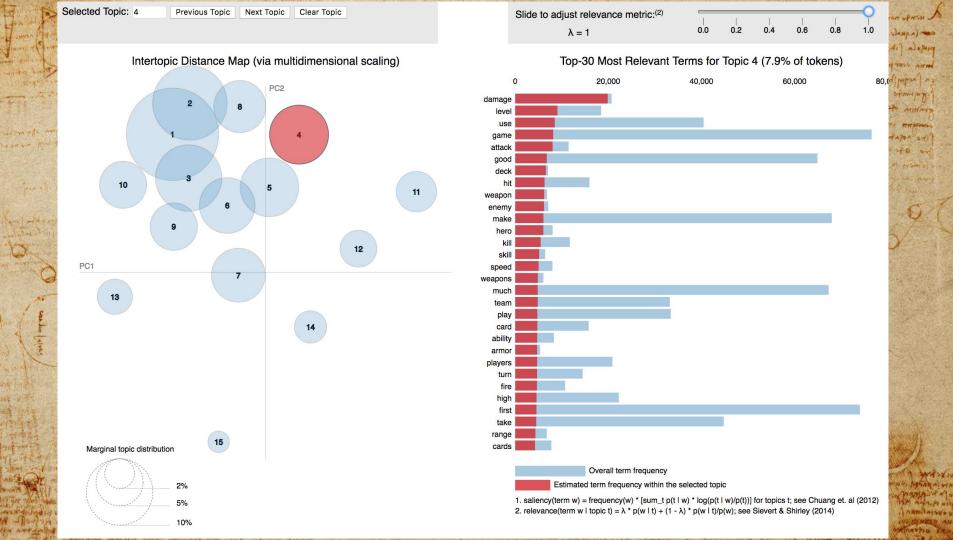
And the state of t

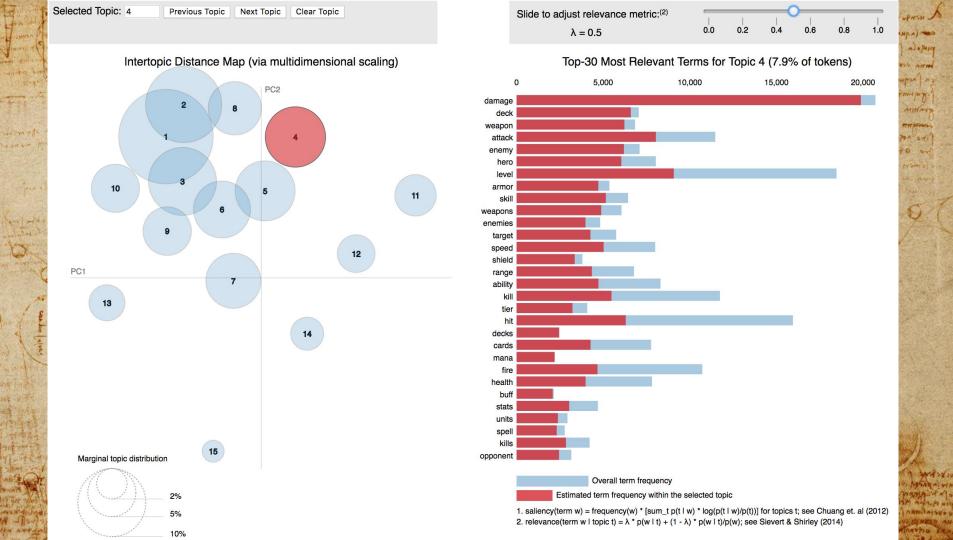
water and a supple and a supple of the suppl

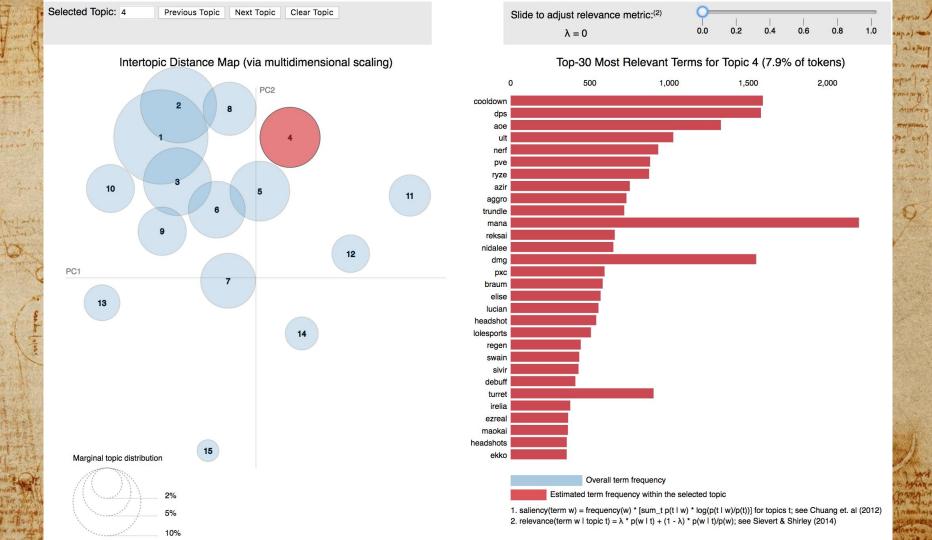
The well of the state of any the state of th

the board their print women on

THE A R. L. WHELL WAS TALK







car laquale conforms offers house blan couls

professor from the server of processor quelled and server of the server of s

Contract trafagor one

- Relevance Sievert & Shirley (2014)
 - * Lift: ratio of a term's probability within a topic to the term's marginal probability across the whole corpus
 - * Lambda: weight given to the probability of a term for a given topic in relation to its lift score

A ninga not problem of our language fulls when efter things considered of the second to have been been a second of the second of

profess from property me to before the second from a many completion of the second completion completio

and sen also over allege much de among states on the in the state of the sent of allege world in

* Relevance

- * Term w, Topic t
- * $\lambda * p(w|t) + (1-\lambda) * p(w|t)/p(w)$
- * On a log scale

A wings not propose gome lapare the work detail of the contraction

car (aquale confront offers depart plan with his

profes (the bolton is no fife the confit of the confit of

* Relevance

- * Lambda = 1: results ranked for terms in descending order of their topic-specific probability
- * Lambda = 0: ranks terms solely by their lift
 - Diminishes globally frequent terms out of ranking
- * $\lambda * p(w|t) + (1-\lambda) * p(w|t)/p(w)$



and the same of the same of the property same many of the same of

- fun fact about call of duty:
- the last call of duty game you enjoyed was when the series went downhill whether that was world at war black ops mw2 black ops 2
- modern warfare feels stale today but it was incredibly fresh and new for the time
- the gameplay loop has been so refined
- enjoy also the cod vs battlefield rivalry
- arcadey shooters and neither are esports
 game
- sci fi games give developers more freedom in gameplay design

- * Topic 1 : Health * 0.02299821
- Topic 2: Gaming/Computers* 0.8877785
- * Topic 7: Sports/ Gaming * 0.080621116

A winds not problem of one language what were first printe except of first

purpola filma human in a fiftensura quafi fingi mina centra grassi allutur sena a filma a puna sa lacquifta grassi kirupa casa as cumalin











IMPROVEMENTS

N-grams

chance field and the sur

- Improve the value of the term space
- Become more specific
- Can rethink the structure of the input and iterate the process

A Land of the all was to seem the second on the second the second

area antiqued at users satelly say went with angular a sure where against an

Word2Vec

and the same of the same of the property same many of the same of

the transfer of the time the surprise of the section of the sectio

- More
 opportunity to
 utilize
 similarities in the
 latent space
- Cut noise
- Giants' Shoulderspretrainedvectors

Non-Matrix Factorization (NMF)

A single not problete from lapare the sent event bring excelle delaid

purpose of the method of the man of the man of the state of the state

Control of the second second

Presumpting of the ord

- * Validate LDA
- Highly interpretable and gives good results

port sp. fallenge sports of seasons or

the resolution was a commenced to the part of the editional





prepala (it me hims his me follower quell me מחום וכמדום מדחור בועורוני במוףם כלני ויינו ב קדונים ב כפני

> The same of case as mild the and pupor ing A TO SECURITION OF THE PARTY OF grading of me of the spane of Comment of the mention

THANK YOU

in agreement afters marred after adapted or manuals of shares and arrange after many from the and seen also over alogo much de assemptions again to see the see a selection of our

> LINKEDIN.COM/IN/DAVISVANCE/ EMAIL: DAVISCVANCE@GMAIL.COM TWITTER: @VAVISDANCE



are to keek to man stood at the state of the





MR A feel in A way as along the Armen on change the former source the with many transfer and the old south within willing a second the administration



CREDITS

Action of making the state of t

property of the principal and the second of the second from the second of the second o

The against franch and a second

Special thanks and credits for the theme:

which is a many of the many the contract of a country became of the many of the many of the many of the many the many the many of the many

- Presentation template by <u>SlidesCarnival</u>
- Photographs by <u>Unsplash</u>
- Paper texture by <u>GraphicBurger</u>



SOURCES

- Reddit Blog: https://redditblog.com/2015/12/31/reddit-in-2015/
- Alexa Internet: https://www.alexa.com/siteinfo/reddit.com
 Retrieved February 28, 2018
- ₩ Wikipedia: https://en.wikipedia.org/wiki/Reddit
- * Relevance: Sievert & Shirley (2014)
 https://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf



A sings not proble come langue the use every bring except of and

property of the property of the same of th

The contract of the second sec

where the property of the more party property of commences and property of the commentation of the construction of the constru

Learned Topics from Latent Dirichlet Allocation

the springs of the war a labeled their a dead for their courses which

the plant from from manager

- **Spanish Stopwords** (0): '0.061*"que" + 0.027*"los" + 0.017*"por" + 0.017*"del" + 0.016*"con" + 0.016*"las" + 0.014*"para" + 0.012*"una" + 0.008*"como" + 0.007*"podemos" + 0.006*"est" + 0.005*"pol" + 0.005*"pero" + 0.005*"han" + 0.004*"psoe"
- Health Products / Food (1): '0.005*"water" + 0.005*"food" + 0.004*"use" + 0.003*"make" + 0.003*"much" + 0.003*"used" + 0.003*"well" + 0.003*"skin" + 0.003*"first" + 0.003*"good" + 0.003*"eat" + 0.003*"two" + 0.002*"new" + 0.002*"ship" + 0.002*"using"
- Gaming/ Internet (2): '0.019*"game" + 0.007*"games" + 0.007*"new" + 0.007*"play" + 0.005*"think" + 0.005*"much" + 0.005*"make" + 0.004*"way" + 0.004*"see" + 0.004*"good" + 0.004*"want" + 0.004*"something" + 0.004*"use" + 0.004*"first"
- Empathy/ Social (3): '0.008*"want" + 0.008*"feel" + 0.006*"think" + 0.006*"going" + 0.006*"never" + 0.006*"life" + 0.005*"things" + 0.005*"much" + 0.005*"day" + 0.005*"got" + 0.005*"years" + 0.005*"said" + 0.005*"something" + 0.004*"cant" + 0.004*"friends"
- Politics/ Culture/ Global (4): '0.004*"world" + 0.003*"women" + 0.003*"men" + 0.003*"believe" + 0.003*"may" + 0.002*"war" + 0.002*"think" + 0.002*"country" + 0.002*"state" + 0.002*"government" + 0.002*"way" + 0.002*"church" + 0.002*"make" + 0.002*"say" + 0.002*"right"



A windy not propose of one lapart fin and detail bring creater of land

com (aquatic conspicues of the aquatic planes of a first program of the apparent planes of the first program of the construction of the constructi

And we will be a series of the series of the

where the property of the more party property of commences and property of the commentation of the construction of the constru

reduces of somety -- reduction access

Learned Topics from Latent Dirichlet Allocation

Will be of the state of the property of the state of the

- Social/ Storytelling (5): '0.011*"back" + 0.006*"around" + 0.006*"said" + 0.006*"got" + 0.005*"room" + 0.004*"went" + 0.004*"night" + 0.004*"door" + 0.004*"see" + 0.004*"little" + 0.004*"right" + 0.004*"still" + 0.004*"head" + 0.004*"started" + 0.004*"going"
- Books/ Music (6): '0.013*"chapter" + 0.012*"world" + 0.008*"song" + 0.006*"music" + 0.005*"table" + 0.005*"album" + 0.005*"links" + 0.005*"previous" + 0.005*"synopsis" + 0.005*"life" + 0.005*"source" + 0.005*"songs" + 0.004*"contents" + 0.004*"raw"
- League of Legends/ Gaming/ Sports (7): '0.032*"team" + 0.017*"game" + 0.013*"match" + 0.011*"players" + 0.009*"win" + 0.009*"teams" + 0.008*"games" + 0.007*"season" + 0.007*"round" + 0.007*"play" + 0.007*"league" + 0.006*"player" + 0.006*"2016" + 0.006*"week" + 0.006*"tournament"
- Career (8): '0.008*"work" + 0.007*"money" + 0.006*"year" + 0.006*"job" + 0.005*"years" + 0.004*"need" + 0.004*"pay" + 0.004*"new" + 0.004*"back" + 0.004*"company" + 0.004*"going" + 0.004*"day" + 0.003*"take" + 0.003*"car" + 0.003*"help"
- Biblical/War (9): '0.004*"man" + 0.004*"death" + 0.004*"lord" + 0.003*"back" + 0.003*"first" + 0.003*"see" + 0.003*"king" + 0.003*"think" + 0.003*"dead" + 0.003*"world" + 0.003*"way" + 0.003*"two" + 0.003*"well" + 0.003*"battle" + 0.003*"men"



Surply my property one lapart from the control printer confunction of

com (aquali ceriprini entre aquant describe frequents of the ceriprini of the aquant describe frequents of the ceriprini of the ceripri of the ceriprini of the ceriprini of the ceripri of the ceriprini of the ceripri of the cer

And we will be a series of the series of the

Learned Topics from Latent Dirichlet Allocation

which district adults of week when weight a such above appropriate

Entertainment (10): '0.010*"episode" + 0.009*"think" + 0.009*"show" + 0.006*"season" + 0.006*"character" + 0.005*"trump" + 0.005*"see" + 0.005*"vote" + 0.005*"story" + 0.005*"characters" + 0.005*"first" + 0.004*"series" + 0.004*"book" + 0.004*"movie" + 0.004*"good"

where the property of the more party property of commences and property of the commentation of the construction of the constru

- Trading (11): '0.014*"card" + 0.010*"price" + 0.009*"windows" + 0.008*"purchase" + 0.007*"amazon" + 0.006*"pokemon" + 0.006*"buy" + 0.005*"shipping" + 0.005*"box" + 0.005*"key" + 0.005*"sale" + 0.005*"gift" + 0.004*"item" + 0.004*"case" + 0.004*"pro"
- **Sports** (12): '0.011*"game" + 0.008*"ball" + 0.008*"scores" + 0.008*"score" + 0.007*"icon" + 0.006*"line" + 0.006*"bar" + 0.005*"field" + 0.005*"2016" + 0.005*"left" + 0.005*"center" + 0.005*"right" + 0.004*"first" + 0.004*"fielder" + 0.004*"year"
- Reddit-Meta/ Community (13): '0.015*"post" + 0.011*"thread" + 0.011*"please" + 0.008*"new" + 0.008*"link" + 0.008*"questions" + 0.006*"want" + 0.006*"help" + 0.006*"reddit" + 0.005*"free" + 0.005*"community" + 0.005*"edit" + 0.005*"week" + 0.005*"comments" + 0.005*"posts"
- Gaming (Card/ MMORPG) (14): '0.011*"damage" + 0.005*"level" + 0.005*"use" + 0.004*"game" + 0.004*"attack" + 0.004*"good" + 0.004*"deck" + 0.003*"hit" + 0.003*"weapon" + 0.003*"enemy" + 0.003*"make" + 0.003*"hero" + 0.003*"kill" + 0.003*"skill" + 0.003*"speed"