



## REDDIT



#### Home to microcultures:

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

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

#### As of February 2018, Reddit boasted:

- · 6<sup>th</sup> most visited website in the world
  - \* 4<sup>th</sup> 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

com (aquels conspects of the appart plants the

## GOOGLE CLOUD PLATFORM & DATA

and was the aut along and to a mind the sample and the transfer of the the transfer of along the

A comment of the second of the

ב מות מובים לה כני מיקום או למוחווום ביני

case (a quali crestocido effera brane l'alan se

megala fitmo horrior ma

Comment in the personner of the comments of th



# Google BigQuery



- ❖ 321 GB of Reddit posts available on BigQuery
- Extracted a small subset of self posts with filters for good content
- ♣ 150k posts of endless possibilities to explore in the cloud



Google Compute Engine

## PROCESSES & PAIN POINTS

- \* Setup
- Exploratory Data Analysis
  - \* Low hanging fruit
  - \* Build intuition
- \* Preprocessing
- Iterating EDA & Preprocessing
- Singular Value Decomposition / Latent Semantic Analysis
- Similarity Metrics
- \* KMeans Clustering
- \* Latent Dirichlet Allocation

- Working on the cloud
- Biases/ assumptions in the data acquisition & cleaning

A new for nor propose come topper the way there of the contraction of the

pergola from horner or a fillenter out from

semila the and pupor a:

and property of the second

- \* Queries, Stopwords, Tokenization
- \* Bot activity/ Noise
- Poor results with Term
   Frequency-Inverse Document
   Frequency vectorizer
- KMeans could not find meaningful clusters
- · No pain, no gain

# LATENT DIRICHLET ALLOCATION (LDA)

case (aquale conforme office beautifular colores

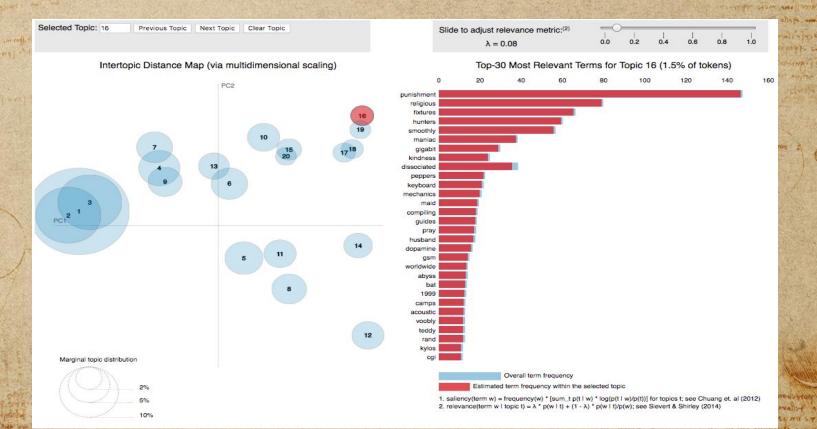
people (three people on a photomer appeal to the series of the series of

And the section of th

- LDA gave distributions for each document
- Map of documents -> topics
- Collection of topics per user
  - \* User profiles useful for marketing & more!
- \* Bonus: probabilistic outcome

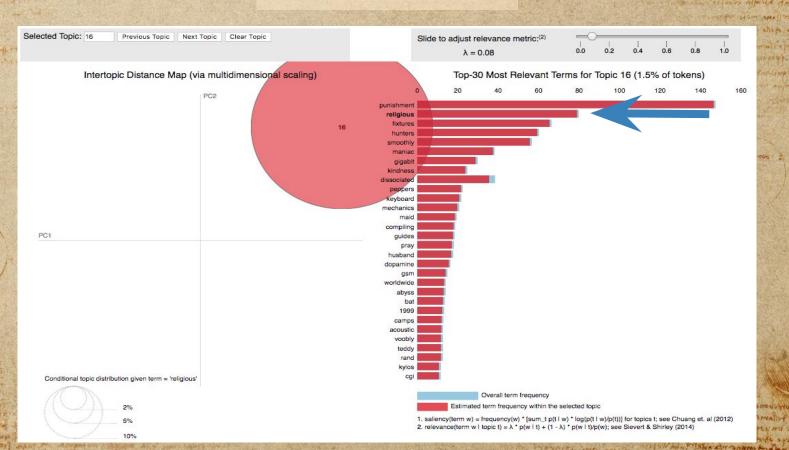
## **PYLDAVIS**

com (aquali cresiprine after base plans siche torre



## **PYLDAVIS**

car (aquali cresiprine after brown folian chila tim





where the many the many of the control of some property of the many open the control of the cont

- 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 esportsgame
- sci fi games give developers more freedom in gameplay design

\* Topic 1 : Health 
\* 0.02299821

A wings not propose of one lagrant the man of collection except of form

mil April a come of the second of the second

Control of the second second

- Topic 2: Gaming/Computers\* 0.8877785
- Topic 7: Sports/Gaming\* 0.080621116











## **IMPROVEMENTS**

#### N-grams

shearen male sent she sust

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

area antiqued at users satelly say court enter congress a such charte agreement and

#### Word2Vec

where the most the more than a most of some most of some and the most of the most of the most of the most one

- 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

purpola firms principe no Africamor qualifica mana compo grade allette sopo official fine grans en forgon ha givet divine com so

A manufa francis in manufacture and a state of the manufacture of the

שונים ספר אף ביאחורים

- Validate LDA
- Highly interpretable and gives good results

print sp. intersprintings. Amallet a

the resolution was a commenced to the part of the editional



MR A TO THE STATE OF THE PROPERTY OF THE PROPERTY OF THE PROPERTY OF with many transfer and the old south within willing a second the administration



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 of party with took come State children in the taken in who were a property and a small and and and and the number of city at primition got

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

> and which will be a see to the see the which the last with the last grading of me of the spane on Comment of the mention LINE REPARTS OF PRINCES PROBLEM OF SHIP ON

Survey on Aiddle Antis all tout They at ranger for for per per marries jour rige | assemblished apartition of a sun or sunder our



### CREDITS

A wings not proble come langue the was every bring cooper of and

and organized profession of the state of the

margings some age committees spagganting on a sunggior

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

All and the state of the second secon

Special thanks and credits for the theme:

with the same of the control of a same control of same of poster of the control of the control of the control of

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



the April 19 of the total state of states on the son Mil son south

with surveyor with your a happen in the flow po effete cuture being the para

## 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



A winds not propose of our lapart file was detail plant content colored

com (aquele cur print of me fighter and form the form and frage form from the me fighter and the fighter and fight

A design of the second of the

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

Well to a file out many the account of the same of the same about the

THE POTE THE PURE WARRENT

- **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 winds not propose of one lapart from men detall strate creates of land

com (aquatic conspicues of the aquatic planes of a first program of the apparent planes of the first program composition of the composition of the

And we will be a series of the series of the

where the provided interest place which the common of the place of the same could be strong in

#### Learned Topics from Latent Dirichlet Allocation

which days adoles of west with anything and their against the

THE WAY THE PURE WARRENT

- 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"



A winds not propose of one lapart file was detal plane creates of land

com (aquali ceriprini entre aquant elevanica frequential of the second of the second elevanical frequential of the second elevanical of the second

A delivery of the second of th

#### Learned Topics from Latent Dirichlet Allocation

all the state of the party of the state of the state of the state of the state of

THE PART PROPERTY MANAGER

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"