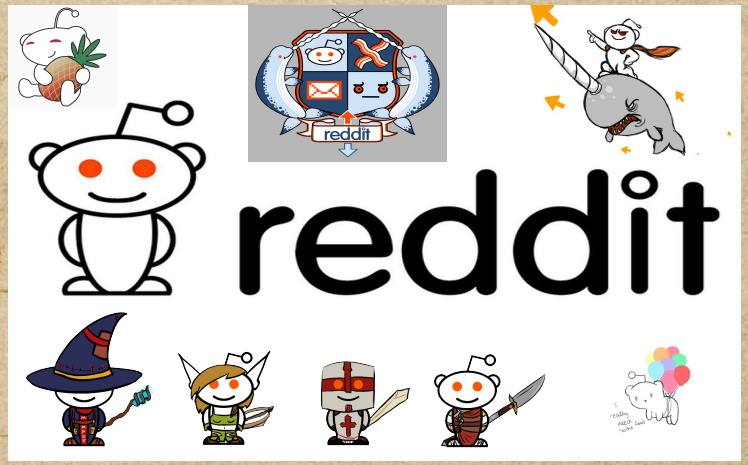
Topic Modeling Reddit Posts

Reddit

- *American social news aggregation, web content rating, and discussion website*
- *Founded June 23, 2005*

Reddit



Home to microcultures:

- Academics
- Movie Geeks
- Music Lovers
- Political Activists
- Sexual Deviants
- Gamers
- Foodies
- Gym Rats
- Meme Aficionados
- & More

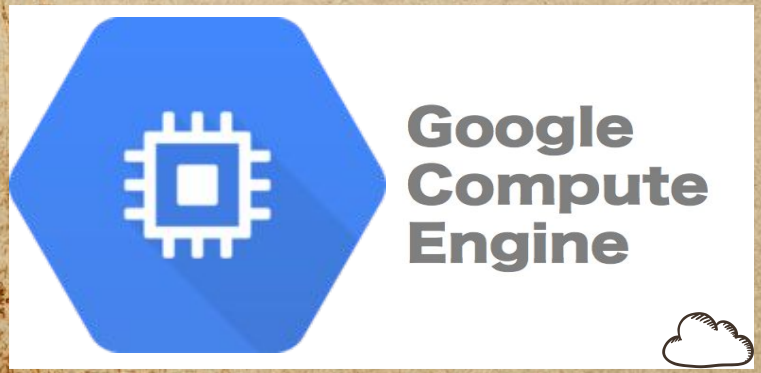
*As of February 2018, Reddit boasted:

- 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**

* Alexa Internet

Google Cloud Platform & Data

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

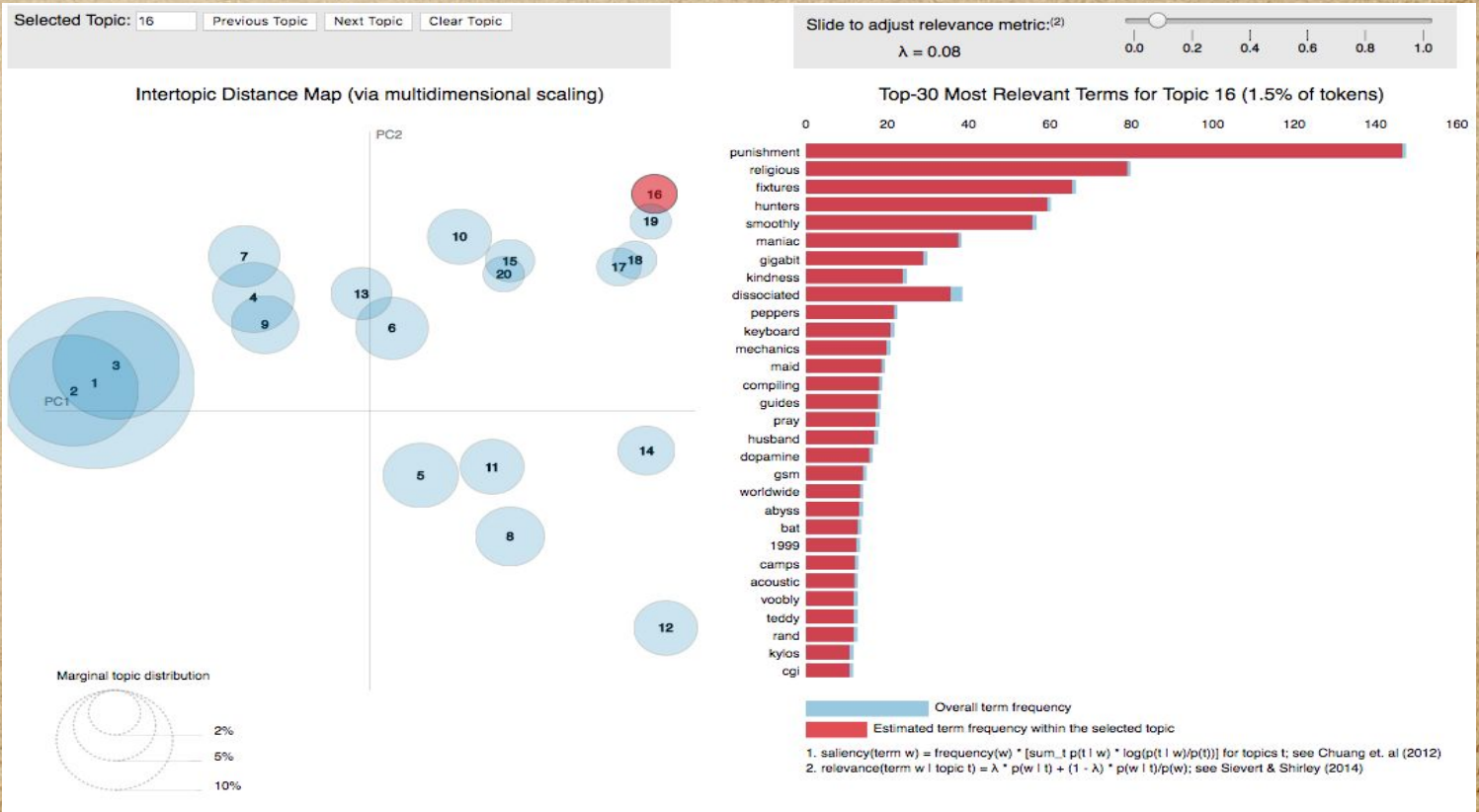
Processes & Pain points

- 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
  - 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

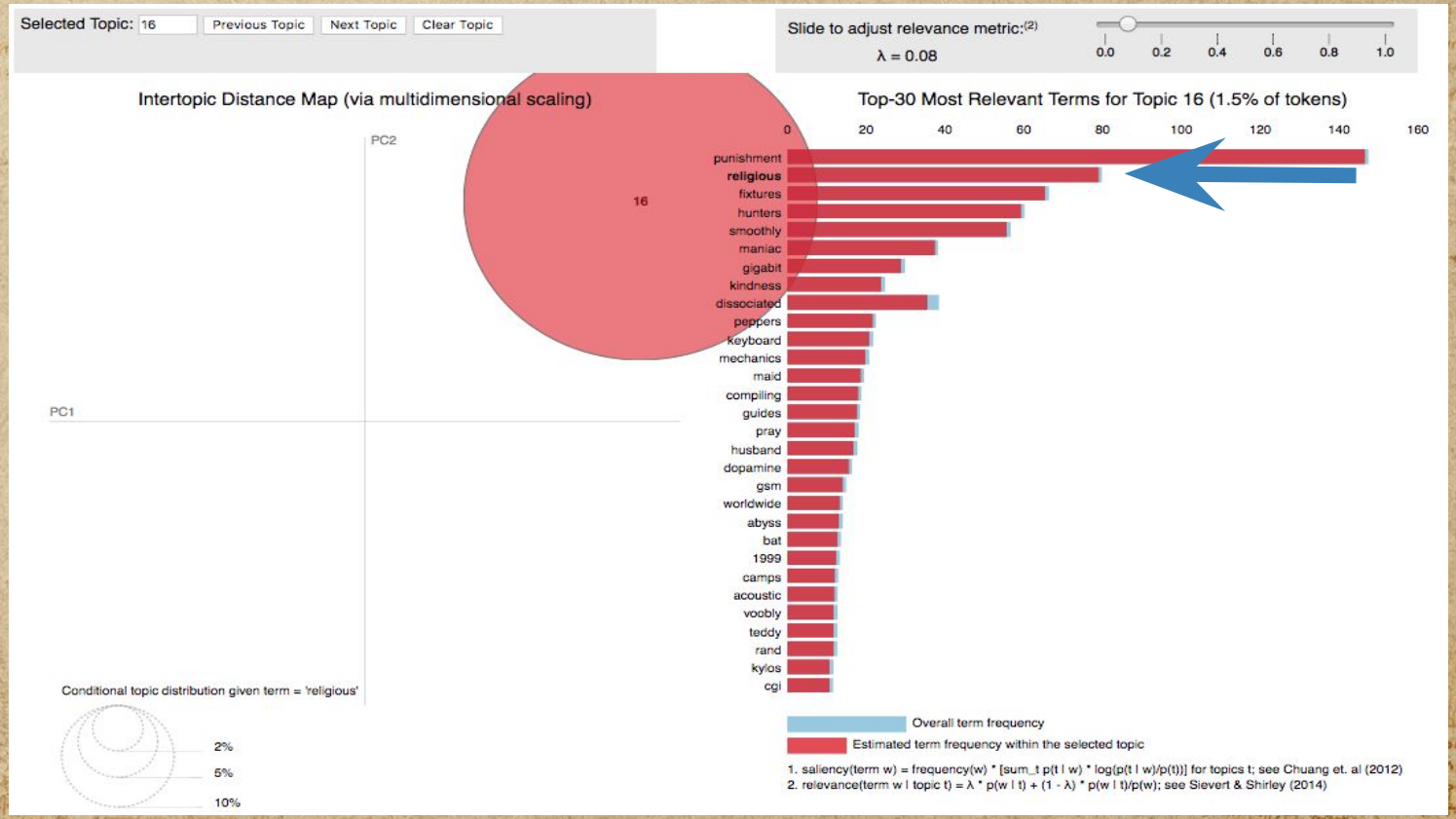
Latent Dirichlet Allocation (LDA)

- 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


Selected Topic: 16
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0.08
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
16
19
10
7
15
20
17
18
4
13
9
6
3
1
2
5
11
14
8
12
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 16 (1.5% of tokens)
0 20 40 60 80 100 120 140 160
punishment
religious
fixtures
hunters
smoothly
maniac
gigabit
kindness
dissociated
peppers
keyboard
mechanics
maid
compiling
guides
pray
husband
dopamine
gsm
worldwide
abyss
bat
1999
camps
acoustic
voobly
teddy
rand
kylos
cgi
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

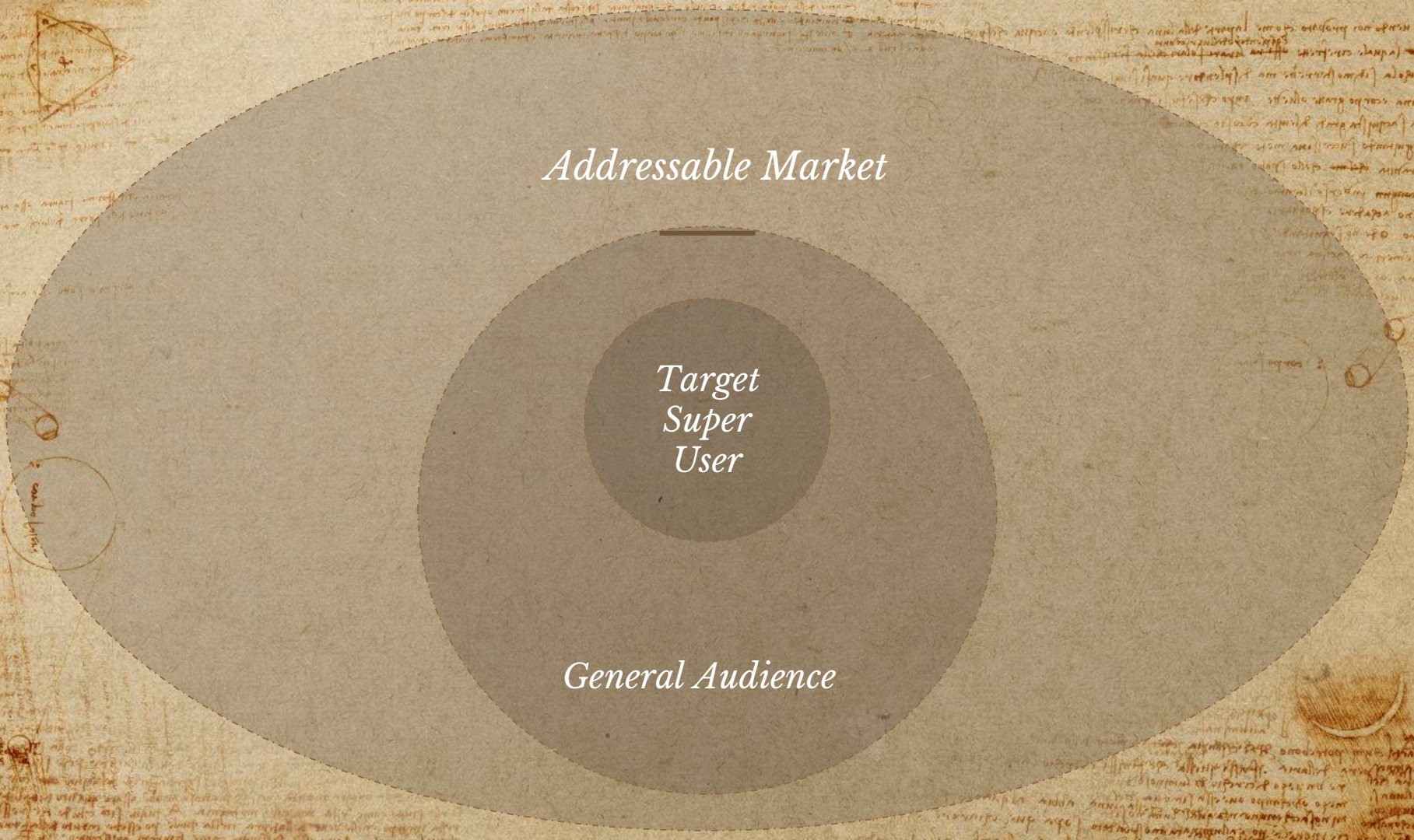
PyLDAvis

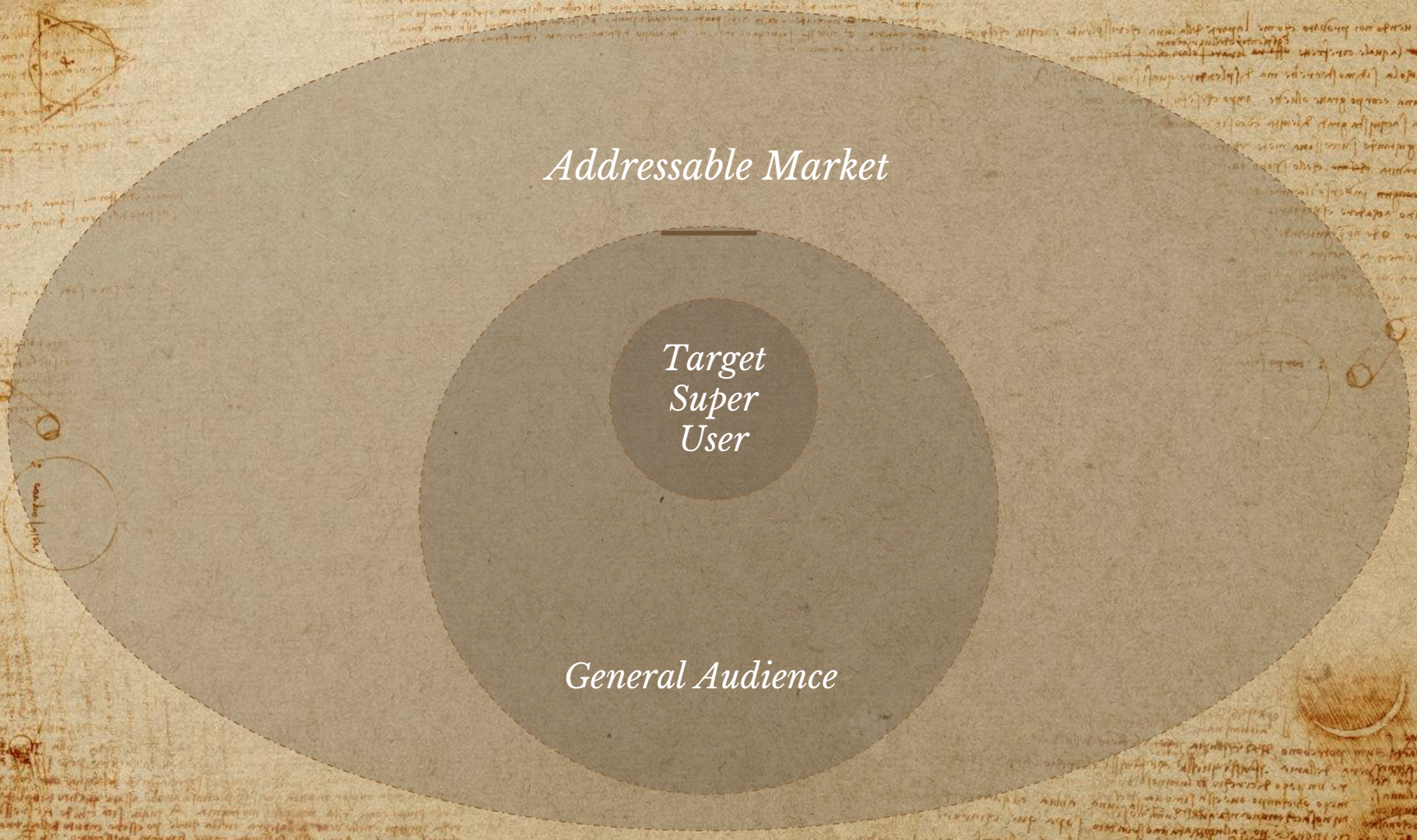

Selected Topic: 16
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0.08
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
16
Conditional topic distribution given term = 'religious'
2%
5%
10%
Top-30 Most Relevant Terms for Topic 16 (1.5% of tokens)
0
20
40
60
80
100
120
140
160
punishment
religious
fixtures
hunters
smoothly
maniac
gigabit
kindness
dissociated
peppers
keyboard
mechanics
maid
compiling
guides
pray
husband
dopamine
gsm
worldwide
abyss
bat
1999
camps
acoustic
voobly
teddy
rand
kylos
cgi
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

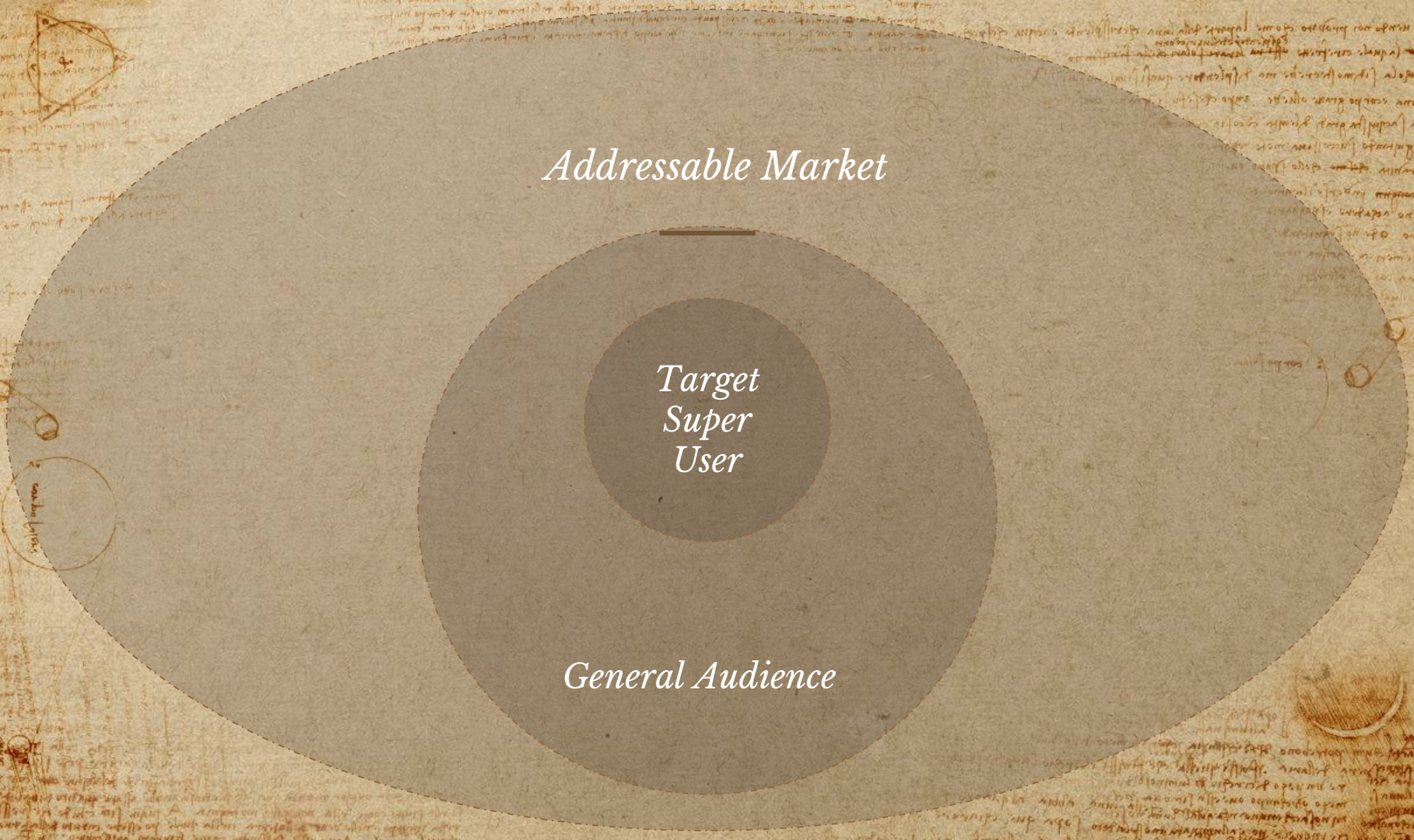
Application

- 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


Addressable Market
Target
Super
User
General Audience


Addressable Market
Target
Super
User
General Audience


Addressable Market
Target
Super
User
General Audience


Addressable Market
Target
User
General Audience


Addressable Market
Target User
General Audience

Improvements

N-grams

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

Word2Vec

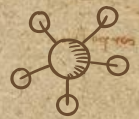
- More opportunity to utilize similarities in the latent space
- Cut noise
- Giants' Shoulders - pretrained vectors

Non-Matrix Factorization (NMF)

- Validate LDA
- Highly interpretable and gives good results







Thank You

linkedin.com/in/davisvance/

Email: DavisCVance@gmail.com

Twitter: @VavisDance


Credits

Special thanks and credits for the theme:

- Presentation template by SlidesCarnival
- Photographs by Unsplash
- Paper texture by GraphicBurger

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

Appendix

Learned Topics from Latent Dirichlet Allocation

- 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*"playing" + 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"'

Appendix

Learned Topics from Latent Dirichlet Allocation

- 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*"new" + 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"

Appendix

Learned Topics from Latent Dirichlet Allocation

- **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"'
- **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"'