Florian Philippi

06/02/2017

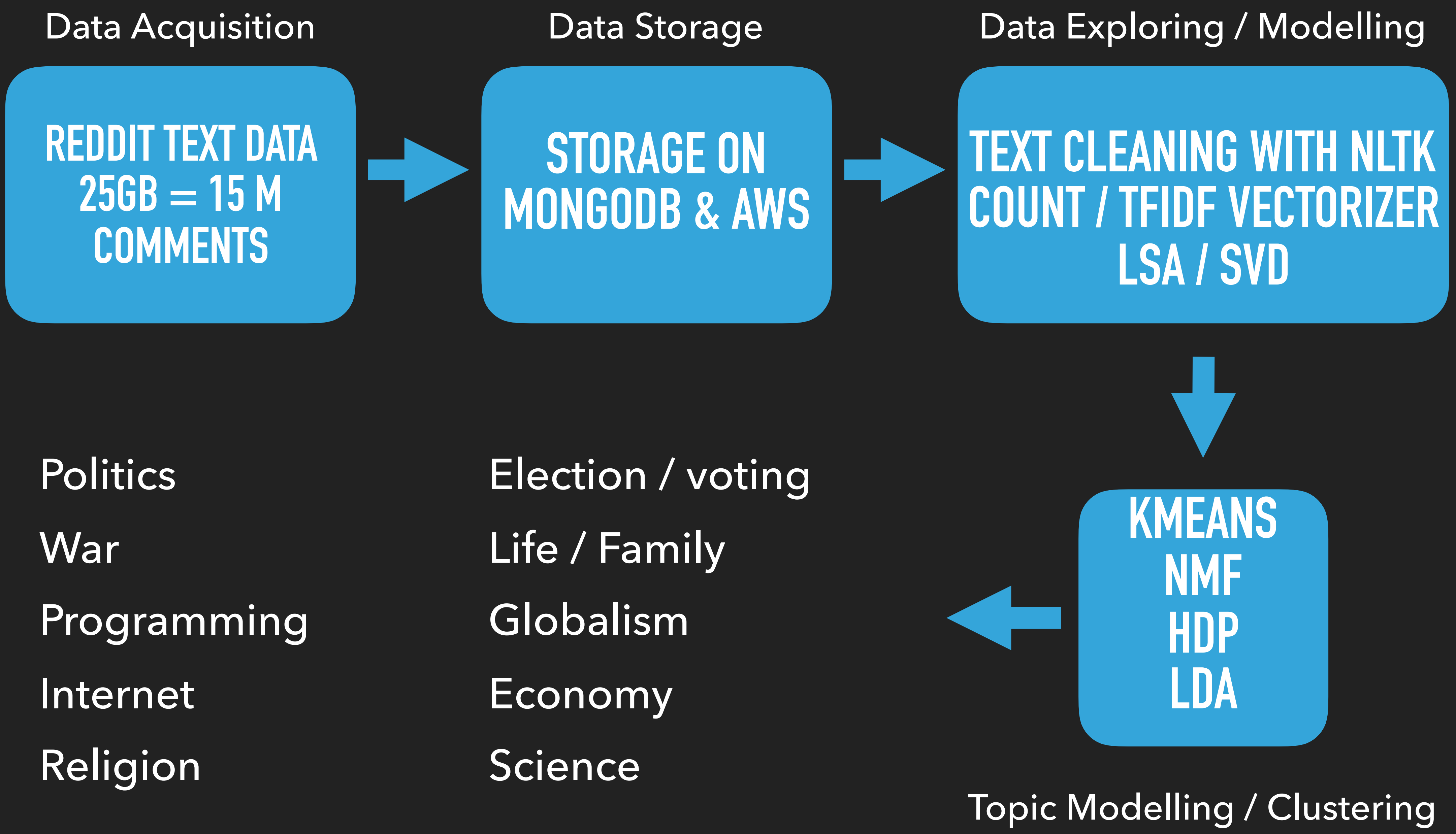
REDDIT USER ANALYTICS

NATURAL LANGUAGE PROCESSING

OBJECTIVES

- Investigate reddit comments data on a user basis
- 300GB of text data available online
- Create a tool to summarize user activities using NLP
    - Topics of interest
    - Connection to users with similar interests
    - Suggestions for new connections and topics

PROCESS



Data Acquisition

REDDIT TEXT DATA
25GB = 15 M
COMMENTS

Data Storage

STORAGE ON
MONGODB & AWS

Data Exploring / Modelling

TEXT CLEANING WITH NLTK
COUNT / TFIDF VECTORIZER
LSA / SVD

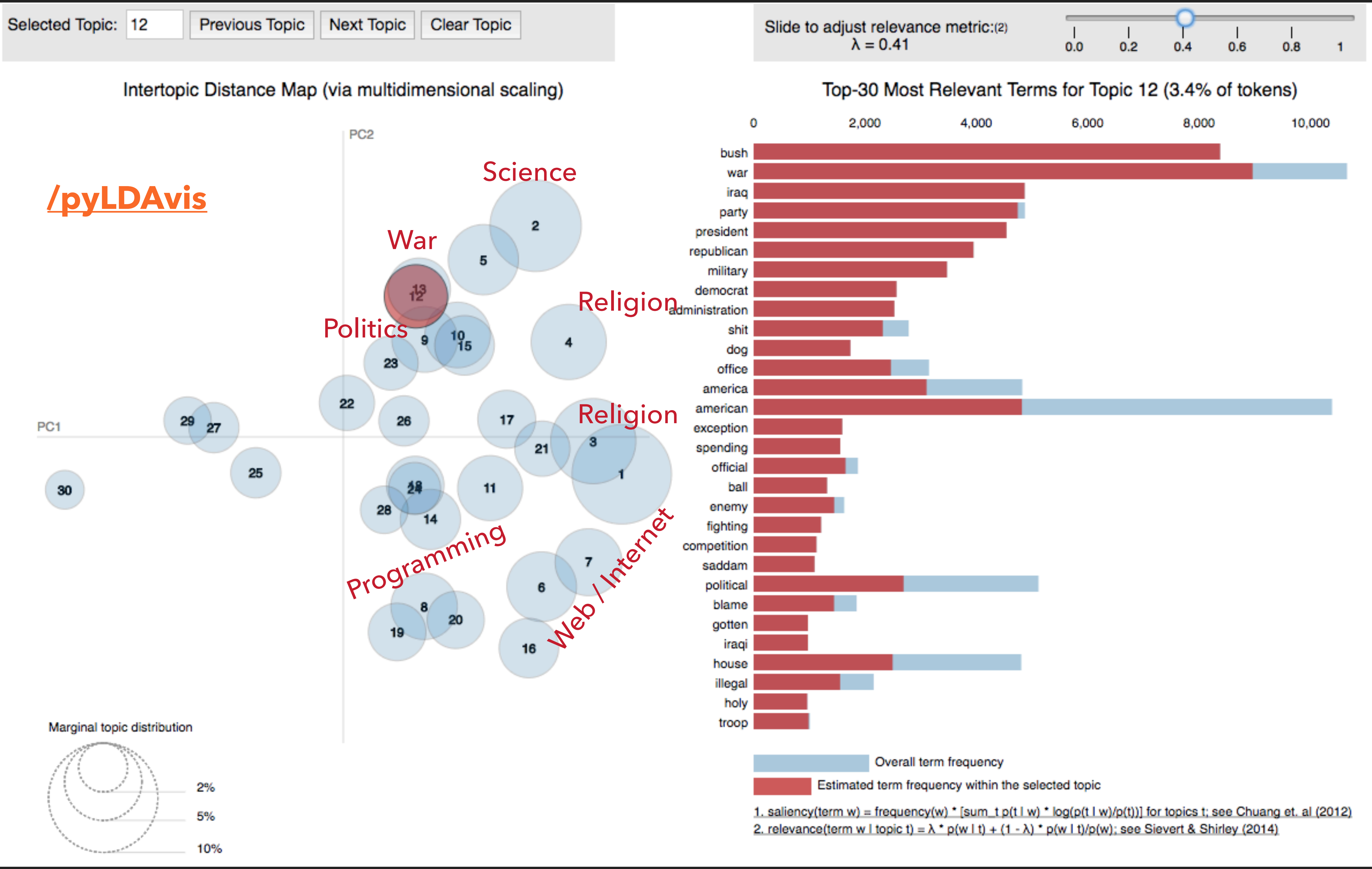
KMEANS
NMF
HDP
LDA

Topic Modelling / Clustering

Politics
War
Programming
Internet
Religion

Election / voting
Life / Family
Globalism
Economy
Science

TOPICS FOUND WITH LDA



/pyLDAvis

Selected Topic: 12 | Previous Topic | Next Topic | Clear Topic

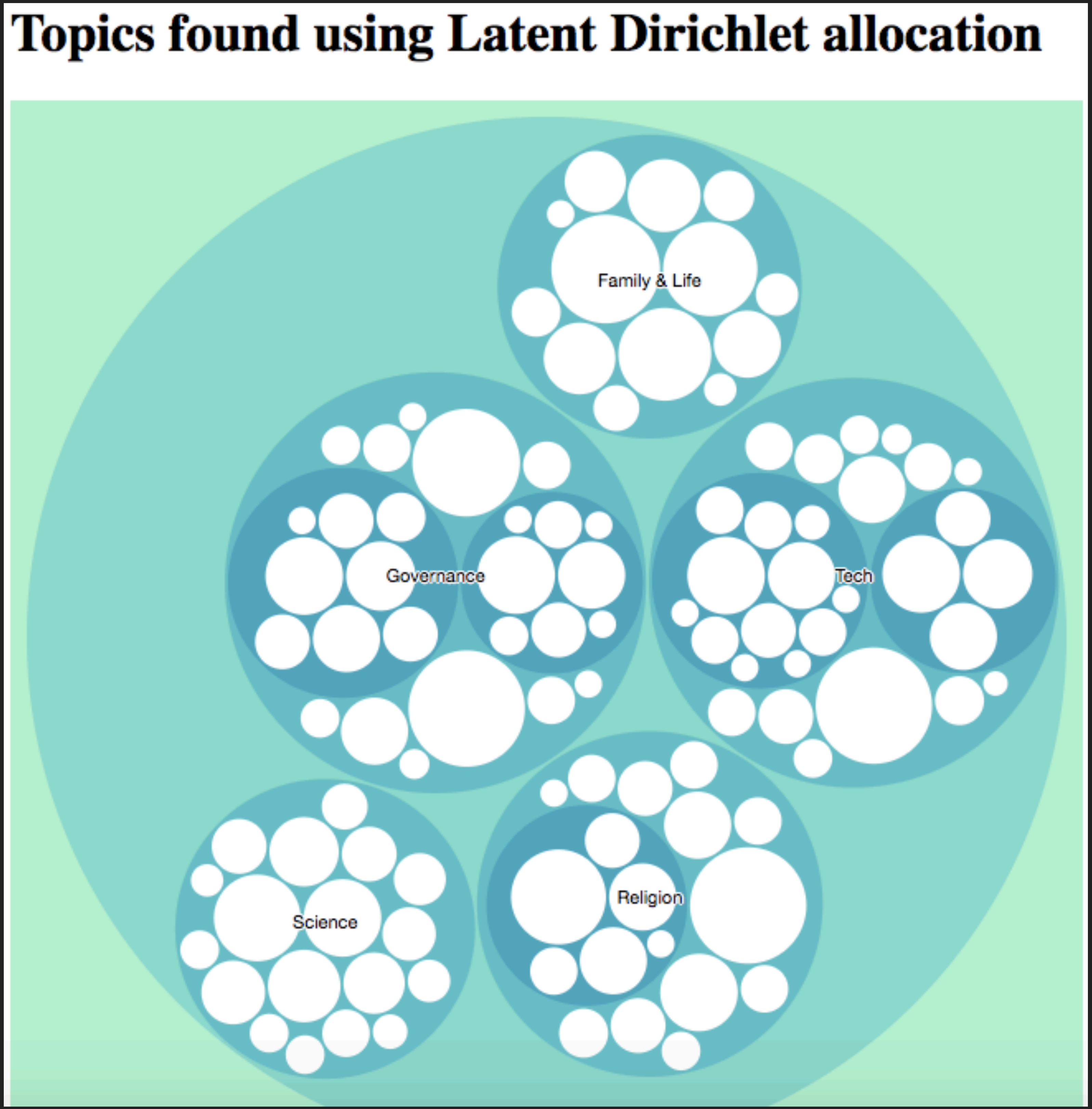
Slide to adjust relevance metric:(2)
λ = 0.41

🟦 Overall term frequency

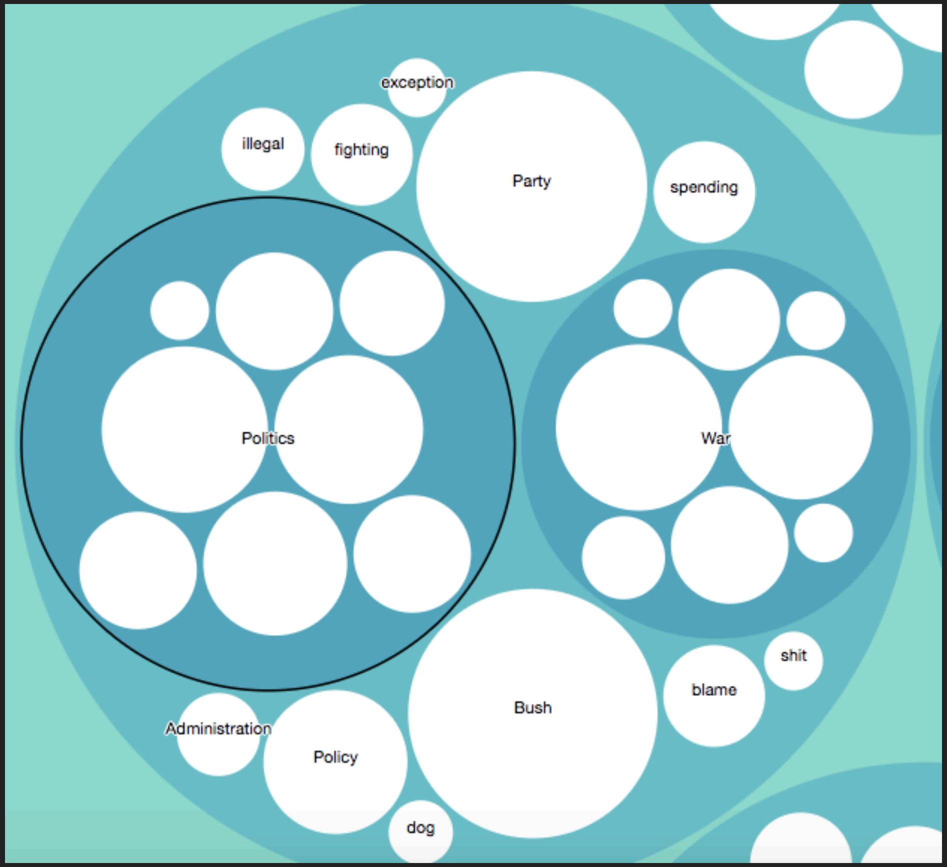
🟥 Estimated term frequency within the selected topic

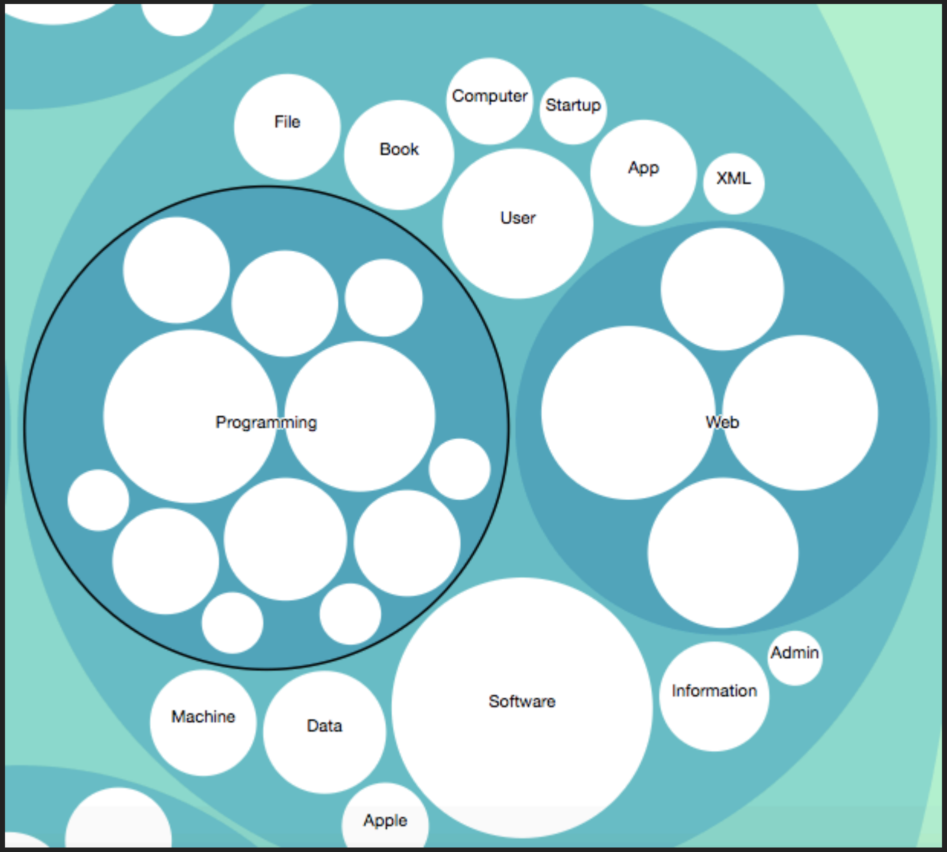
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

VISUALIZATIONS - D3

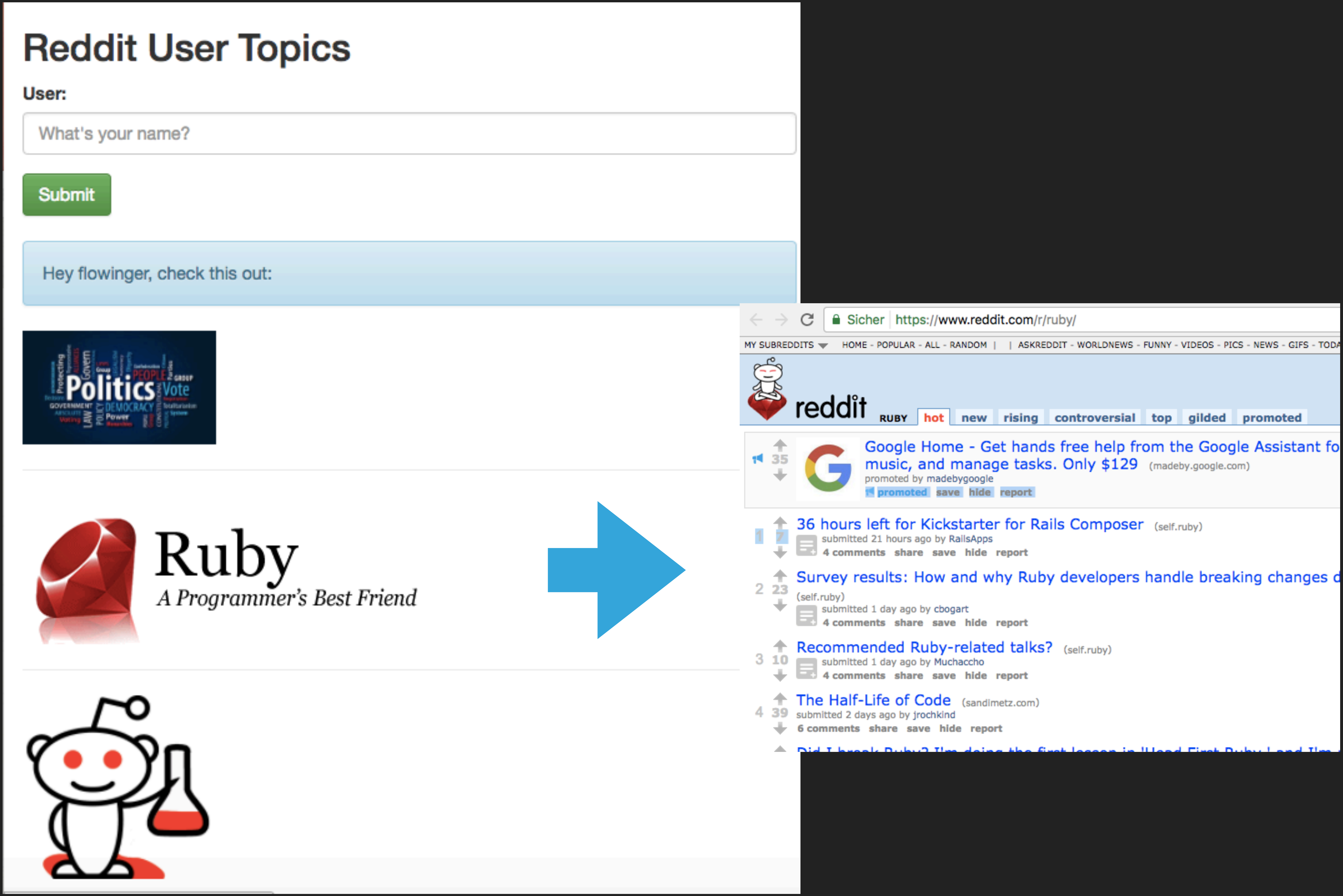


Topics found using Latent Dirichlet allocation





VISUALIZATIONS - FLASK APP



Reddit User Topics

User:

What's your name?

Submit

Hey flowinger, check this out:

Sicher https://www.reddit.com/r/ruby/

MY SUBREDDITS HOME - POPULAR - ALL - RANDOM | | ASKREDDIT - WORLDNEWS - FUNNY - VIDEOS - PICS - NEWS - GIFS - TODA

reddit RUBY hot new rising controversial top gilded promoted

35 Google Home - Get hands free help from the Google Assistant fo music, and manage tasks. Only $129 (madeby.google.com)
promoted by madebygoogle
promoted save hide report

1 7 36 hours left for Kickstarter for Rails Composer (self.ruby)
submitted 21 hours ago by RailsApps
4 comments share save hide report

2 23 Survey results: How and why Ruby developers handle breaking changes d (self.ruby)
submitted 1 day ago by cbogart
4 comments share save hide report

3 10 Recommended Ruby-related talks? (self.ruby)
submitted 1 day ago by Muchaccho
4 comments share save hide report

4 39 The Half-Life of Code (sandimetz.com)
submitted 2 days ago by jrochkind
6 comments share save hide report

Did I break Ruby? I'm doing the first lesson in 'Head First Ruby ' and I'm

FUTURE WORKS

- Use more meta data as features
- Use all of the available data (submissions etc.)
- Word2vec

SOURCES

https://files.pushshift.io/reddit/