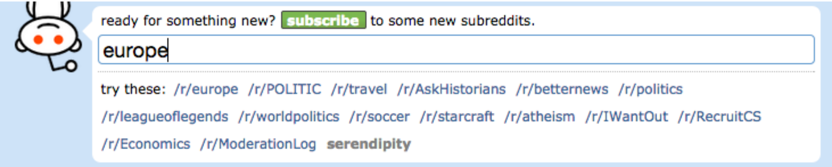
haventreddityet.com

A reddit ~~recommendation~~ *discovery* engine

James Douglas Pearce

HOW CAN I DISCOVER NEW SUBREDDITS?



ready for something new? to some new subreddits.

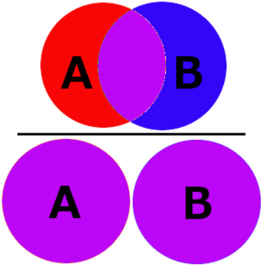
europe

try these: /r/europe /r/POLITIC /r/travel /r/AskHistorians /r/betternews /r/politics /r/leagueoflegends /r/worldpolitics /r/soccer /r/starcraft /r/atheism /r/IWantOut /r/RecruitCS /r/Economics /r/ModerationLog **serendipity**

- Results seem to be skewed by a few very popular subreddits...
- Simple list
- No context
- Does not encourage discovery!

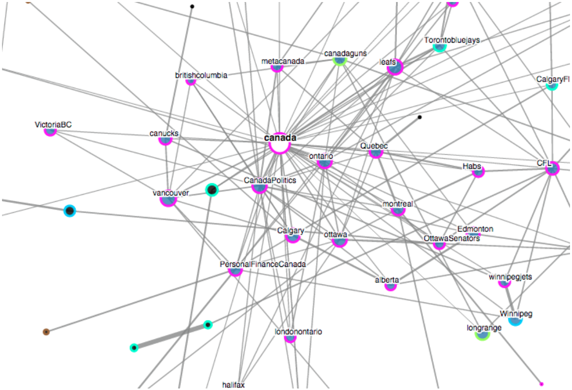
Live Demo

ALGORITHM



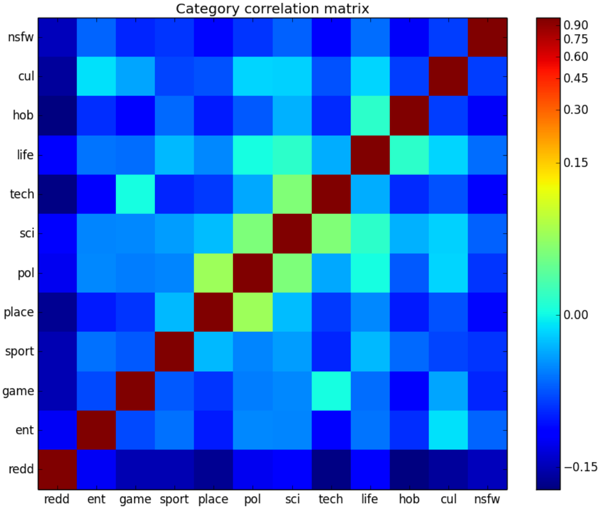
- Jaccard similarity: overlap of user in subreddits divided by total
- Categories are defined as a set of example subreddits

DATA

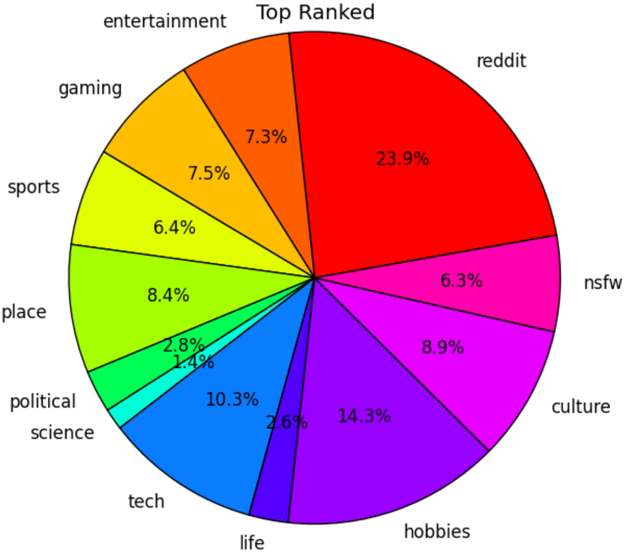


- Collect data by crawling over reddit
- Random walk: subreddit → redditor

UNIQUE CATEGORIES?



REDDIT BY CATEGORIES



MORE ABOUT ME

Research:

- Searching for Dark Matter at the LHC
- Big Data!
- Specialize in data mining and machine learning

Hobbies:

- Kaggle competitions
- Poker
- Motorcycles
- Boardgames

Questions?

JACCARD COEFFICIENT

The Jaccard coefficient can be used as a user-user type similarity measure.

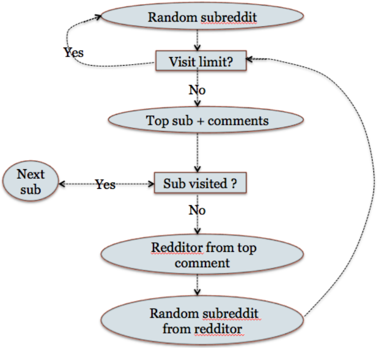
$$J_B(A) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

- A is the set of redditors in subreddit A,
- B is the set of redditors in subreddit B.

$$J_C^{Cat}(A) = \frac{1}{|\mathrm{C}|} \sum_{B_i \in C} J_{B_i}(A) \tag{2}$$

- C is the set of sets of redditors in subreddits $\mathrm{B_i}$ in category C,
- $|C|$ is the size of set C (number of sets).

CRAWLING REDDIT WITH THE PRAW API

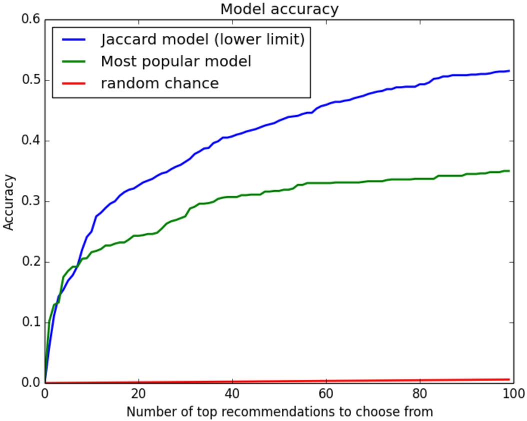


- reddit.com is HUGE, I can only take a small sample
- The redditor-subreddit matrix I am sampling from is sparse
- I want to collect information about as many subreddits as possible
- I want my redditor sets to overlap

Strategy:

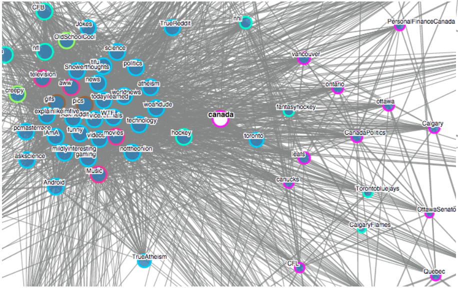
- Collect “smart” data
- Use “dumb” (Jaccard) algorithm

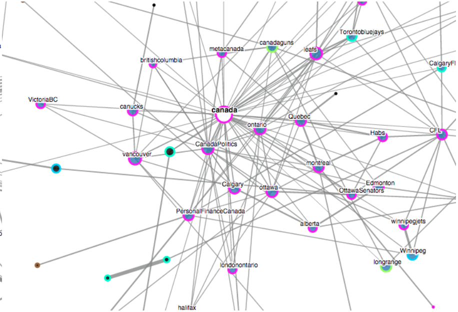
MODEL VALIDATION



- For each redditor hold out one subreddit they are part of
- Make a list of the top *N* subreddits based on Jacquard similarity
- Calculate the accuracy of that recommendation as a function of *N*

SUBGRAPH GENERATOR





$$\mathrm{P}_{\mathrm{transition}}(\mathrm{n}_{\mathrm{i}}) \propto J_{canada}(n_i) \cdot \alpha^{n_{trans}} \cdot \beta^{n_{con}} \tag{3}$$

- $\alpha \in (0,1)$ is a transition decay factor
- n_{trans} is the number of times n_i has been traversed
- $\beta \in (0,1)$ is a connectivity decay factor
- n_{con} is n_i number of connections (degree)