**General**

* Online vs **offline** model and disadvantages – offline model may miss new subreddits
* Select user list from latest month and backtrack to build training set
* How much data is enough data – need to do EDA

**Idea**

* Utilize historical reddit comment data to build subreddit recommender system

**Data Sources**

<https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/>

http://files.pushshift.io/reddit/comments/

https://praw.readthedocs.io/en/stable/

**Objective**

* Build a neural computational based recommendation system that provides suggestions of potential subreddit subscriptions to a user based on their historical site usage that reports reasonable results as decided by test subjects.
* Investigate methodologies using neural computation for recommendation systems and compare their results when applied to this specific project
* Compare to other more traditional RS models?

**Design Questions**

* Will we build a UI/product, or just report experimental results? - Stages
* If UI, a bot or webpage or …?
* Success Criteria – better than random? User feedback? Exploring NN and comparing.
* Limit number of subreddits? (top 100,1000,1000000?)
* How many months of data do we need – do it iteratively or grab list of users and backtrack.
* Data structuring sparse matrix? How to incorporate timescale
* Compare to more conventional techniques?

**Feature Engineering**

* Order of user added subreddits
* Weighted subscribed subreddit by total comment count
* NLP of user comments – **No -** Future work

**ScratchPad**

* Measurement of success could be user engagement in suggested subreddits after suggestion
* Proposal could be phased – Phase1 is experimental results and optional phase2 is UI/product development

**Research Papers**

<https://cs224d.stanford.edu/reports/LiuSingh.pdf>

<https://arxiv.org/pdf/1606.07792v1.pdf>

http://benanne.github.io/2014/08/05/spotify-cnns.html

**Tooling**

* Python – data-munging and exploring
* <https://www.tensorflow.org/> - NN model development
* mongoDB - database
* flask – webapp
* d3.js – visualization
* PRAW reddit python API