# Subreddit Recommendation System

Team: r/datascience Members: iting, klee50, mjacks12, yma37

## Goal

As Reddit currently offers no in-house recommendation system, our goal is to test two algorithms, Alternating Least Squares and Bayesian Personalized Ranking, in order to create a subreddit recommendation system that can help users discover subreddits they might enjoy but have never interacted with, based on their current interactions with other subreddits.

## Data

We first used calls to the Pushshift API to gather a selection of 5,000 users from the most frequent commenters in the top 100 subreddits in January 2020. This was not enough data to train a well-functioning model, so we augmented our data with a Pushshift directory of all comments from January of 2013, only including our original observations for unique users and newer subreddits. We excluded comments from deleted users and known bots. Our final dataset tracks the number of comments left by each user in each subreddit, spans 8,181 subreddits and 355,639 users. Bias in this dataset consists of potentially missed bots and the bias towards frequent commenters found in our original (appended) dataset.

## Model+Evaluation Setup

As Reddit provides no explicit way to rate subreddits, our systems rely on implicit feedback in the form of comments. Given the abundance of implicit recommendation systems, we test two algorithms. Alternating Least Squares decomposes our large matrix of user/subreddit interaction into a user matrix and a subreddit matrix, which we use to recommend subreddits. Bayesian Personalized Ranking aims to tailor recommendations for users by considering the level of interaction of subreddits during training.

Our chosen evaluation method is AUC, presented as the mean AUC across users. We trained our models with 20% of our original non-zero data masked off. For testing we used a binarized version of our original data set. In this evaluation setting, false positives are values for which the model incorrectly predicted a positive interaction (the user never interacted with the subreddit), while true positives were correctly predicted (the user did interact with the subreddit). We also offer as comparison the AUC score for suggesting only the 100 most popular subreddits.

## Results and Analysis

**Claim 1:** Both ALS and BPR perform better than a baseline recommendation of just the most popular subreddits.

**Support for Claim 1:** As we can see from the table, both ALS and BPR have higher AUC scores than our baseline model, which simply suggests the 100 most popular subreddits every time. ALS has a higher AUC score than BPR, implying that it is better at distinguishing between relevant and irrelevant subreddits.

| Algorithm | AUC Score |
|---|---|
| Alternating Least Squares | .802 |
| Bayesian Personalized Ranking | .753 |
| Baseline (most popular) | .553 |

**Claim 2:** ALS works better than BPR.

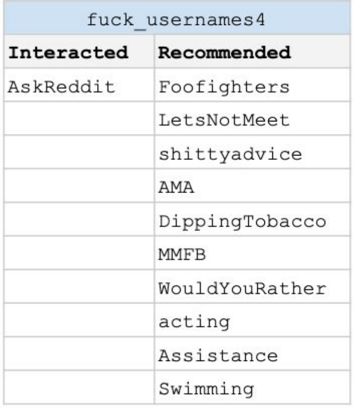
**Support for Claim 2:** As a sanity check, we also wanted to take a look at the outputs of our models. We randomly selected 20 subreddits and based on their ALS and BPR scores and found their top 10 related subreddits.

| Subreddit | | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| **simpleliving** | **ALS** | minimalism | Anticonsumption | PhysicGarden | Frugal |
| | **BPR** | AskPhysics | ontario | SelfSufficiency | Anticonsumption |
| **animenews** | **ALS** | demonssouls | darksouls | Blazblue | anime |
| | **BPR** | PERU | MouseReview | AverageMisfires | PSO2 |

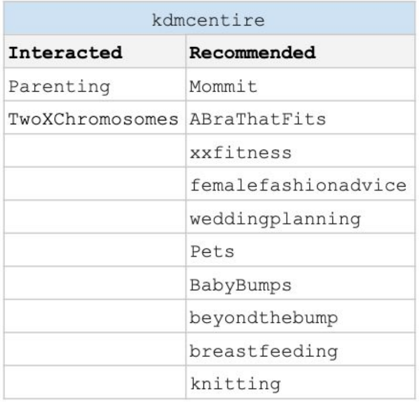
BPR tends to output unrelated subreddits, such as r/AskPhysics for r/simpleliving and r/PERU for r/animenews. ALS seems to be much more consistent with outputs that relate to the topic of the input subreddit. This, combined with its higher AUC score, ultimately led us to choose the ALS matrix factorization for the basis of our recommender.

**Claim 3:** Our ALS recommendations qualitatively look good, especially for users showing a major interest; however, if users are subscribed to only default or "unspecialized" subreddits, our recommendations become more difficult to judge.

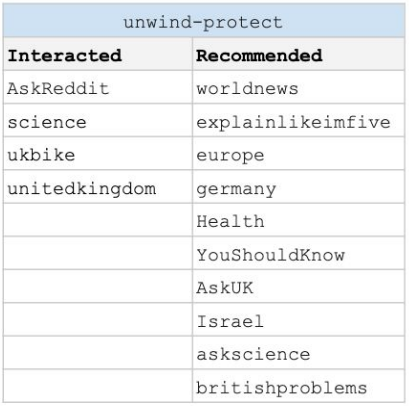
**Support for Claim 3:** By "unspecialized", we mean subreddits that are open-ended or centered around a very broad subject like AskReddit. Furthermore, it is a default subscription subreddit, meaning that every new user automatically subscribes to it. We analysed our data qualitatively by randomly selecting 20 users and creating recommendations for them, determining that the results reflected the performance we expected from AUC scores.



| fuck_usernames4 | |
|---|---|
| **Interacted** | **Recommended** |
| AskReddit | Foofighters |
| | LetsNotMeet |
| | shittyadvice |
| | AMA |
| | DippingTobacco |
| | MMFB |
| | WouldYouRather |
| | acting |
| | Assistance |
| | Swimming |



| kdmcentire | |
|---|---|
| **Interacted** | **Recommended** |
| Parenting | Mommit |
| TwoXChromosomes | ABraThatFits |
| | xxfitness |
| | femalefashionadvice |
| | weddingplanning |
| | Pets |
| | BabyBumps |
| | beyondthebump |
| | breastfeeding |
| | knitting |



| unwind-protect | |
|---|---|
| **Interacted** | **Recommended** |
| AskReddit | worldnews |
| science | explainlikeimfive |
| ukbike | europe |
| unitedkingdom | germany |
| | Health |
| | YouShouldKnow |
| | AskUK |
| | Israel |
| | askscience |
| | britishproblems |

User *fuck_usernames4* has interacted only with r/AskReddit, so our recommendations are somewhat all over the place, due to how widely-interacted-with r/AskReddit is. User *kmdcentire*, meanwhile, has interacted with both r/parenting and r/twoxchromosomes, which implies that they are likely a mother. Our recommender seems to pick up on this subject of interest, suggesting subreddits that make sense for a mother to be subscribed to.

User *unwind-protect* seems more representative of a typical user: they interact with both generic (r/AskReddit) and specific (r/science, r/ukbike) subreddits, and our recommender makes a good mix of recommendations that mostly seem relevant. r/worldnews and r/explainlikeimfive are more default-subscription subreddits; though they don't seem related to r/AskReddit in topic, it's possible that many users only interact with the default-subscribed subreddits, which is why they are often suggested when given one as input. Subreddits like r/health, r/askscience, and r/britishproblems seem directly related to this user's specific interests, and despite the few seemingly out-of-place suggestions, our recommender definitely looks to be working well for this user.

## Sources

"AlternatingLeastSquares." AlternatingLeastSquares - Implicit 0.4.0 Documentation. https://implicit.readthedocs.io/en/latest/als.html.

Huang, Lin. "Stacking Collaborative Filtering for Implicit Feedback." doi:10.14711/thesis-b1106722.

Steffen, Freudenthaler, Christoph, Gantner, and Lars. "BPR: Bayesian Personalized Ranking from Implicit Feedback." ArXiv.org. May 09, 2012. https://arxiv.org/abs/1205.2618.

"BayesianPersonalizedRanking." BayesianPersonalizedRanking - Implicit 0.4.0 Documentation. https://implicit.readthedocs.io/en/latest/bpr.html.