r/ecommender: A Personalized Subreddit Recommendation Engine

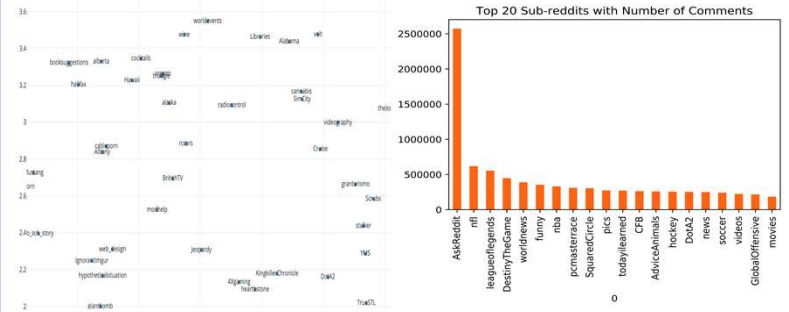
Abhishek Das , Janvi Palan
Sukanto Guha, Nikhil Bhat

Introduction

Reddit, with more than 330 Million Active users and 138K+ subreddits, lacks in its subreddit recommendation system, which usually provides generic recommendations. In this project, we propose a novel personalized recommender system that learns from both, the presence and the content of user-subreddit interaction.

Dataset

- Our dataset comprises of user comments on Reddit from the month of January 2015.
- It contains 57 million comments (32 GB) from reddit users. We pre-processed the data to remove bots, [deleted] comments, comments with less than 30 characters, and users with less than 5 comments.
- Final dataset: **735834 users, 14842 Subreddits, 28 million comments.**
- We used Stratified Sampling for splitting data.



Methodology

Using presence of comment of a user on a subreddit as implicit signal of interest:

- **Matrix Factorization using ALS:** We used Matrix factorization along with ALS as the learning algorithm to decompose the user-subreddit interaction matrix into the product of two lower dimensionality rectangular matrices.
- **BPR**: The Bayesian Personalized Ranking optimizes the Matrix Factorization techniques on Implicit Data to rank the recommendations rather than just giving absolute recommendation scores.

Using comment text along with presence of comment to include user-subreddit interaction:

- **textual-BPR (t-BPR)**: Inspired by the V-BPR paper, we generated embeddings of the comments made by the users as textual factors for our model.

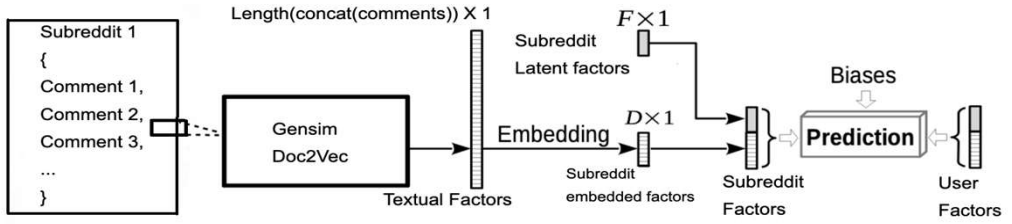
$$\hat{x}_{u,i} = \alpha + \beta_u + \beta_i + \gamma_u^T \gamma_i + \theta_u^T \theta_i \quad \ldots\ldots (i)$$

where, α → global offset, β_u, β_i → user bias, subreddit bias,
γ_u^T, γ_i → Latent factors for user u, subreddit i,
θ_u^T, θ_i → Textual factors for user u, subreddit i

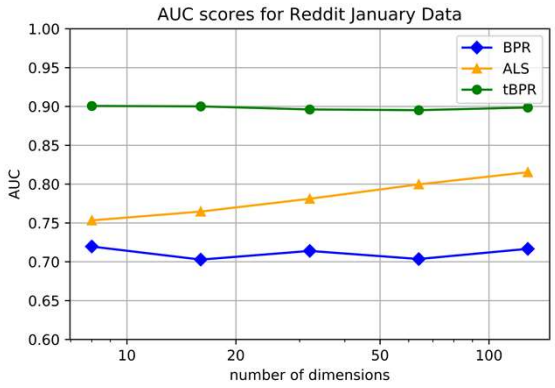
We create 2 sets of textual factors: **Subreddit2Vec (θ_u)** and **User2Vec (θ_i)**

- Vanilla t-BPR: We use (i) to calculate the predictor for each user-subreddit interaction.
- Learnt t-BPR: The textual factors of user embeddings, θ_u are learnt using the loss function:

$$\sum_{(u,i,j) \in Ds} \ln \sigma(\hat{x}uij) - \lambda_\theta \|\theta\|^2$$



Evaluation and Results



Model	AUC	Dims
ALS	0.815	128
BPR	0.717	32
t-BPR	**0.901**	**16**

How did we do?

u/TallnFrosty
Follows: nba, Gunners, 49ers
Our Recommendations: realmadrid, sportsbook, lakers, NYCFC

Subreddits recommended for r/gameofthrones: asoiaf, ArcherFX, CK2GameOfthrones, thewalkingdead

Discussion

- We were able to successfully demonstrate that embeddings of the comment strongly indicate user preference.
- ALS & BPR give good recommendations in practice, but recommendations by t-BPR are more specific.
- We can further explore and improve the novelty, diversity and serendipity of our recommender, as just using the AUC as an evaluation metric does not show the complete picture.