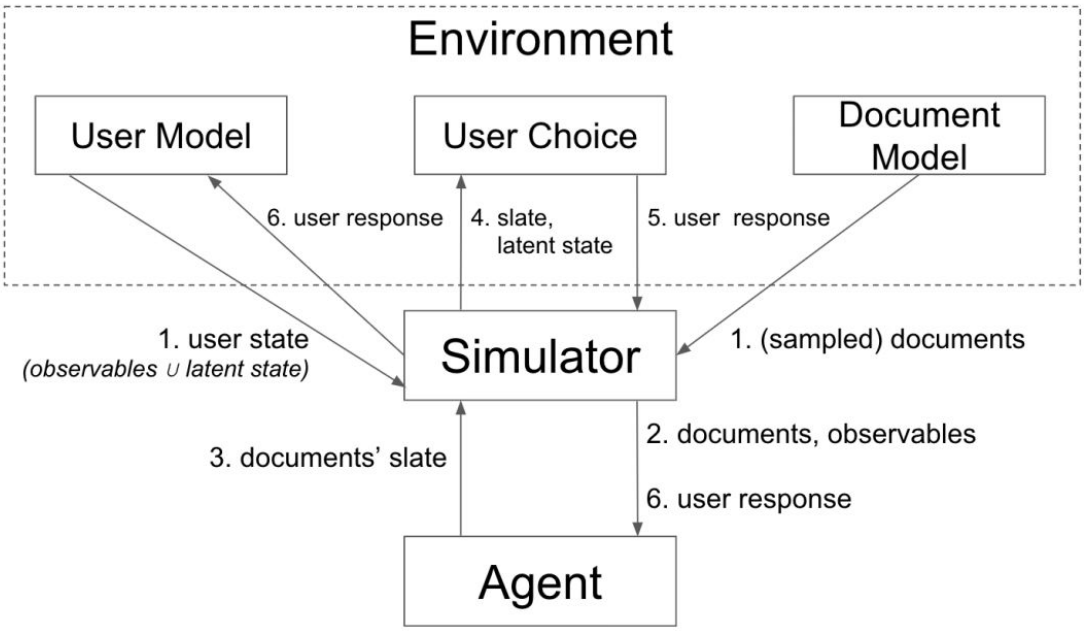
Project Proposal

Name: Collin Prather and Shishir Kumar

Research Question:
Most practical recommender systems focus on estimating immediate user engagement without considering the long-term effects of recommendations on user behaviour. Reinforcement learning (RL) methods offer the potential to optimize recommendations for long-term user engagement. However, since users are often presented with slates of multiple items—which may have interacting effects on user choice—methods are required to deal with the combinatorics of the RL action space.

Google's SlateQ algorithm addresses this challenge by decomposing the long-term value (LTV) of a slate into a tractable function of its component item-wise LTVs. For our project, we want to compare the efficiency of SlateQ to other RL methods like Q-learning that don't decompose the LTV of a slate into its component-wise LTVs.

Reinforcement Learning problem formulation:



Sequential: A user makes a series of selections (i.e. clicks) from a slate of recommended documents. Each user is allocated an initial budget of time to engage

with content during an extended session. Each document consumed reduces the user's budget by a fixed amount. But after consumption, the user's budget is replenished by an amount that depends upon the appeal of that document for the user. Hence the time steps are discrete and finite.

Environment: We will use the recsim environment for training models. It consists of:

- Set of documents in slate, selected by a document model from a fixed corpus
- User interacting with the given slate, defined by its interest, satisfaction and choice models.

Agent: The recommender system is our agent in this case. At each stage of interaction with a user, 'm' candidate documents are picked from 'P' corpus of documents and the recommender system chooses a slate of size 'k' for the recommendation.

Example state: The agent recommends a slate of 'k' documents depending upon the user's interests, satisfaction and choice models.

Example reward: Each user is allocated an initial budget of time to engage with content during an extended session. Each document consumed reduces the user's budget by a fixed amount 'l'. But after consumption, the user's budget is also increased by a bonus 'b' < 'l' that increases with the document's appeal to the user.

Data: Data will be generated from the environment by selecting appropriate model parameters for the user model, user choice model and document model.