**Pocket Gems**

**Data Science Homework**

Thank you for your interest in Pocket Gems. This is the Data Science Homework. You may use whatever programming language you want (R, Python, Excel, TI-85, etc.). Please send your code and answers to the questions at the end.

Please try not to spend more than 6 hours on this problem. If you do, that’s OK, but you could easily spend a week on this problem. This is a homework problem and we are more interested in seeing how you think and work. If you are confused or need help, do not hesitate to call Ryan Maddux at 650-387-3253 or e-mail [ryanmaddux@gmail.com](mailto:ryanmaddux@gmail.com)

Good luck and have fun!

Also, please read the entire problem before beginning.

**The Context:**

Playing Episode may help you gain context for this question. Episode is a story telling game here at Pocket Gems. In the game, we ask players to start reading content in 2 onboarding stories (they can switch between the stories if they like). After they’ve gotten through 5 chapters of their onboarding material, we open up more material to them. Players then can move to other stories and start consuming other content. Players consume content one chapter at a time.

The way that Pocket Gems makes money in Episode is by either:

1. Showing advertisements at the beginning of chapters
2. Having players pay to access additional content beyond what is given away for free

In this sense, Episode profits when players read more content, which, hopefully, they enjoy! So, one way to make Episode a more profitable product is to match users with content that they will engage in.

In this homework, we’d like you to assess what’s going on in stories and put together some simple recommendations for different groups of players (i.e. a baby recommendation system based on limited data).

**The Data:**

The data consists of two tables:

1. A table that includes a user’s identifier, their install date, and their device type (i.e. which Apple device they are using) for a sample of users who installed the game in May of 2015. User identifiers have been anonymized. Data has been provided for a little over 14,000 users. This data and time constraints are likely not sufficient for great recommendations, so don’t worry about putting together anything amazing.
2. A data set with the read history for all users in the sample. This data set lists the user ID, the story name, the chapter name, the time stamp from when they read the story, and the date they read the story.

**The Homework**

You may use any programming language you like. Please send the code that you used to generate your results as well as answers to the following questions:

1. Which stories are “good” stories?
   1. What’s your metric for “good”?
   2. What’s your support for this conclusion?
   3. Of the stories that are “good” stories, are there any factors that should be noted?
2. Build a set of recommendations based on the data. Your recommendations may be very complex and can be reported in any way you feel expresses what you’re trying to do
   1. Explain your methodology.
   2. Describe a couple of the results for context.
   3. How confident are you in your results?
3. Outline one of the assumptions that you made when building your model. What’s the danger in using this assumption?
4. What other data would you want and why? How would it affect your analysis?
5. Briefly explain how you would scale this/implement it into a system.

Thank you again for interviewing with Pocket Gems!!!  
  
Ryan