Digital Media Laboratory

Weekly Report August 30, 2021

Ali Jandaghi Alaee
alialaee98@gmail.com

Report No. 6
00/06/02- 00/06/08

Abstract

First of all, the whole input data was divided into training and validation sets, then Skip-Gram algorithm was executed on the training data and the output matrix saved. Moreover, a piece of code was written to extract the most repeated clicks for each validation query. This program will be used to compute nDCG in the immediate future.

Description

- The input data was split into training set and validation set with an approximate ratio of 85:15. After that, Skip-Gram algorithm was executed on the training data and the results saved into drive storage for further evaluation. Two saved files are named as learned_weights.tsv, which represents the embedding, and vocab.json, which contains the vocabulary.
- A new google collaboration was set up to evaluate the outputs of the model using validation data set. To begin with, a function was constructed to calculate the K nearest points to a particular point in the embedding matrix, using cosine similarity. It then improved to deliver the K nearest categories to a given point. Furthermore, another piece of code was developed to find the frequency of clicks corresponding to each query in the validation set. With these two, the nDCG assessment is possible if an effective discrete rating policy applies, with respect to the frequency of clicks.

Next Week

1. Finding and implementing the best discrete policy to complete the nDCG.
2. Construction of a frame for managing the tale queries.
3. Deciding what to do with products that have not been learned because they do not have any clicks associated with them.

References