Vijay Sarathy

Vimeo Coding Exercise

Approach

The main manipulation that I have done to the datasets involved mapping the category names to each movie. I then concatenated the tile, caption, and categories into one field and used the bag of words model, which justifies the validity of the concatenation mentioned- the order of words and semantic relations are not a factor. I used TF-IDF to generate a sparse matrix of each possible word appearing in each movie listing. I then used cosine similarity among each entry in the matrix to determine the 10 most related other titles for each movie. Finally, I manipulated the results for the movies of interest into a single json file.

What I Liked/Disliked

I liked that this is a real application and there are a lot of different skills that it involves. I liked building something from the ground up. I think for further improvement it would be nice to have a module to start with and to build on top of that as it could work as both a way to get to know the team’s coding/ problem solving style and to be assessed on how to improve/ use someone else’s code for a slightly different problem.

Scaling Up

Building it for real

I think my approach may not scale well because more memory would be consumed to store a larger sparse matrix if many more movies were to be added. I think I would instead switch to a deep learning method, continuous bag of words to still not worry about word order, and to leverage the GPU to perform the large number of matrix multiplication calculations upon averaging the weights instead of having a sparse matrix.

Additional Techniques

I think what I mentioned above may also qualify in this category. To go a bit further, I think using a distributed computing framework is also necessary to distribute the load between the possible calculations.