Deliverable 3

Civera

April 7, 2021

1. All data is collected
   1. We’ve collected all the data from the SQL table with the relevant information needed for normalizing the actor and action from the description.
   2. This table was pre-processed, reduced to its relevant features and vectorized for supervised and unsupervised learning.
2. Refine the preliminary analysis of the data performed in PD1&2
   1. In working with this dataset and attempting to manipulate it to meet the client’s needs,
   2. First, we clustered on descriptions missing the actor, then
   3. We’ve tried both supervised and unsupervised learning methods.
      1. Supervised Method we have tried:
         1. RandomForest Classification
         2. Naive Bayes Classification
      2. Unsupervised Method we have tried”
         1. Kmeans classification with tfidf
         2. Kmeans classification with GloVe embeddings
   4. It appears, at this point, that clustering is more advantageous but with the volume of data, finding the optimum number of clusters is a key objective and very important before proceeding with the cluster inference and action/actor extraction using semantic role labelling.
   5. Converting PHP regex code to Python showed us that regex is very limited to giving a generalized result. The existing regex code is hardcoded with select/certain actors and actions to search for, which does not help when new distinct actors and actions are added to the data. Although the existing regex code covers the most frequent actors and actions, it is not sustainable for future new data and requires manual additions to regex code for new distinct entries
3. Answer another key question
   1. Our analysis and pre-processing indicates that clustering and semantic analysis will yield the best results. The passive nature of the descriptions in this dataset seems to work best with GloVe embeddings.
4. Attempt to answer overarching project question
   1. While still under analysis, it appears that unsupervised learning will be our choice for meeting this client’s objectives. We are using k-means++ and SSE graphs to determine the optimum number of clusters.
   2. Have converted existing regex code to python (per clients request), but decided to focus on supervised and unsupervised learning models and move away from the hard coded classification.
5. Create a draft of your final report

\*see “Project Report”

1. Refine project scope and list of limitations with data and potential risks of achieving project goals.
   1. The volume of the dataset is difficult and time-consuming to handle on a fixed schedule. We’re hopeful that the SCC will be able to pre-process it more quickly.
2. Submit a PR with the above report and modifications to original proposal