Civera - Project Deliverable 2

More data should have been collected to perform a more thorough analysis of the data and attempt to answer one additional question relevant to your project proposal which you will submit as a pull request.

Checklist

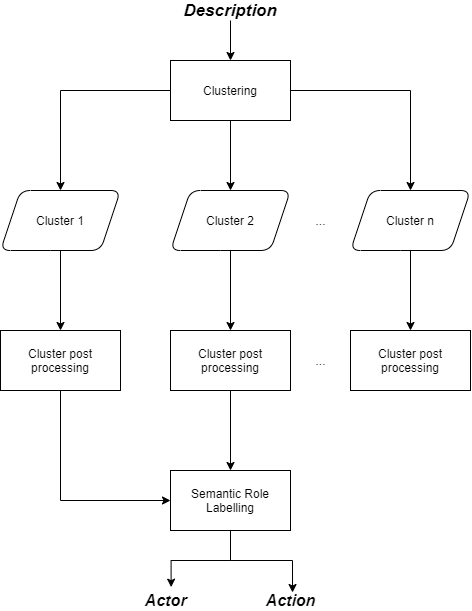
As part of this deliverable, we have also pushed our code. See the *Code* folder in *Civera*.

1. Collect and pre-process a secondary batch of data
   1. We made some data analysis on the samples that have both action and actor pairs. We gathered information that could answer these questions.
      1. How many unique actors? How many unique actors?
      2. How many unique pairings?
      3. What pairing is most common?

This Information is available in the data folder of our github.

* 1. Half of the group is working on unsupervised learning strategies to cluster the data. Specifically we are trying Glove embeddings and TFIDF to cluster the data and then try Semantic Role Labelling to extract action and actor from the descriptions.
  2. Half of the group is trying supervised learning strategies to classify the data to extract actor and action from the descriptions.
  3. Another group member is working on converting existing php regex code from the client into python, to give us a third method for processing the data. The regex PHP code has to be grouped in Python and we are using example strings/sentences to test if the functionality of the regex code works the same in Python as in PHP.

1. Refine the preliminary analysis of the data performed in PD1
   1. We have written code to pull data in chunks from the database.
   2. We are continuing to transform/update the clients existing regex code.
   3. After these steps were complete from deliverable 1, we have begun to classify the data using different strategies to identify the critical action and actor fields.
2. Answer another key question - How we plan to normalize the data...
   1. We found that because case descriptions are written in the passive voice, semantic analysis was going to be even more difficult. We observed case actors and case actions were best clustered using unsupervised learning methods like KMeans on tfidf vectors of text descriptions or KMeans on glove embeddings of text descriptions.
   2. Once we have the clustered texts, we do some post processing to improve the clusters.
   3. After the post processing step, we use semantic role labelling to extract the actor and action from the passive voice text descriptions.



* 1. We are working on identifying case\_action using forest classification.