Project 3 Report

Kofi Adu-Gyan (z1723260)
Joshua Boley (z1698317)

April 20, 2021

Analysis

The analysis phased was composed of several steps. An initial LDA analysis revealed that, using only the standard word vectorization techniques, there was little significant cross-document clustering around the entities of interest in the reports, such as the individuals and events being reported. In addition, named entity recognition with standard pre-built English vocabularies did not do an adequate job extracting entities of interest for a threat analysis. This required additional steps to overcome.

First, we used the open-source Doccano tagging software to create custom entity definitions and extract the entities from the report documents. Unfortunately this was a largely manual process, however the resulting vocabulary of custom entities was found to work well. A special category of entity, "Red Flags/Suspicious" was defined for special terms in the corpus that would likely merit close scrutiny, such as "explosives training facility". Supplementing this was an "Event" category used to capture words that might draw the investigator's interest such as "arrested".

Latent Dirichlet Allocation was performed on the document corpus using word vectorizations generated from the custom vocabulary and a custom tokenizer that filtered all words from the report documents except the tagged entities. Contrary to original expectations, the LDA does not reveal information or patterns relevant to specific threats. However, it can be used to identify statistically-significant clustering ("themes", if you will) among the entities tagged in the documents that are ideal for giving an investigator an idea of where to begin looking for threats.

Design Considerations

The Jigsaw software possessed some useful tools for finding entity relationships across a document corpus, however its biggest drawback is that unless the investigator already has some familiarity with the case there is no clear place to start. Our tool is largely intended to address this in an intelligent fashion by allowing the investigator to explore not only the entities identified in the corpus but be able to get a high-level idea of how user-defined groups of entities cluster across documents.

The main view of the tool is composed of three areas. The first allows the user to investigate the entities by type (i.e., category) via a drop-down selector that populates a selection box with all of the entities belonging to that category. Both the drop-down and entities populating the selection box are color-coded by category, using both background and text color. The same color-coding is applied to the entities in the document viewer, thus guaranteeing semantic consistency between views.

In terms of marks and channels, this part of the interface is very simple. We use color channels (foreground and background) to identify the type of the entity (category). The marks are the text of the entities and the rectangular blocks of background color on which they sit in the selection box.

The second area of the main interface contains a heatmap matrix relating selected entities to the topic clusters extracted by the LDA. The design here is also intentionally simple, using color channel to draw attention to entities (on the rows) that more strongly associate with a cluster (i.e., column). Colder colors indicate weaker relationships while warmer correspond to stronger associations. Entities and clusters are separated of course, so we can state that location is also used as a channel.

The third area of the main interface contains a list of documents, sorted by date, that most strongly associate with a selected cluster. Location may be considered a channel here as well, as the closer to the top of the list a document falls, the earlier the date of the report. Color is used as a channel here in the same way as the heatmap, with colder colors indicating weaker relationships and warmer colors stronger relationships.

Our document view aims to build a display of the original documents with highlighted views. Different areas of interest will be highlighted with varying highlight and text colors, and will even differ on color intensity based on the weight of the entity.

The user interface is very straightforward. After moving to this view, it will provide the display of the document selected from the previous view loaded up to the user after their import.

The data is processed by using jquery in order to load up the results of the LDA that is stored into a json file, and will dynamically build the document, paragraph by paragraph, while including the highlighted areas of interest.

Usage

Exploration begins in the entity selection panel. A drop-down allows for the selection of entities by their type. The default entities displayed in the selection box below are the "Red Flag/Suspicious", which should make for a good starting place. Selecting an item in the entities selection box and then clicking the left-arrow to the right adds the selected entity to the selection box on the far right. You may left-click and add multiple items at a time. Click on the "Update" button below to populate or change the entities displayed below in the "Cross-Document Term Clustering" view's heatmap. You may also remove entities from the right-most selection box by left-clicking the right-arrow button. The associated entries in the "Cross-Document Term Clustering" view will also be removed, if present.

The "Cross-Document Term Clustering" view shows the strength of association between selected entities and topics as a heatmap. The heatmap cells are annotated for disambiguation. Clicking on a column allows the user to examine the documents that most strongly associate with a topic.

The "Cross-Document Term Clustering" view is most effectively used to explore clustering of terms, beginning with those most strongly associated with likely threats (the "Red Flag/Suspicious" entities). The user may at their discretion add or remove entities from the view to determine which tend to cluster together. For instance, if the user were to select the "forged" Red Flag entity, then as a matter of course it could make sense to next add the "passport" and "passports" Artifact entities to determine if either is related, and then view the documents that most strongly correspond to that cluster.

The final view of the main interface, to the far right, is populated with the documents that most strongly associate with the topic whose column has been left-clicked. Reports are listed in ascending chronological order, and are color-coded based on their strength of association with the topic. A document relevance score is also given for each, for the purposes of disambiguation. Clicking on a report in the document selection view will bring up the contents of the report in a new view.