UI Labeling Interface Requirements

Overview

The ability to manually annotate documents and to review existing tags is crucial to evaluating and improving NLP applications. The manually created tags can be used as training data to improved Supervised Learning algorithms. They can also be used as validation and test sets to compare performance across different approaches, hyperparameters, and timeframes (model drift).

This tool and its output could be used during model/algo development prior to production deployment. It could also be used periodically after production deployment for ongoing model/algo validation and QA.

An annotation tool could also be used to correct mislabeled data in the Production database, but our understanding is this is out of scope for our purposes.

Several open-source tools exist offering varying subsets of the desired annotation functionality e.g. [INCEpTION](https://inception-project.github.io/), [BRAT](https://brat.nlplab.org/), and [Doccano](https://doccano.github.io/doccano/), and perhaps a survey of what is available is warranted. We have direct experience with entity labeling in INCEpTION. However, it also supports entity linking and Elasticsearch connectivity.

High Level Requirements

1. Load, Configure and Save the set of tags, and the set of resolution services to use.
2. Load a File or a Set of Files from File System (Text or JSON) or from Data Store e.g. Elasticsearch.
3. View Documents with Existing Tags and Resolved Entity Information
4. Add, Change, And Remove Tags
5. Add, Change, And Remove Resolved Entity Information
6. Save/Persist File(s) and Updated Tags and Resolved Entity Information as JSON to File System or to Data Store e.g. Elasticsearch
7. Compare and Generate Metrics on one set of tags versus a different set of Tags on the same Document or Document Set.
8. Recommenders: Given a document or document set to be labeled, prepopulate recommended entity labels to accelerate the manual process.

Requirements

1. Load, Configure and Save the set of tags, and the set of resolution services to use.
   1. We can load a preconfigured list of tag names/types allowed for this session via a

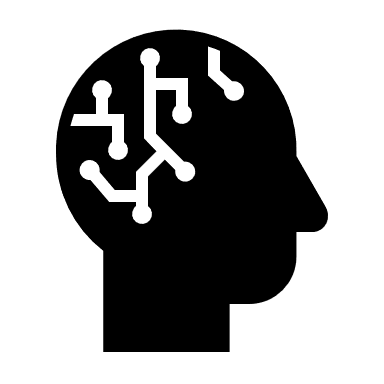
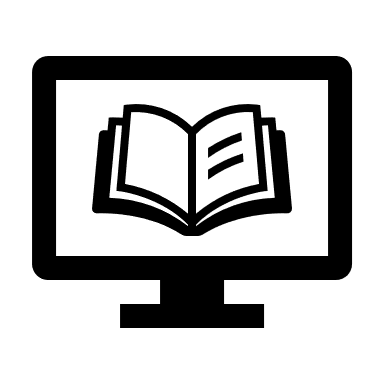
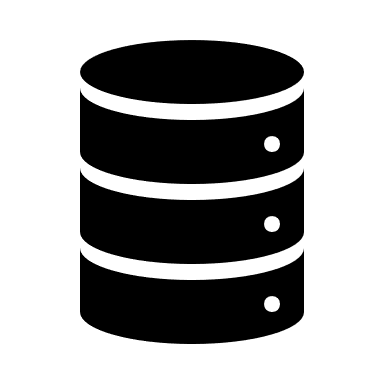
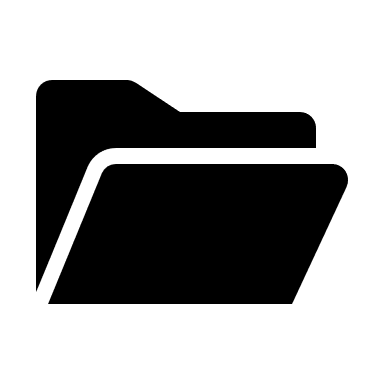
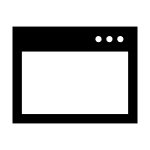
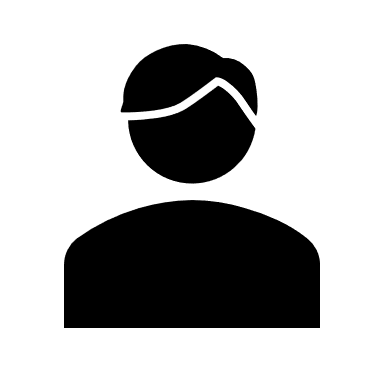
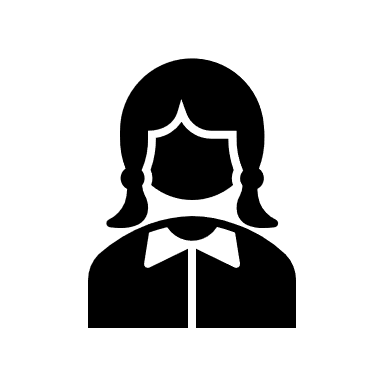
Load->Config menu option.

* 1. We can create or edit the list of tag names/types allowed for this session.
  2. We can save the list of tag names/types allowed for this session via a Save->Config menu option.
  3. We can load the resolution service URI for each tag type for this session via a

Load->Config menu option.

* 1. We can manually enter/edit the resolution service URI for each tag type for this session.
  2. We can save the list of resolution service URIs for this session via a Save->Config menu option.

1. Load a File or a Set of Files from File System (Text or JSON) or from Data Store e.g. Elasticsearch.
   1. Load plain text (untagged) file(s) from the file system via a File->Open->Text menu option.
   2. Load specially formatted JSON file(s) from the file system that include raw text, existing tags and resolved entity information via a File->Open->JSON menu option.
   3. Load documents, with existing tags and resolved entity information, from Elasticsearch via Document Id(s), or Search Filters.
   4. The loaded documents are then displayed in a list for individual selection. Alternatively, the first document is displayed with forward and backward buttons to navigate through the set.
   5. Question: Would it be useful to load CONLL format?
2. View Documents with Existing Tags and Resolved Entity Information
   1. When a document is selected from the loaded document list, it is displayed in a document viewer window.
   2. The text is not editable.
   3. If tags are already present, they are highlighted in different colors according to tag type.
   4. If entity information is already present it is displayed next to the tagged entity.
   5. The user can navigate across pages of documents using “Previous Page” and “Next Page” buttons. Alternatively, a scroll bar will allow scrolling through a document that is too large to display all at once.
   6. The user can navigate across documents using “Previous Document” and “Next Document” buttons.
3. Add, Change, And Remove Tags
   1. The user can highlight an untagged word or phrase and be presented with a dropdown featuring the list of tags available in this session.
      1. When a tag is added, the user will have the opportunity to resolve the entity as well (See 5 below)
      2. There will be a text box for the user to add a completely new tag which will be added to the session configuration.
   2. The user can highlight a tagged word or phrase and be presented with a dropdown featuring the list of tags available in this session. The existing tag would be preselected.
      1. When the tag is changed the user will have the opportunity to resolve the new entity as well (See 5 below)
      2. There will be a text box for the user to add a completely new tag which will be added to the session configuration.
   3. There will be an X next to each existing tag which the user can click to remove the existing tag.
      1. If a tag is removed any entity resolution information will be removed as well.
   4. Preexisting tags, new tags and edited tags will be presented in differing colors or color intensities to enable identification.
4. Add, Change, And Remove Resolved Entity Information
   1. Upon addition or edit of a tag, the user will be prompted to select the correct resolution.
   2. A dropdown will display the top ‘n’ resolutions for the user.
   3. If resolution is unsuccessful or the user is not satisfied, there will also be a text box with auto-completion that the user can use to manually perform/correct the resolution.
   4. The user can change the resolution information on an existing tag without changing the tag by clicking on the existing entity information box. They will be presented with the same dropdown interface described in 5b and 5c above.
   5. Preexisting, New, and Edited entity information will be displayed/highlighted in such as a way to easily differentiate them.
   6. Question: Should it be possible to leave entity info for new tags blank or remove existing entity info and leave it blank???
5. Save/Persist File(s) and Updated Tags and Resolved Entity Information as JSON or CONLL to File System or to Data Store e.g. Elastic Search
   1. The user can save individual tagged files or batches of files as JSON containing the latest tags and entity resolution information.
      1. It can be saved to the File System via a Save->File(s)->JSON menu option.
   2. The user can save the tagged document to a “scratch” Elasticsearch Index.
      1. There would be a Save->File(s)->Elasticsearch menu option.
   3. The user can export the tagged documents to the file system in a format suitable for NER training e.g. two-column CONLL (or BIO) format.
      1. There would be a Save->File(s)->CONLL menu option.
6. Compare and Generate Metrics on one set of tags versus a different set of Tags on the same Document or Document Set.
   1. The User can request comparison metrics between the tags currently loaded in the interface, and tags for the exact same document set either from JSON/CONLL file(s) on the filesystem or via doc ids in Elastic Search.
   2. The evaluation will provide Overall Accuracy, Precision and Recall.
   3. The evaluation can provide a Confusion Matrix as well.
   4. Question: Should we provide metrics on entity resolution as well?
7. Recommenders: Given a document or document set to be labeled, prepopulate recommended entity labels to accelerate the manual process.
   1. Provide an option to the user to request recommended labels be applied in a distinctive color or other identifiable presentation.
   2. A basic recommender would be a string matcher that automatically highlights string that are already labeled manually in other places in the current session/document set.
   3. More advanced recommenders would use regex matching or NER language models to identify common entity types like person-names and organizations.
   4. The user would have the ability to confirm or remove the recommended tags individually or collectively.



Users

Labeling UI

Tag Recommender

Knowledge Base

Documents & Tags

Filesystem

Elasticsearch

Entity Linking TagsEntity