Tagging Tool How To Guide

KEA Developers

June 7, 2018

1 About

1.1 Introduction

This application was designed to help manufacturers "tag" their data according to the methods described in [1, 2]. The goal of this application is to give understanding to data sets that previously were too unstructured or filled with jargon to analyze. The current build is in very early alpha, so please be patient in using this application. If you have any questions, please do not hesitate to contact Thurston Sexton (`thurston.sexton@nist.gov`) or Michael Brundage (`michael.brundage@nist.gov`). Future changes will be made through our public Github page, which will be available in the near future.

1.2 Terms of Use

This software was developed at the National Institute of Standards and Technology by employees of the Federal Government in the course of their official duties. Pursuant to title 17 section 105 of the United States Code this software is not subject to copyright protection and is in the public domain. `ml-py` is an experimental system. NIST assumes no responsibility whatsoever for its use by other parties, and makes no guarantees, expressed or implied, about its quality, reliability, or any other characteristic. We would appreciate acknowledgement if the software is used. This software can be redistributed and/or modified freely provided that any derivative works bear some notice that they are derived from it, and any modified versions bear some notice that they have been modified.

1.3 Disclaimer

The use of any products described in this toolkit does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

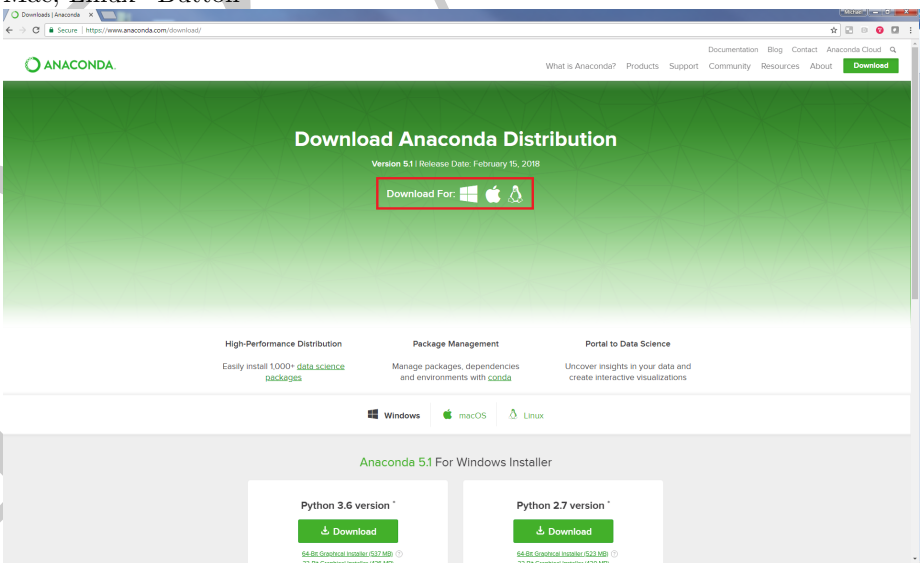
Contents

2 Downloading and Installing Anaconda

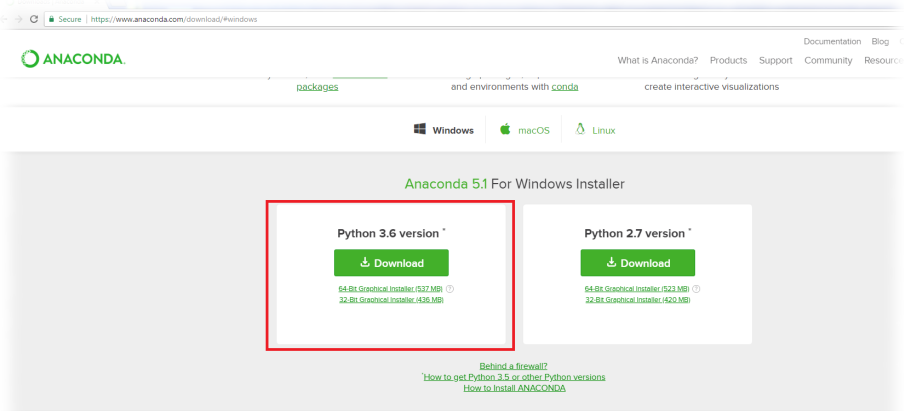
This section will walk through the steps for downloading Anaconda. The steps and pictures are for a Windows machine, but will be similar for other operating systems.

2.1 Downloading Anaconda

1. Visit the Anaconda website and click on the "Download For: Windows, Mac, Linux" Button



2. Click on the appropriate OS and select Python 3.6 version or higher [1]



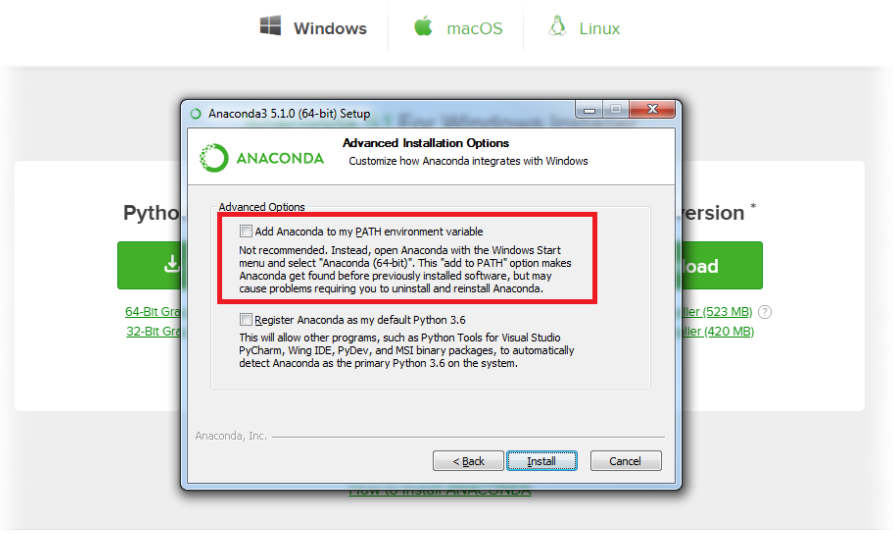
2.2 Installing Anaconda

1. Follow the step-by-step installation from Anaconda until you get to the screen below:

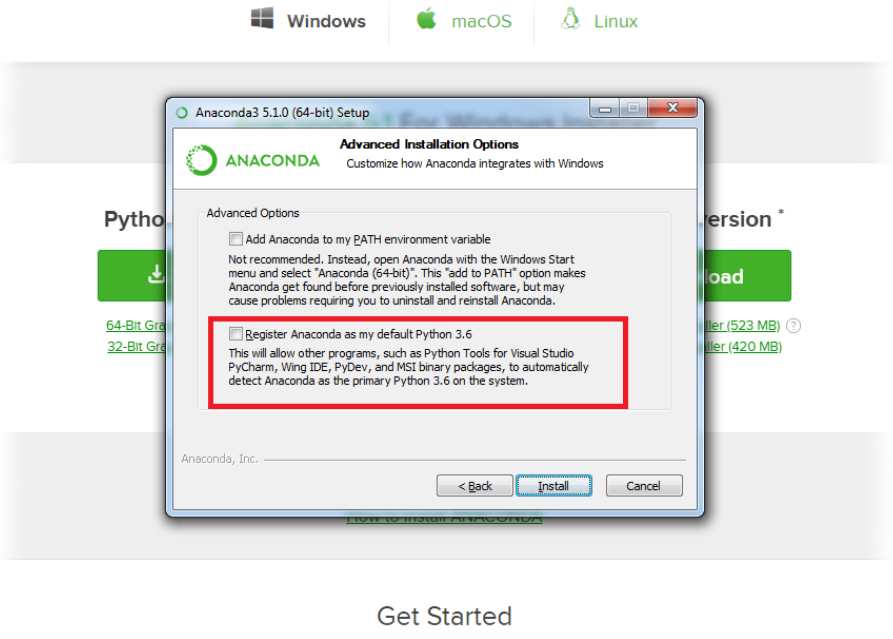
Get Started

2. Enable "Add Anaconda to my PATH environment variable". This registers `conda` as a valid command in any terminal/shell environment. (optional)

[1] This install guide will illustrate the install process for Windows.



3. Enable "Register Anaconda as my default Python 3.6" (recommended)



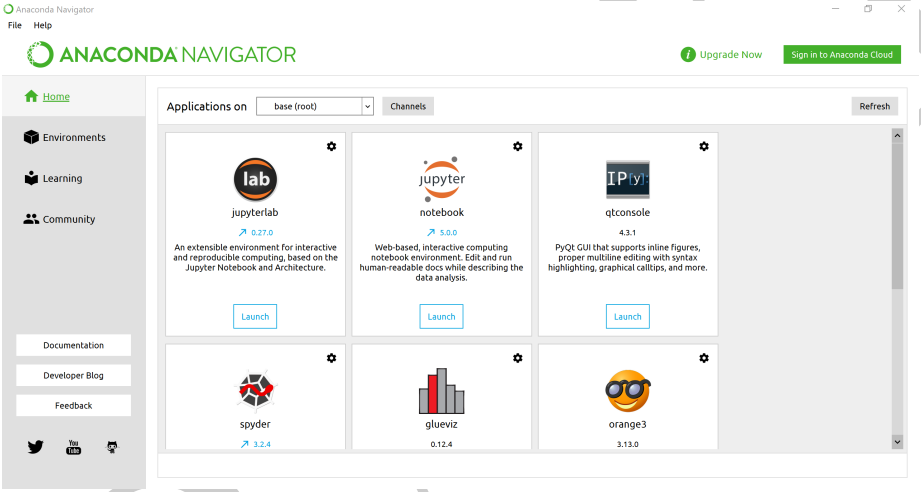
4. Finish Installation

3 Installing the Application and Importing the Environment

This section will walk through the steps for installing the tagging tool application and importing the correct environment to Anaconda.

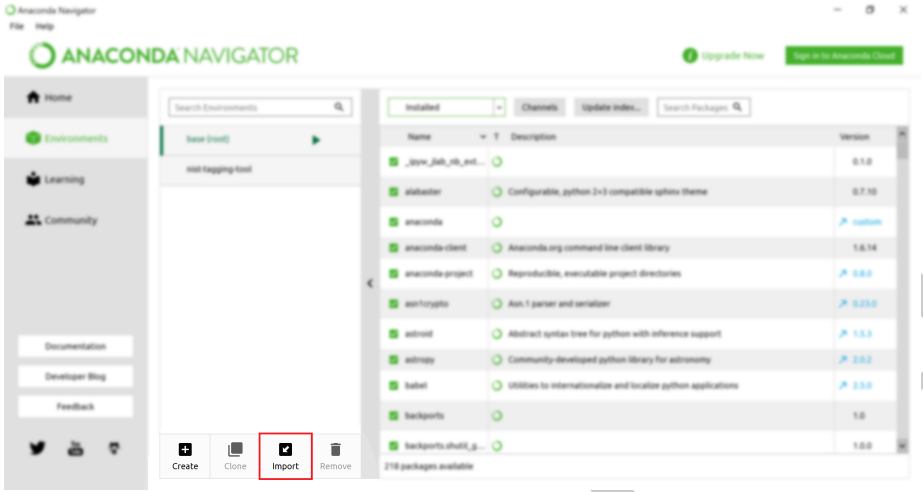
1. Unzip the application .zip file (please note the location of the file)

2. Open Anaconda

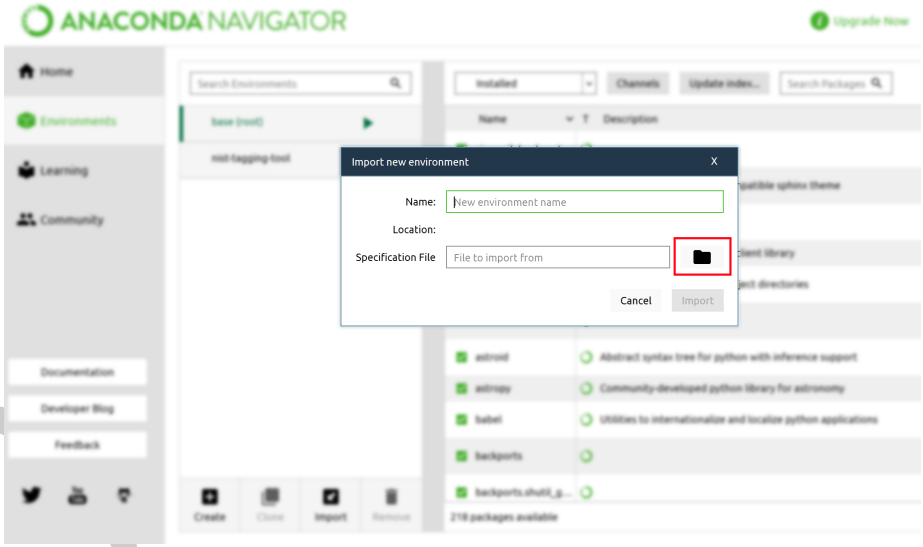


3. Click on the Environments tab

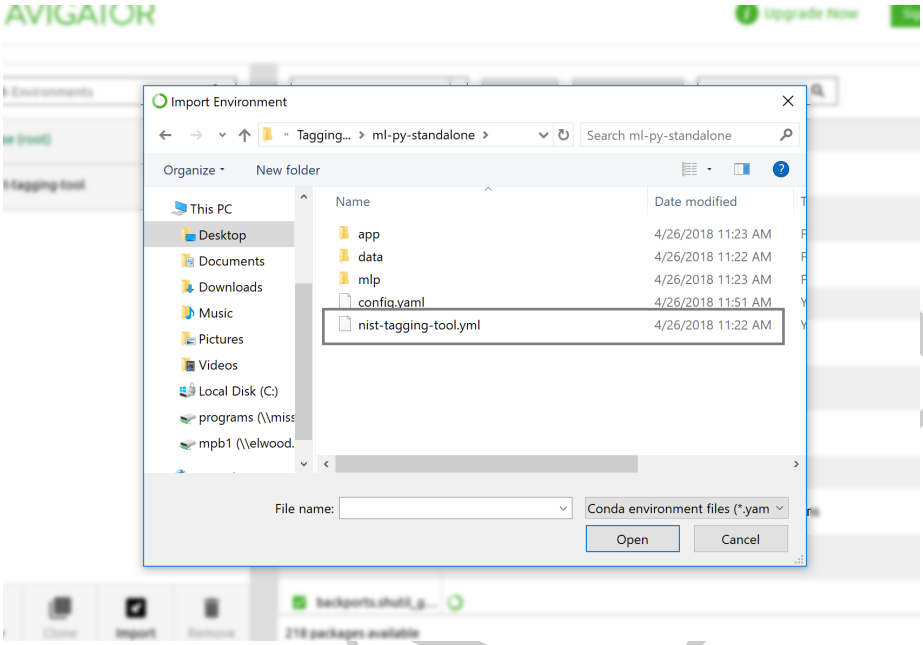
4. Click on the Import button



5. Click on the Browse button



6. Navigate to the folder that has the application stored and locate the **nist-tagging-tool.yml** and select open.



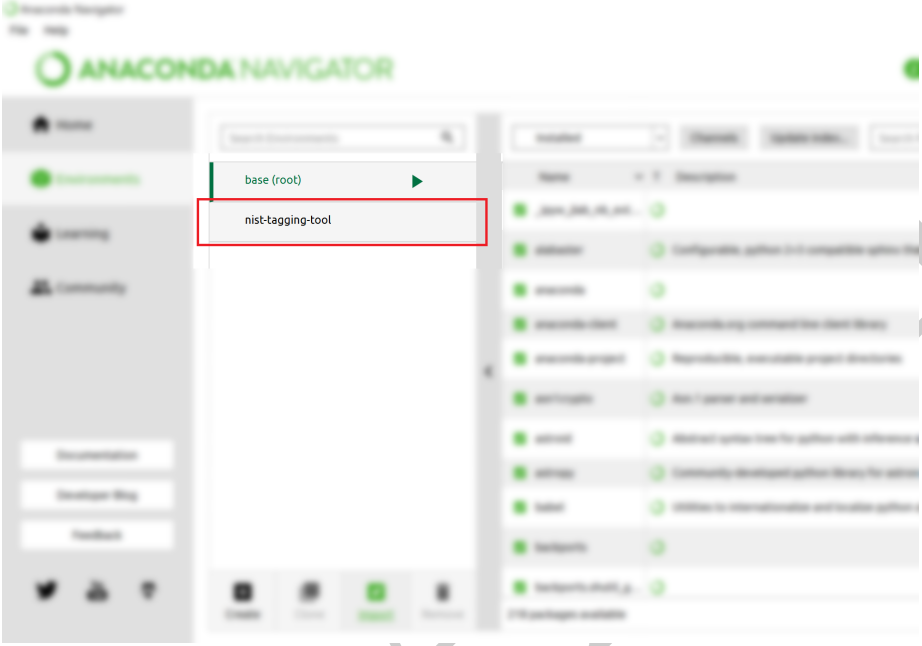
7. Wait for the environment to load (note this can take some time).

4 Using the Tagging Tool

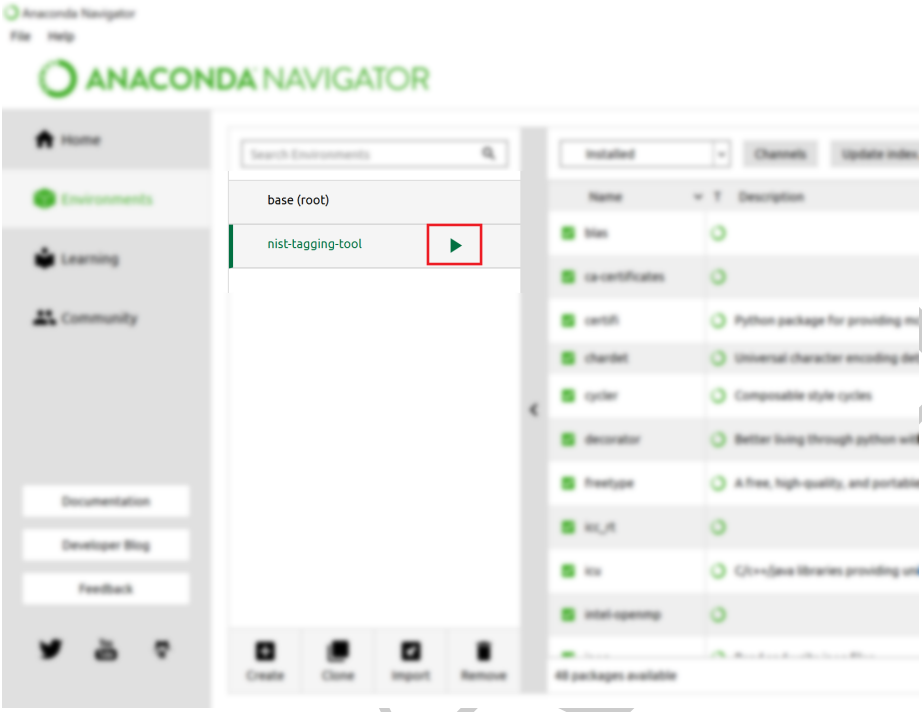
This section will walk through the steps for using the tagging tool application.

4.1 Start the Application

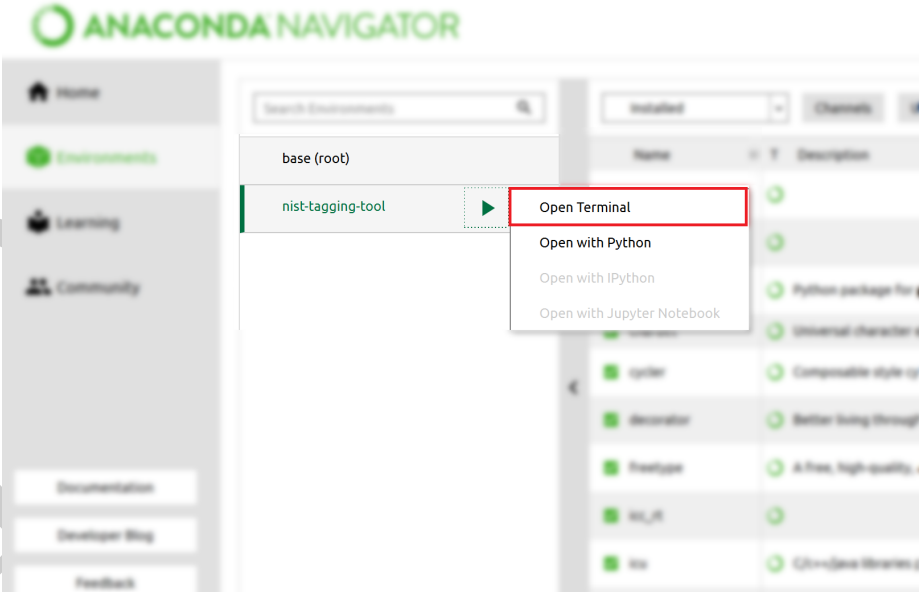
1. Click on the nist-tagging-tool environment button



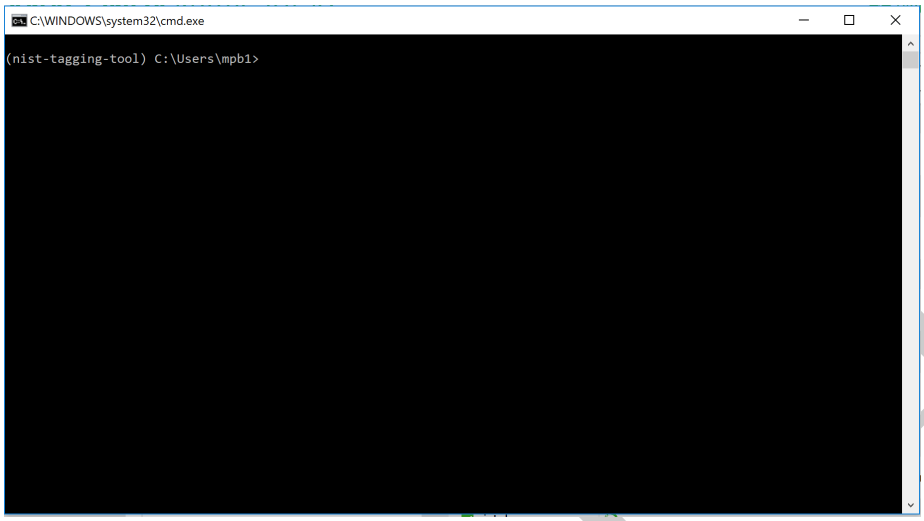
2. Click on the "run" button



3. Select "Open Terminal"



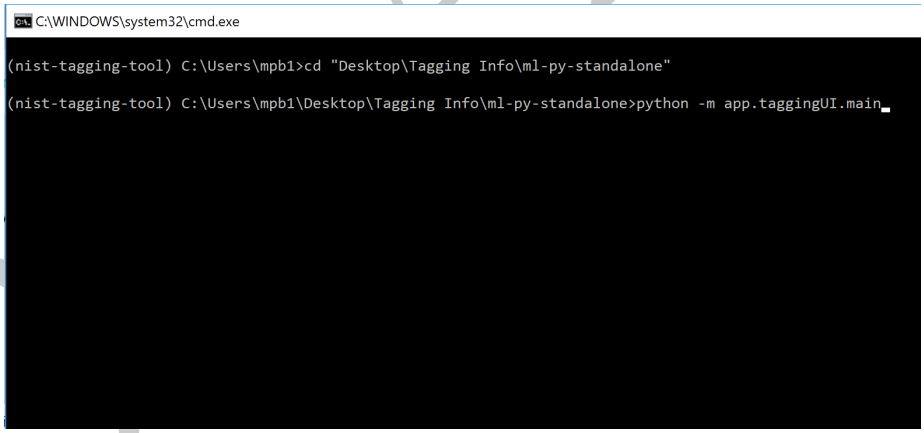
4. A new terminal will open that will look like the below:



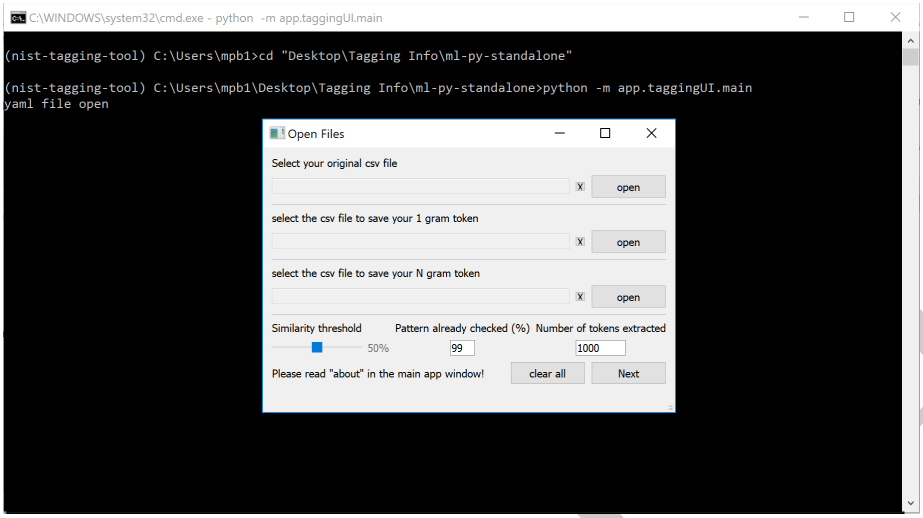
5. Within the terminal, navigate to the top level folder of the application. Here, the folder is located at `"Desktop/Tagging Info/ml-py-standalone"`

6. Type the command

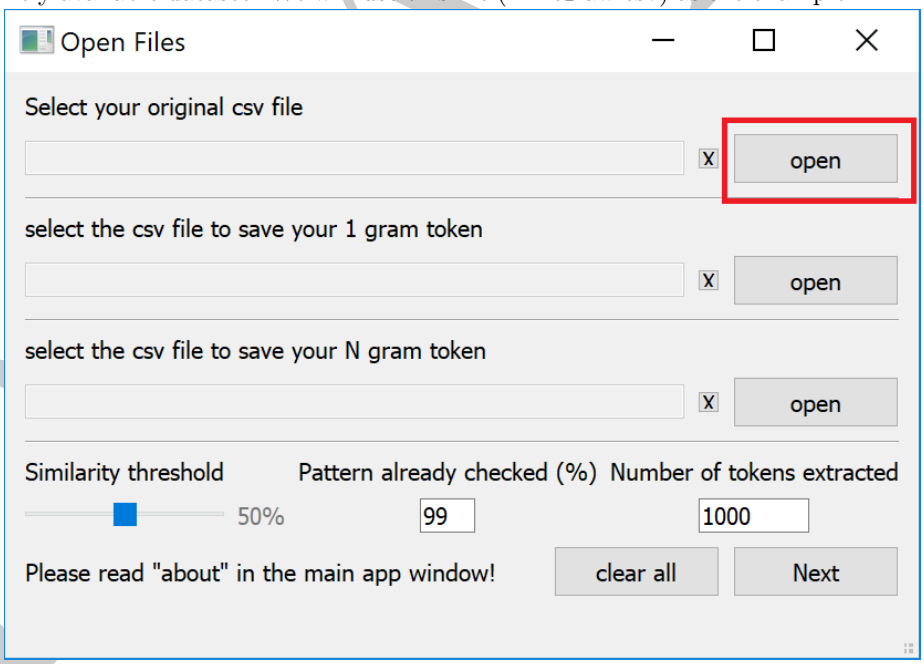
   ```
   python -m app.taggingUI.main
   ```

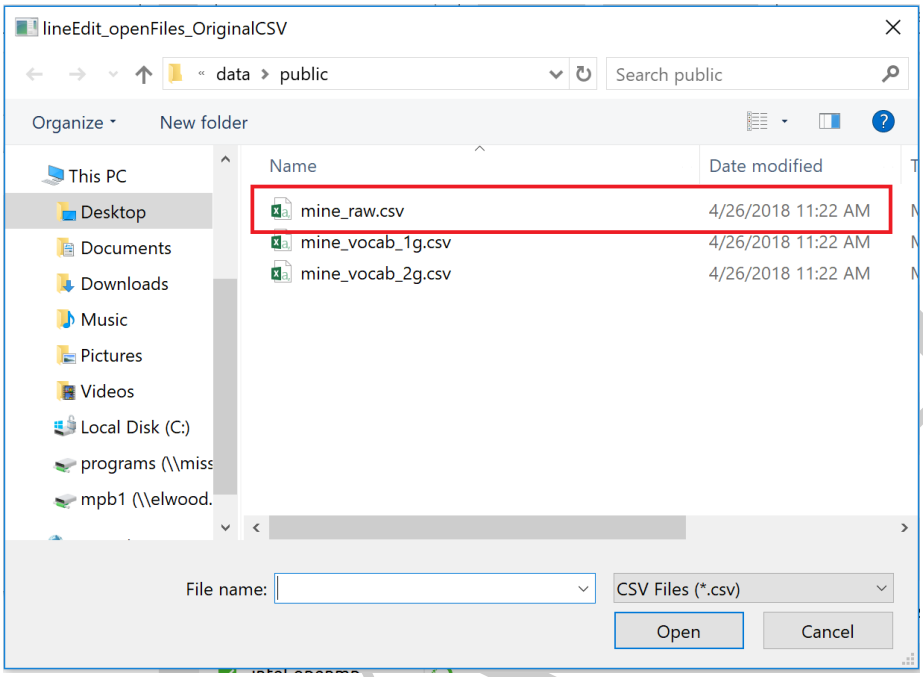


7. The application should open as seen below:

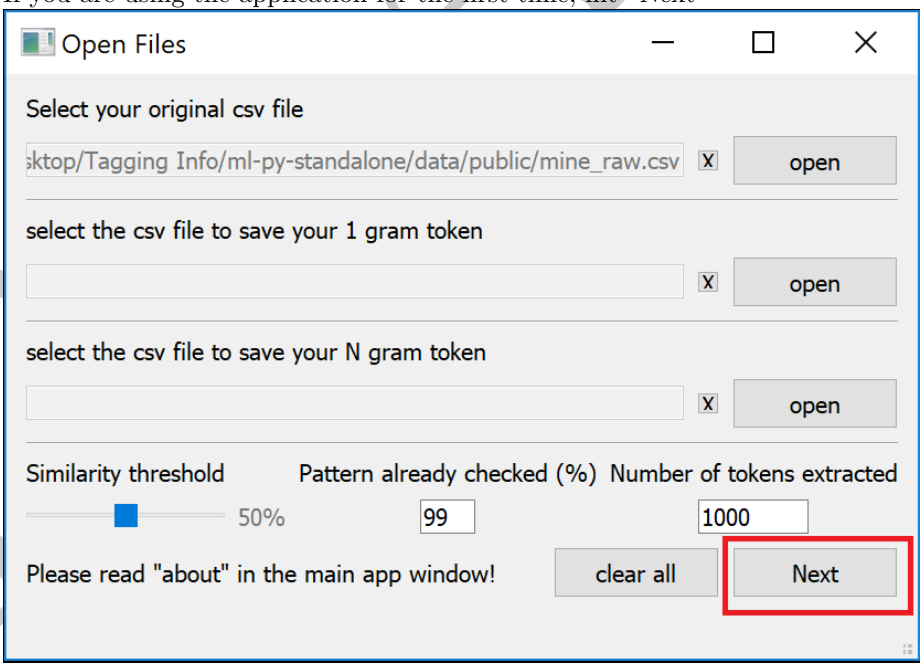


8. Open your .csv file with your MWOs. Included in the application, is a publicly available dataset. We will use this file (mine_raw.csv) as the example.

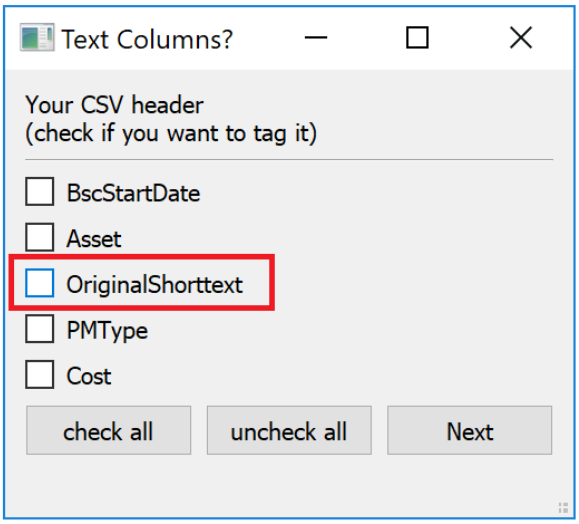




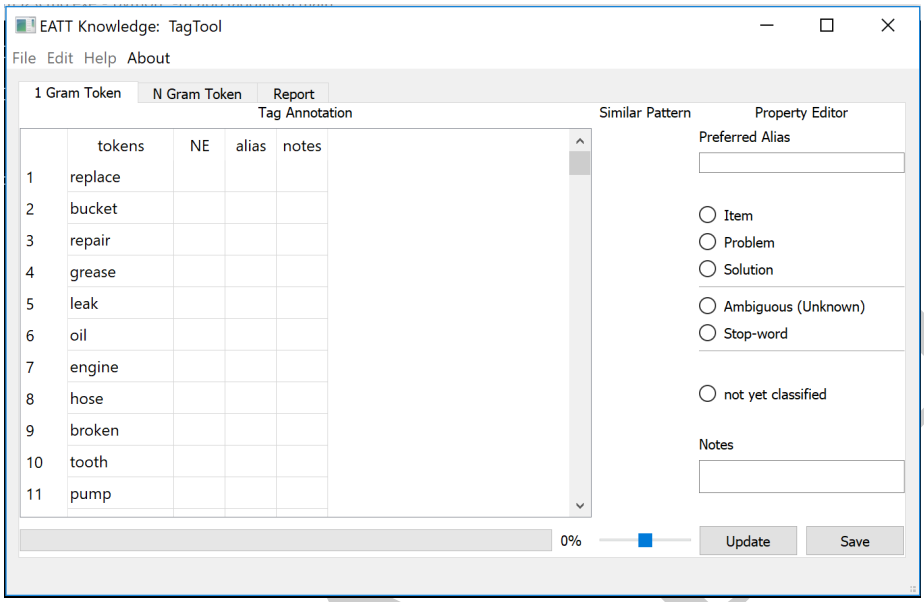
9. If you are using the application for the first time, hit "Next"



10. Select the column(s) that you would like to "tag." In this example, the column is "OriginalShorttext." Hit "Next".



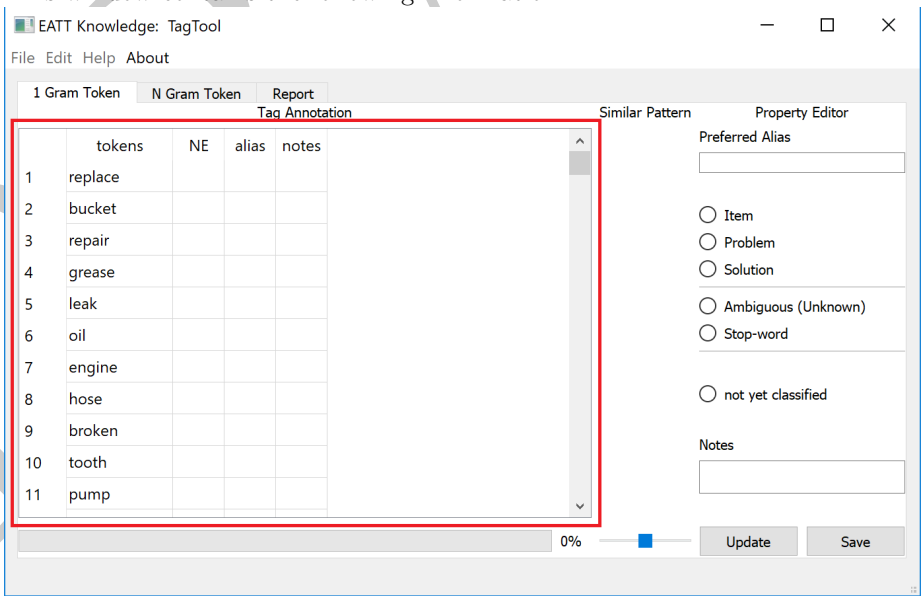
11. The application window will open as seen below:



4.2 Using the Application - 1 Gram Token tab

This subsection will describe the features of the application and goes into detail on the "1 Gram Token" tab.

- This window contains the following information:

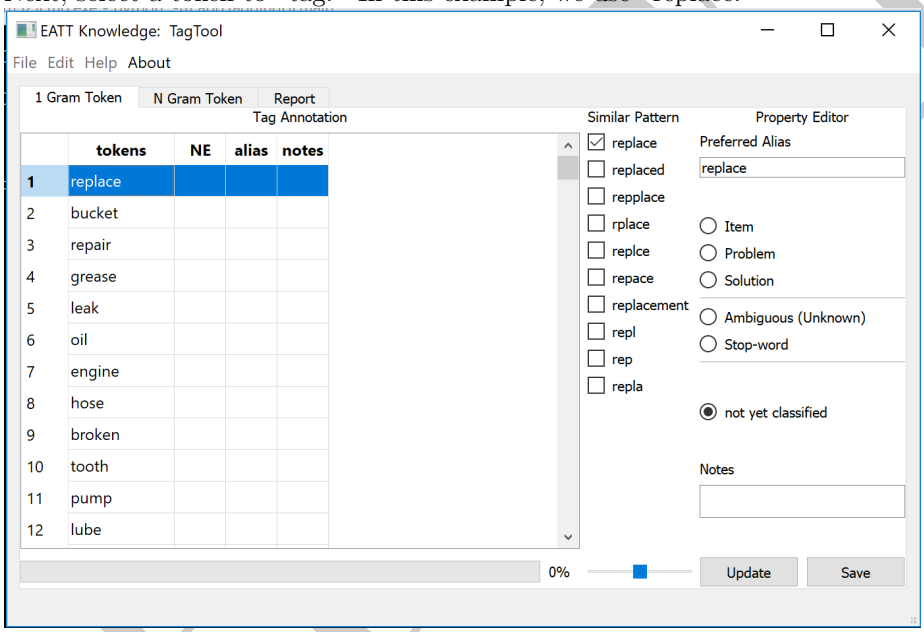


  - "tokens": The token as seen in the corpus and ranked by TF-IDF

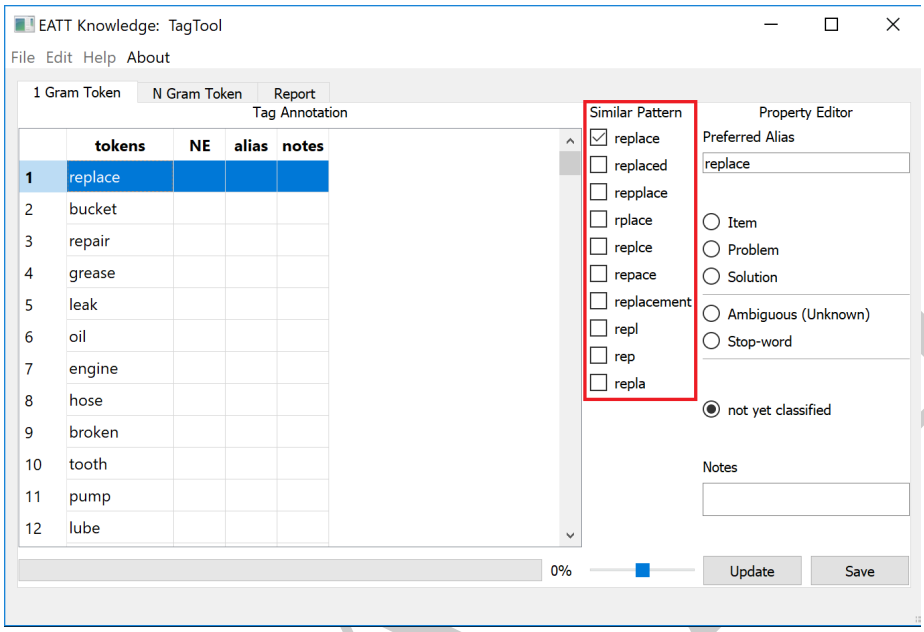
weighting.

  - "NE": This is a "Named Entity." This column will track the classifications of the tokens, which will be explained in more detail later.
  - "alias": This column tracks any aliases for tokens as made by the tool. These represent your new "tags."
  - "notes": This column tracks your notes for any tokens you have mapped to an alias.

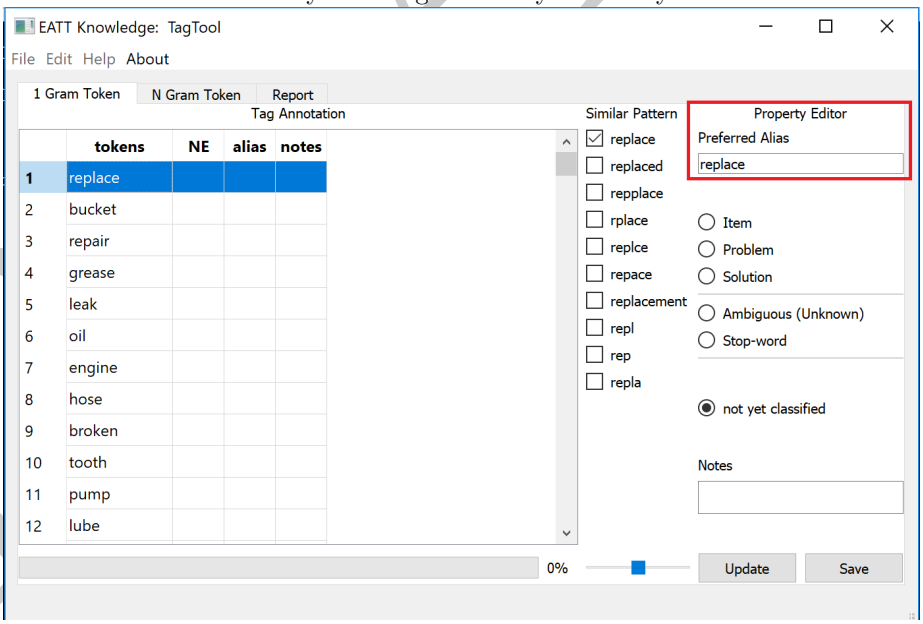
- Next, select a token to "tag." In this example, we use "replace."



- The "similar pattern" field will display words similar to the token using an "edit-distance"-based metric, via *fuzzywuzzy*. Any term that is selected here will be given the same alias and classification as the original token. So in this example, if "replaced" is selected, it will be given the same alias, notes, and classification as "replace"

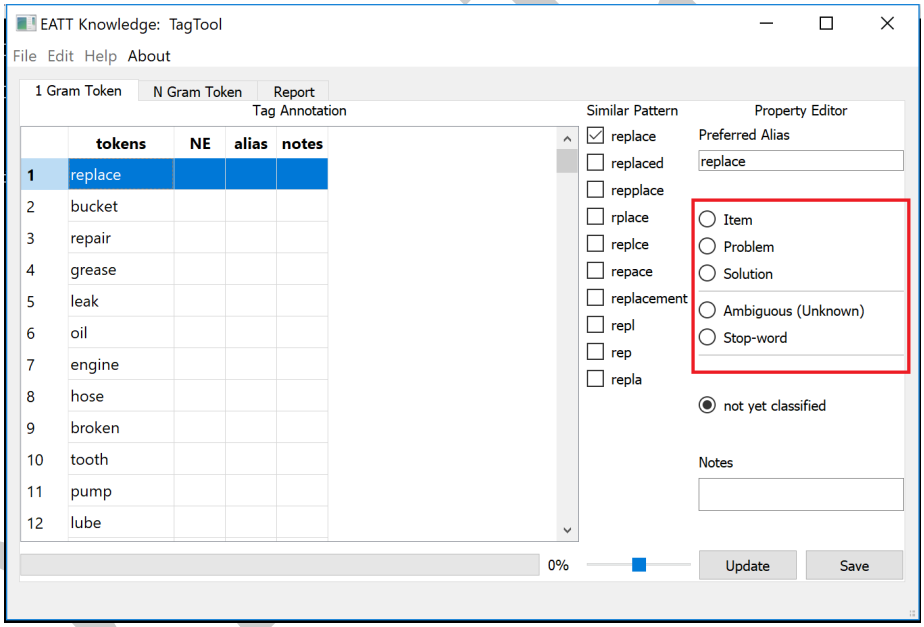


- The "alias" field will allow a user to enter any alias they would like for a token. The field will auto suggest the "token" as-is as the initial alias, but the user has the ability to change it to any alias they desire.

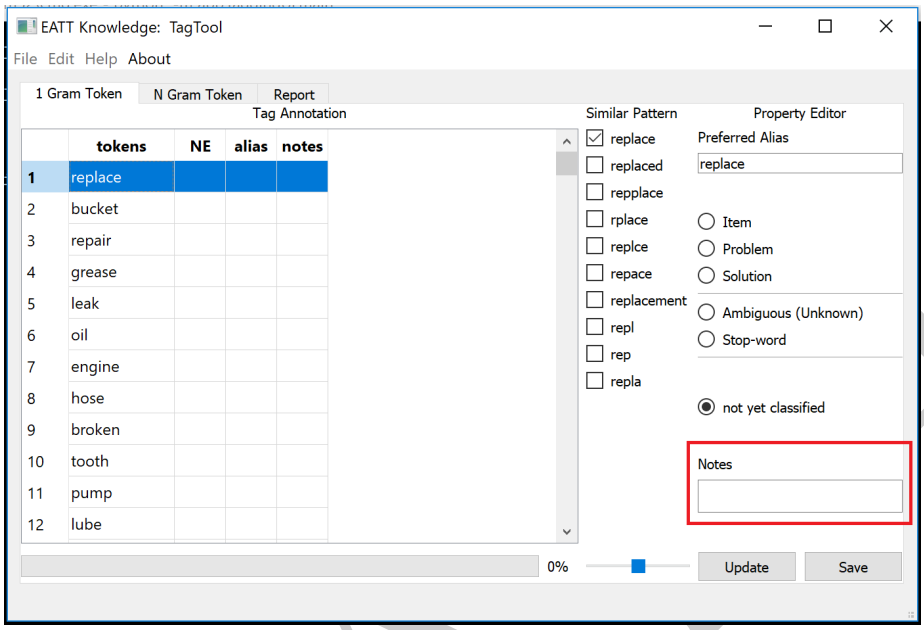


- This field is where the user can classify the "token." The classifications provided are:

- "Item": The objects directly relevant to the issue such as machine, resources, parts, etc. An example is a "pump" is always an item, however, "pumping" would not be an item.
- "Problem": The problem that is occurring at an item. An example is "leak" is always a problem.
- "Solution": The solution action taken on an item. An example is "replace" is always a solution.
- "Ambiguous (Unknown)": Words that are unknown without more context. An example is "oil" as this can be an item or a solution. This is further described in the N Gram Token tab section 4.3
- "Stop-word": A word that does not matter for analysis. For example, "see" or "according" are stop-words.



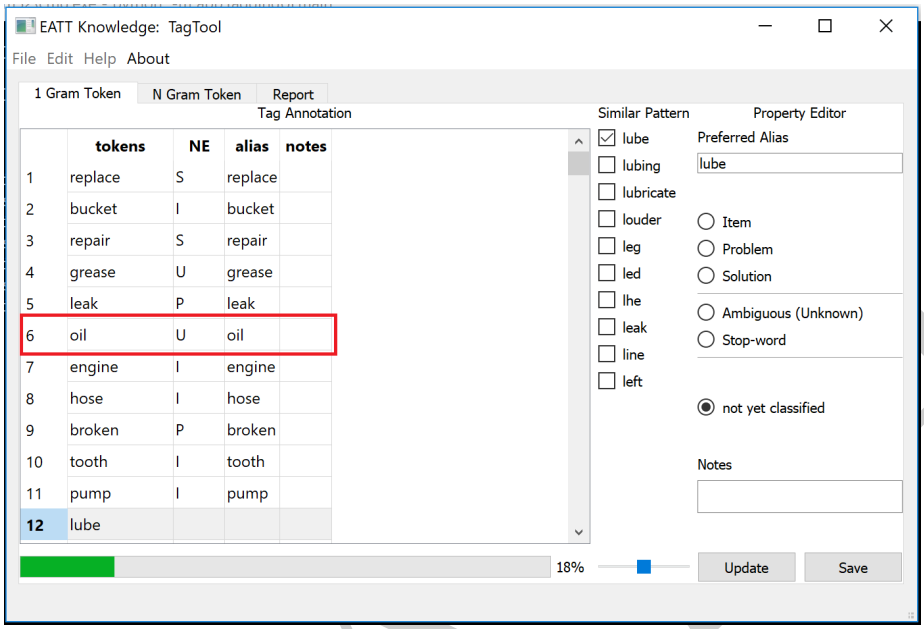
- The "Notes" field allows users to enter notes about the token/classifications.



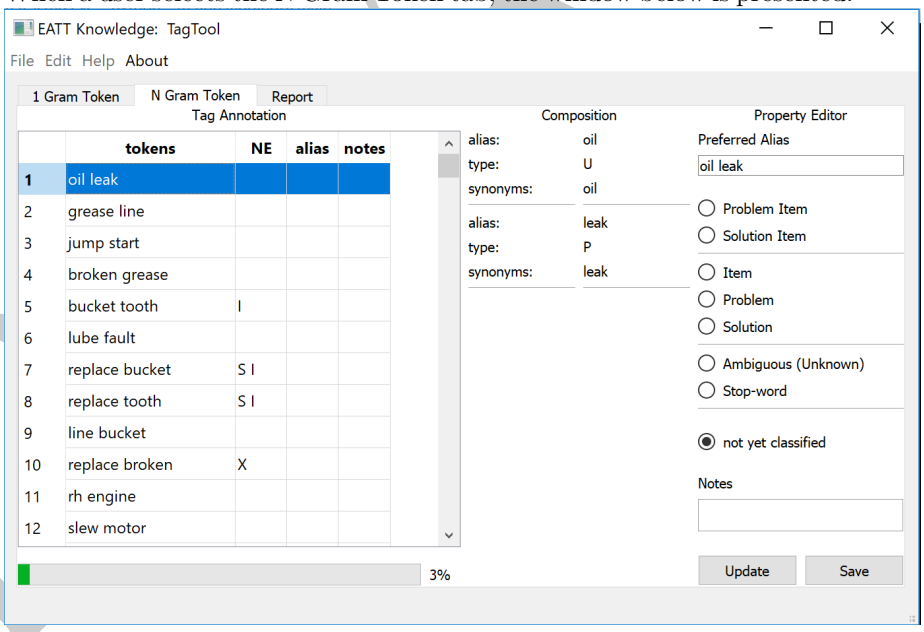
4.3 Using the Application - N Gram Token tab

This subsection will describe the features of the application and goes into detail on the "N Gram Token" tab.

- The N Gram token tab will provide detail on common 2 grams tokens, ordered in TF-IDF ranking, for the corpus (e.g., "hydraulic leak" is a common 2 gram in some data sets). The 2 grams can also provide more context for the "Uknown" classifications from the above section. For example, "oil" is unknown until the user is provided more context.

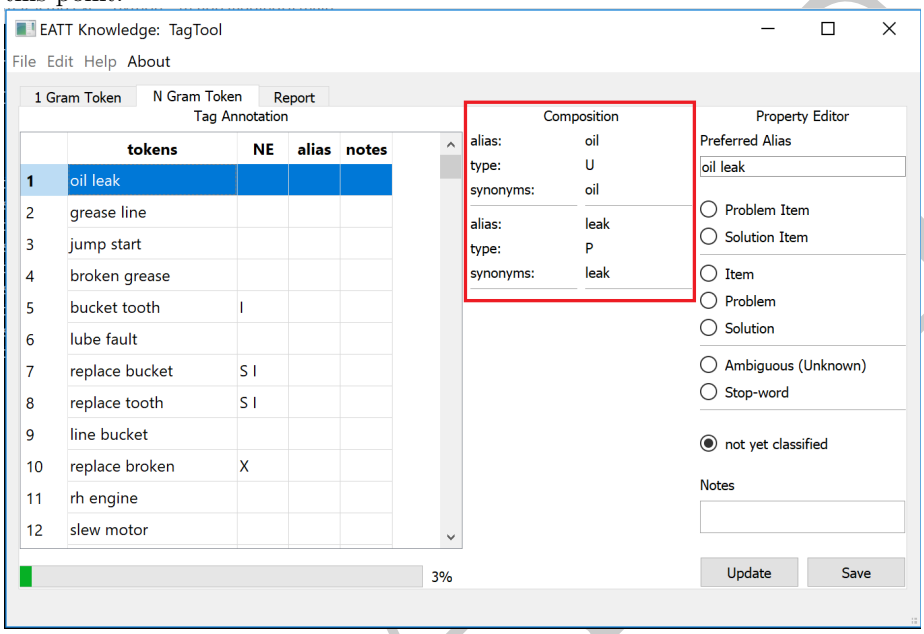


- When a user selects the N Gram Token tab, the window below is presented:

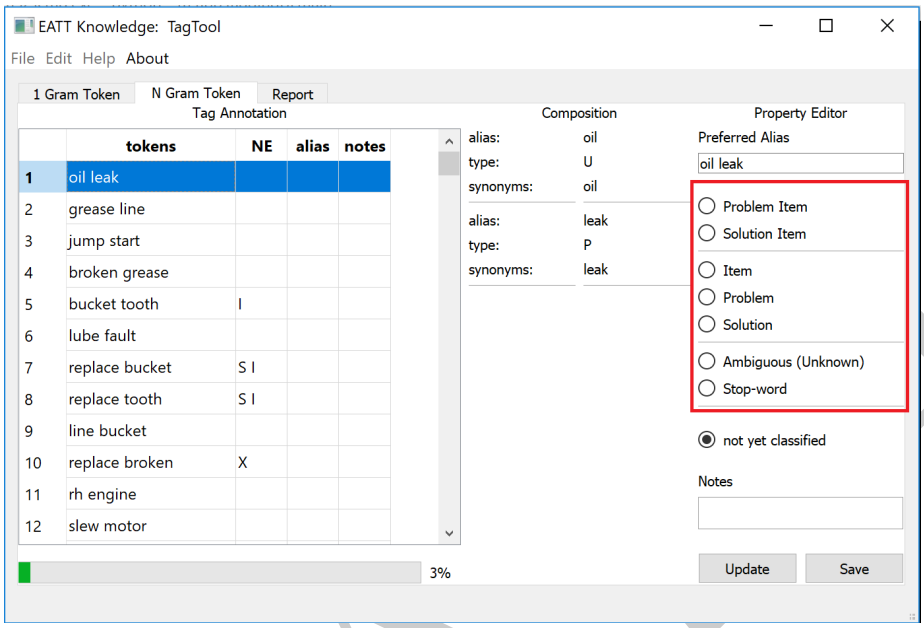


- The user is presented with the Composition of the 2 gram, which are composed of two 1 gram tokens. Each 1 gram is presented, with the classification ("type") and the synonyms (the other words that were linked with the Similar Pattern subwindow in the above section). In this exam-

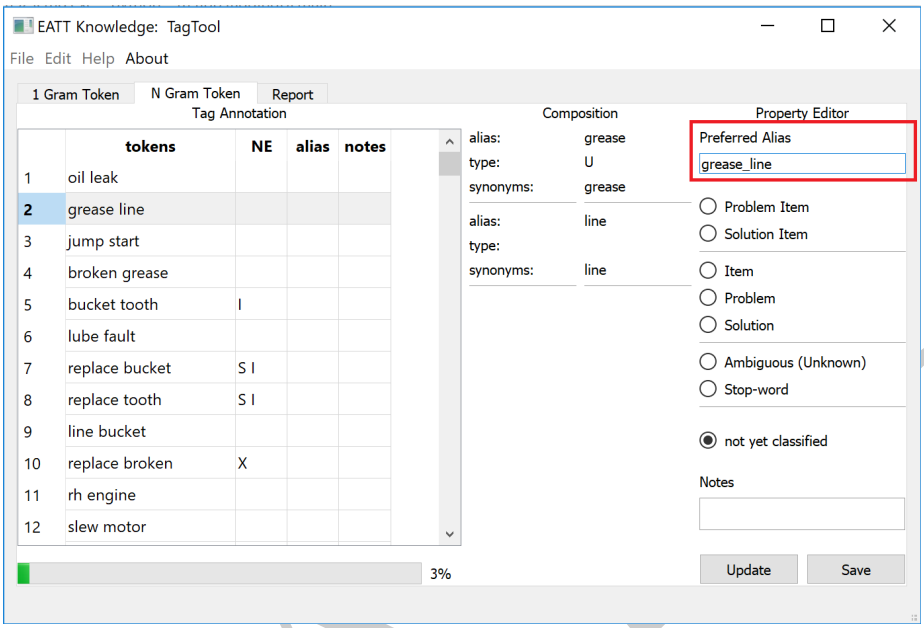
ple, “oil” is an “unknown (U)” classification and has no other synonyms at this point; “leak” is a “problem (P)” and has no other synonyms at this point.



- There are a number of classifications that a user can select for a 2 grams. The user will have to classify any 2 grams that contain an “U” classification. Please note that some 2 grams will be pre-classified based on a ruleset as seen below:



- Problem Item: This is a problem-item (or item-problem) pair. For example, "hydraulic" is an item and "leak" is a problem so "hydraulic leak" is a problem-item pair. The tool will pre-populate some problem-item pairs using the 1 grams that are classified as problems and items.
- Solution Item: This is a solution-item (or item-solution) pair. For example, "hydraulic" is an item and "replace" is a solution so "replace hydraulic" is a solution-item pair. The tool will pre-populate some solution-item pairs using the 1 grams that are classified as solutions and items.
- Item: This is for pairs of items that are de facto 1-grams. For example "grease" is an item, line is an "item", but a "grease_line" is most likely its own "item". The tool will pre-populate some items based on 1 grams that are both items. Please note that 2 gram items, since they are really being treated as 1-grams, must have an underscore (_) in their alias, between the 2 individual items as seen below:

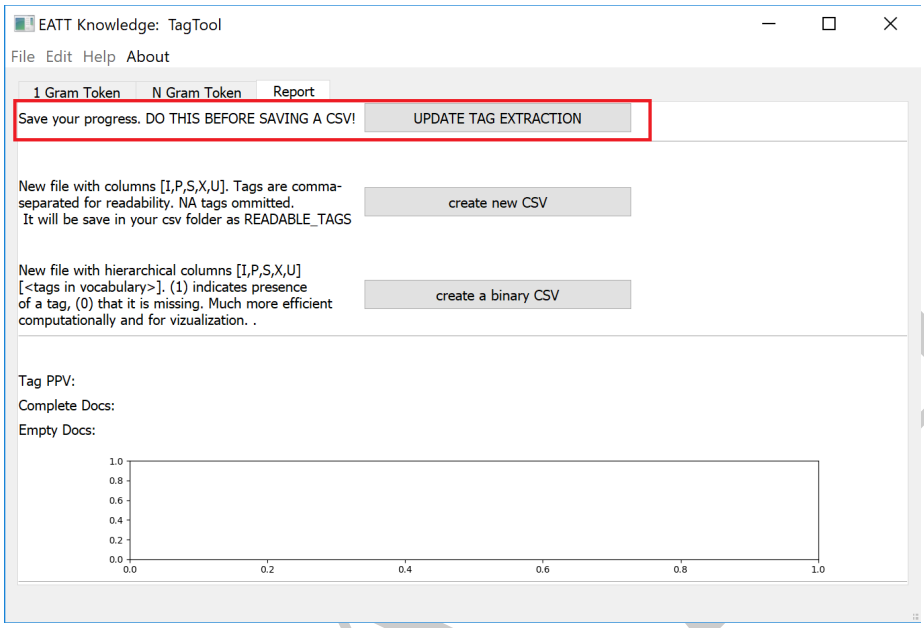


- Problem: This is a problem that is a 2 gram. This will be left up to the user to classify as these will not be pre-populated using 1 gram classifications. Please note that 2 gram problems, since they are really being treated as 1-grams, must have an underscore (_) in their alias, between the 2 individual problems.
- Solution: This is a solution that is a 2 gram. This will be left up to the user to classify as these will not be pre-populated using 1 gram classifications. Please note that 2 gram solutions, since they are really being treated as 1-grams, must have an underscore (_) in their alias, between the 2 individual solutions.
- Ambigious (Unknown): This is an unknown 2 gram that needs more context. This will be left up to the user to classify as these will not be pre-populated using 1 gram classifications.
- Stop-word: This is 2 gram stop-word. This will be pre-populated when a "solution" 1 gram is paired with a "problem" ' gram. The user can decide if any other 2 grams are not useful.

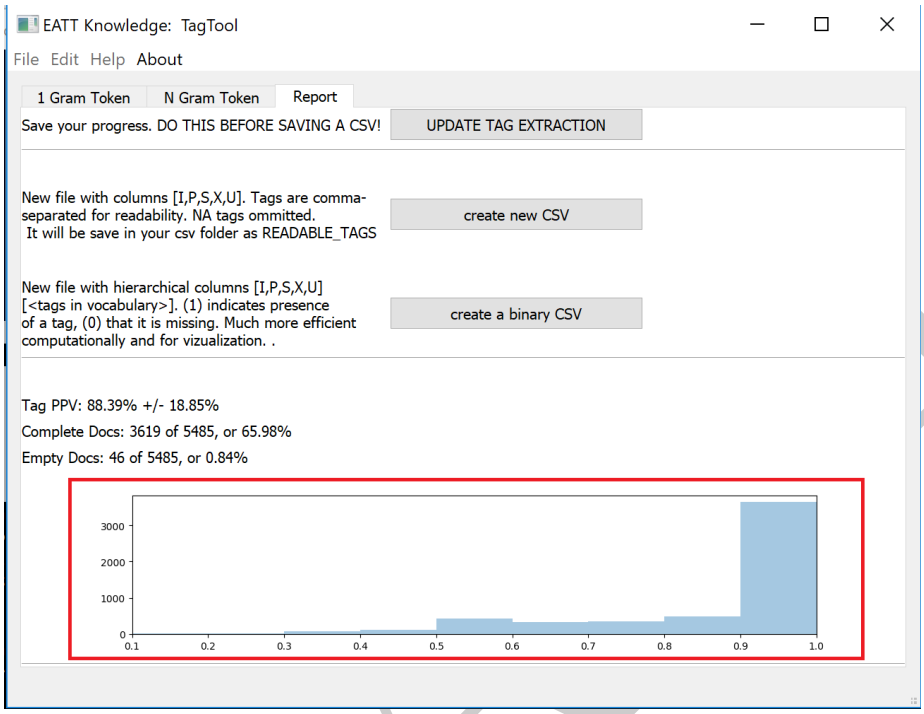
4.4 Using the Application - Report tab

Once the user is done tagging their desired amount of tokens, they can begin using the report tab.

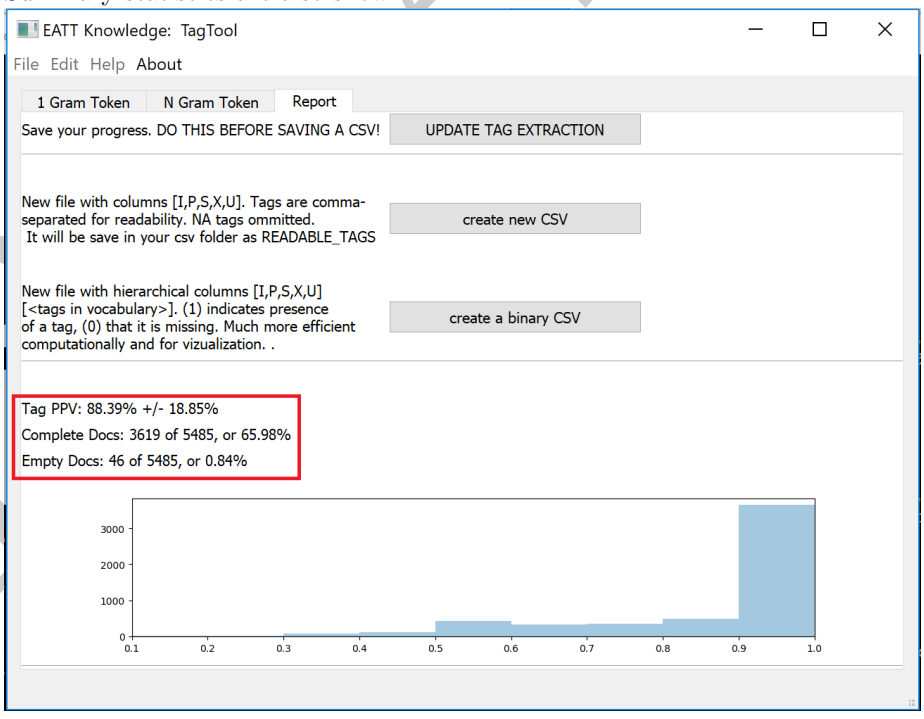
- Please make sure to hit the "UPDATE TAG EXTRACTION" button before proceeding. This may take some time to compute.



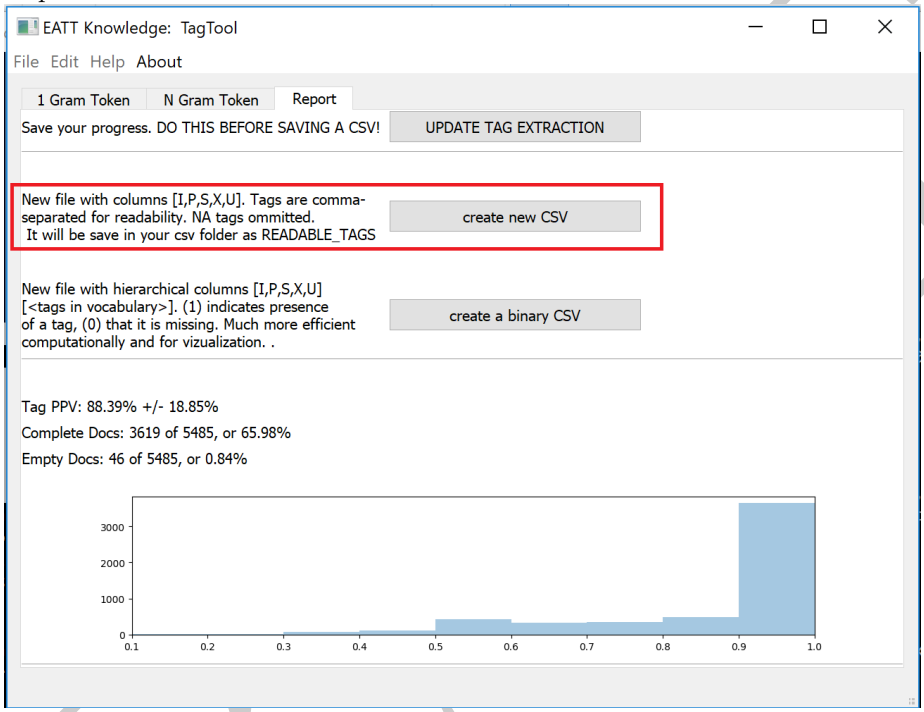
- The bottom graph will update. It explains the amount of tagging that has been completed. The distribution of documents (shown as a histogram) is calculated over the precision for each document (i.e. of the tokens found in a document, what fraction have a valid classification defined).



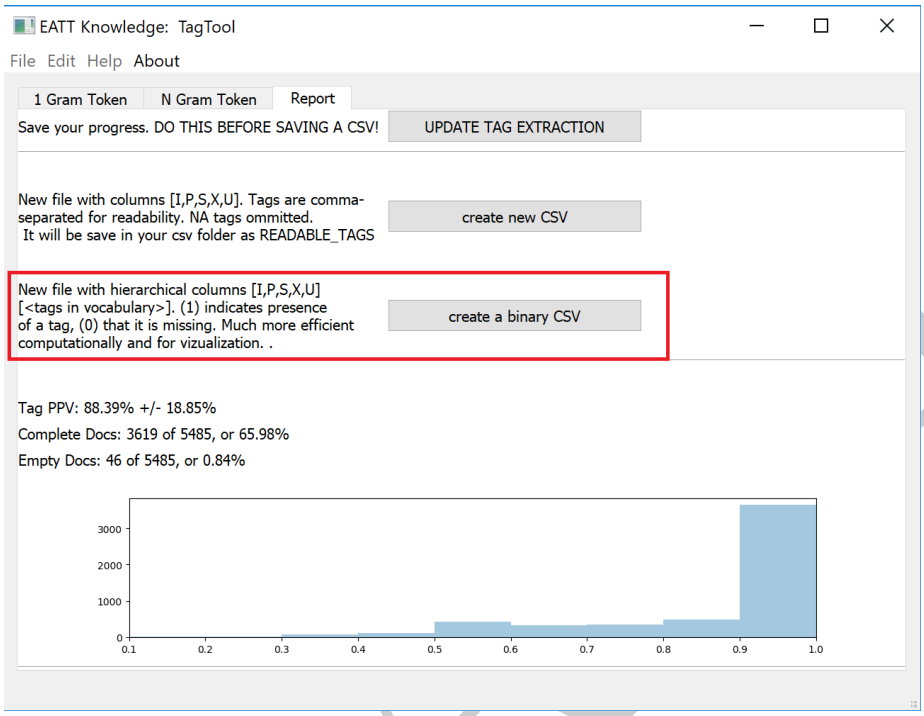
- Summary statistics are also shown.



- The “create new CSV” button will create an .csv with the original dataset and 7 new columns (“I”,“P”,”PI”, “S”,“SI”,“U”, and “X”) , which contain the new tags from each category. Please note that “X” contains any stop words.



- The “create a binary CSV” button will create 2 new .csv files. Each file will contain the work order number (starting with 0), and is ordered identically to the .csv file that was originally loaded. Two new files are created: BINARY_TAGS and BINARY_RELATIONS.



– BINARY_TAGS: The left most column contains the work order number, while the headers contain all 1 gram tags. A "0" is placed when the work order does not contain the tag in the header and a "1" is placed when the tag in the header is contained in the work order.

– BINARY_RELATIONS: The left most column contains the work order number, while the headers contain Problem-Item and Solution-Item tag combinations. A "0" is placed when the work order does not contain the tag in the header and a "1" is placed when the tag in the header is contained in the work order.

References

[1] T. Sexton, M. P. Brundage, M. Hoffman, and K. C. Morris, "Hybrid datafication of maintenance logs from ai-assisted human tags," in *Big Data (Big Data), 2017 IEEE International Conference on.* IEEE, 2017, pp. 1769–1777.

[2] M. Sharp, T. Sexton, and M. Brundage, "Toward semi-autonomous information extraction for unstructured maintenance data in root cause analysis."