Annotation via Dictionary

Dictionary file: What is it?

A dictionary file is perhaps just a fancy name for any csv file that contains a list of terms to be used as the terms to be annotated in an input txt file. So… this is a one-column csv dictionary file

	A
1	Word
2	abort
3	act
4	add
5	ally
6	alternate
7	apply
8	appreciate
9	approach
10	arrange
11	ask
12	astonish

When used as an annotation dictionary, if the input txt file contains any of the words in column 1 (the the header Word), they will be annotated.

And this is another csv dictionary file, this time a two-columns file.

	A	B
1	Word	WordNet Category
2	abort	change
3	act	social
4	add	change
5	ally	social
6	alternate	change
7	apply	consumption
8	appreciate	emotion
9	approach	motion
10	arrange	contact
11	ask	communication
12	astonish	cognition

In this case, you could use the values in the second column (header WordNet Category) to annotate terms in the input txt file for any term appearing in column 1 (header Word).

Dictionary file: Where do I get one?

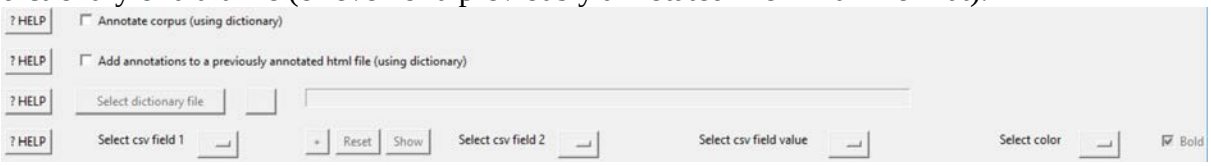
There are several ways of building a dictionary file.

1. The **CoNLL table** is a good start. You may extract from a CoNLL table all tokens/words that have NER values PERSON or LOCATION, or a POSTAG value NN, or a DEPREL value nsubj depending upon your research questions.
   If you use one of the fields of the CoNLL table, if you decide to use lemma values, you would miss a great deal of annotations; for instance, be, the lemma value for is, was, were, being, besides be, would miss out all those other values.
2. You could use the **WordNet** GOING UP option in the NLP Suite WordNet.py to obtain the two-column dictionary file displayed above.
3. You could use **WordNet** GOING DOWN to obtain, for instance, a list of countries, or ethnic/racial groups and then use the list as a dictionary.
4. If you are annotating a book that contains **book index entries**, you could extract the index as a dictionary to be then used to annotate the book.
5. As an extensiion of the index approach, suppose that you investigating New York Times best seller book reviews. You are interested to see whether reviewers use any terms borrowed from literary criticsm. You can use an actual **dictionary of literary terms** (using its index) to construct a dictionary to be used to annotate the reviews. The same would be true if you are investigating newspaper reports on health for which you could use a **dictionary of health and medicine** for ailments and symptoms, or a **list of prescription drugs**.
6. For a restricted domain (e.g., you are investigating a collection of gay people blogs), you may start from a frequency list of terms used (via **n-grams** algorithm for unigrams including or excluding stopwords) and manipulate the output list for youur purposes.

The sky is the limit… Use your creativity!

How does it work? A look at the GUI

The annotator GUI provides a set of widgets designed to make easy the annotation via dictionary of a txt file (or even of a previously annotated file in html format).



Using the GUI widgets you can select the csv file to use as annotation dictionary, then select the combination of columns and colors to be used for annotating (e.g., the values in column 4 of a multi-column csv file).

1. This way, you can choose to annotate the values in the *Select csv field 1* (let's say Words of love) in red, then in another Select csv field 1 (let's say Words of work) in blue, then in yet another Select csv field 1 (let's say Words of leisure) in yellow, …
2. You can choose to annotate the values in *Select csv field 1* by selecting specific values in that field using the *Select csv field value* dropdown menu and associating those selected values to a specific color. You can do that repeatedly, annotating different sets of words in different colors.
3. You can also annotate the values in *Select csv field 1* by selecting specific values (*Select csv field value*) in *Select csv field 2* rather than in *Select csv field 1*. You can associate the selected values to selected colors. When using a second csv column, the actual annotation is carried out using the terms in *Select csv field 1*.

Dictionary annotation is both case sensitive and exact (the dictionary entry 'remember' would miss 'remembered' in the text).

Multiple annotations? Another look at the GUI

You can use the GUI option "Add annotations to a previously annotated html file (using dictionary)" to annotate a previously annotated html file, whether annotated via DBpedia or via dictionary. Using different dictionaries repeatedly using different colors each time on the same annotated file, you can achieve the same result as using the *Select csv field 1* and *Select csv field 2* widgets described above. The advantage of this option is that you can annotate a DBpedia annotated html file a second time around (or third, or fourth…) using dictionary values.

What you cannot do…

1. You cannot use the *Select csv field 1* and *Select csv field 2* widgets on an html file.
2. At the same token, you cannot annotate a previously annotated html file, whether via DBpedia or dictionary, via DBpedia. You will need to do it the other way around, starting from DBpedia.

Output filenames

You can recognize a file that has been dictionary annotated multiple times by specific tags:

NLP_DBpedia_annotated_dict_annotated_Faulkner_Dry September.html
NLP_DBpedia_annotated_multiDict_annotated_Faulkner_Dry September.html
NLP_multiDict_annotated_Faulkner_Dry September.html

In the first case, the filename tells me that the html DBpedia annotated Faulkner's Dry September has been re-annotated via a dictionary.

In the second case, that same html file has been re-annotated multiple times via a dictionary.

In the third case, Faulkner's *Dry September* originally annotated via a dictionary has then been re-annotated multiple times via a dictionary.

Further readings: More TIPS files

TIPS_NLP_Annotator.pdf
TIPS_NLP_Annotator DBpedia.pdf
TIPS_NLP_Annotator extractor.pdf