Word Sense Induction using BERT

Contents

Same Word, Different Senses

A word can mean different things depending on the situation in which it is used. For example the word 'wooden' might be used to talk about the fact that some things are made of wood ('The chair is wooden'). However in another situation 'wooden' might be used to criticise someone's acting ('That actor's performance was very wooden'). So, what 'wooden' means depends on the situation in which it's used.

What is Word Sense Induction?

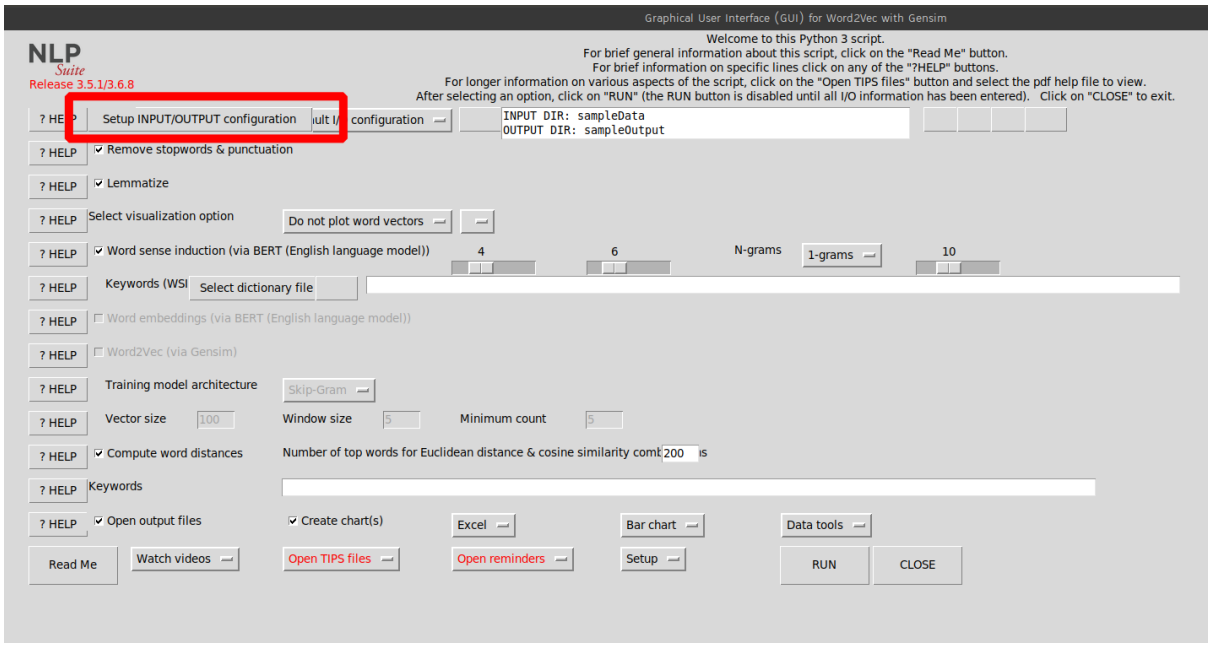
Word sense induction (WSI) is the problem of automatically identifying the different senses expressed by a word used in a collection of documents. For example, if a document contains 10 occurrences of the word 'wooden', an effective WSI procedure would be able to automatically determine which of these 10 occurrences express the *made of wood* sense of 'wooden', and which of the 10 occurrences express the *wooden acting* sense of 'wooden'.

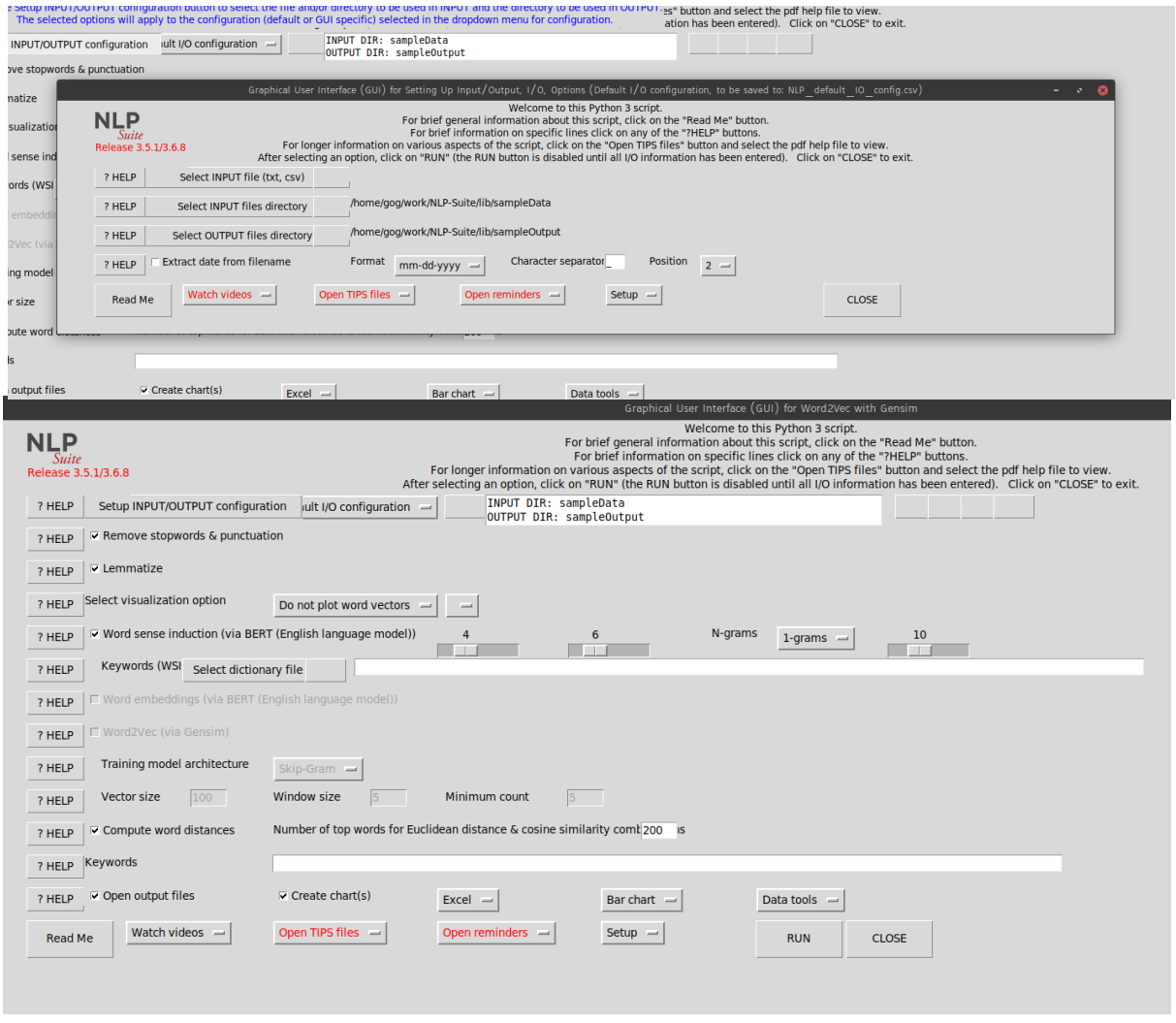
How does Word Sense Induction in NLP Suite Work?

The WSI procedure used in NLP Suite is based on the procedure developed by Lucy Li and David Bamman [1], which makes use of BERT embeddings and k-means clustering. It has three steps. Suppose we want to perform WSI to extract the different senses of 'wooden' expressed by one document. First, we input the document into BERT to retrieve embeddings for every occurrence (token) of 'wooden' contained in the document. Then, we cluster a sample of the retrieved embeddings using k-means and extract the centroids of each cluster. Each centroid is taken to represent a sense. Finally, we match each token of 'wooden' to the centroid closest to its embedding by cosine similarity. Each token matched to a particular centroid is taken to express the sense represented by the centroid.

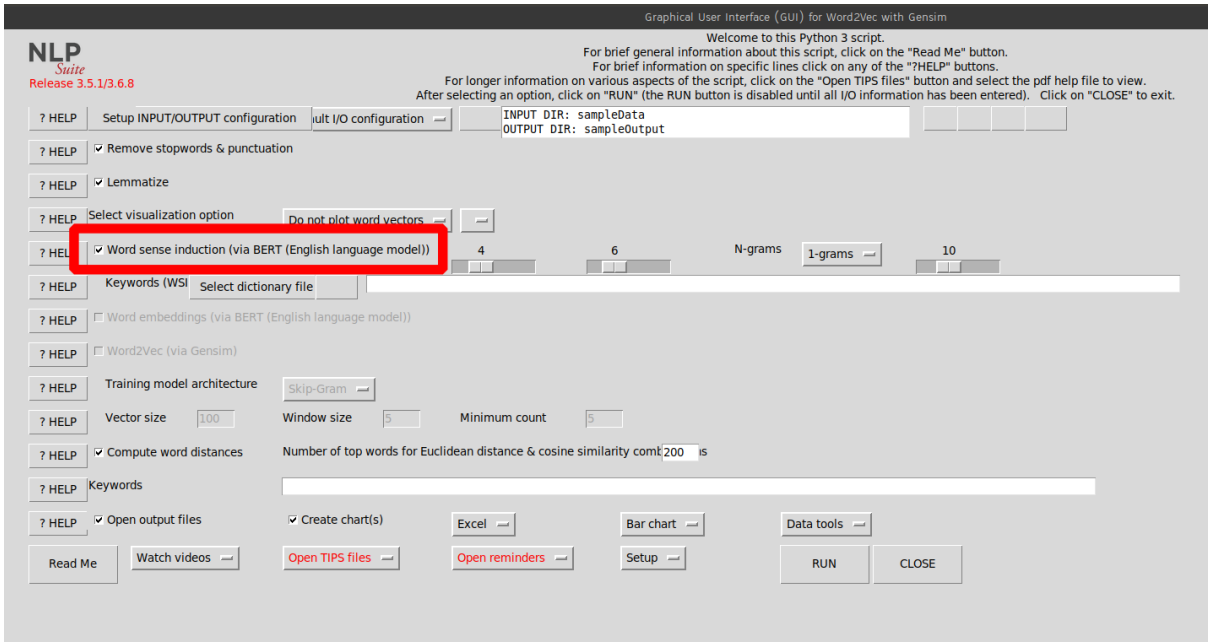
How do I use Word Sense Induction in NLP Suite?

1. Run the menu for Word2Vec with Gensim.
2. Select the documents you wish to run WSI on and what directory you want the output of the procedure to be stored within using ‘Setup INPUT/OUTPUT configuration’.

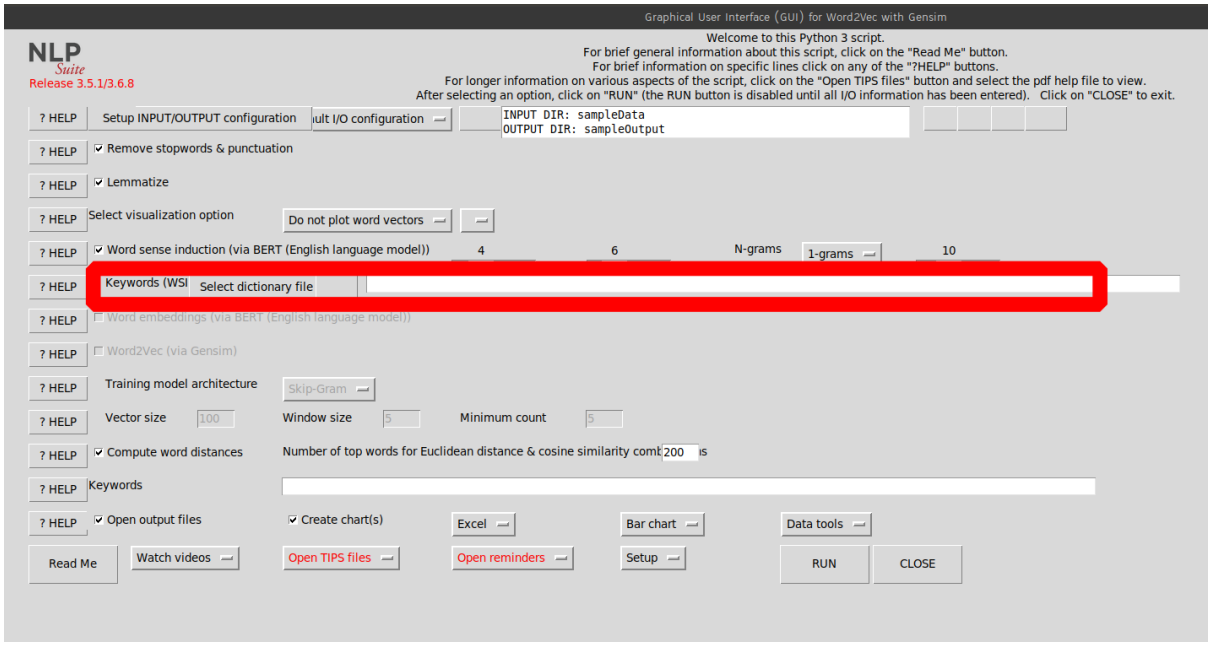




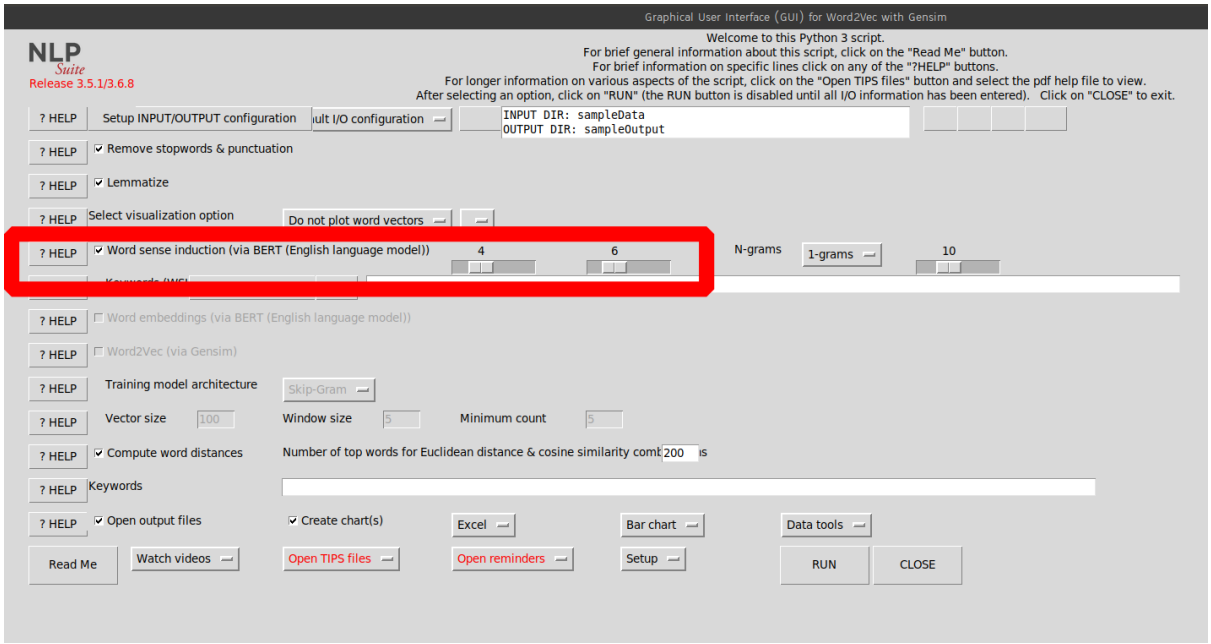
3. Click on the ‘Word sense induction (via BERT(English language model))’ option.



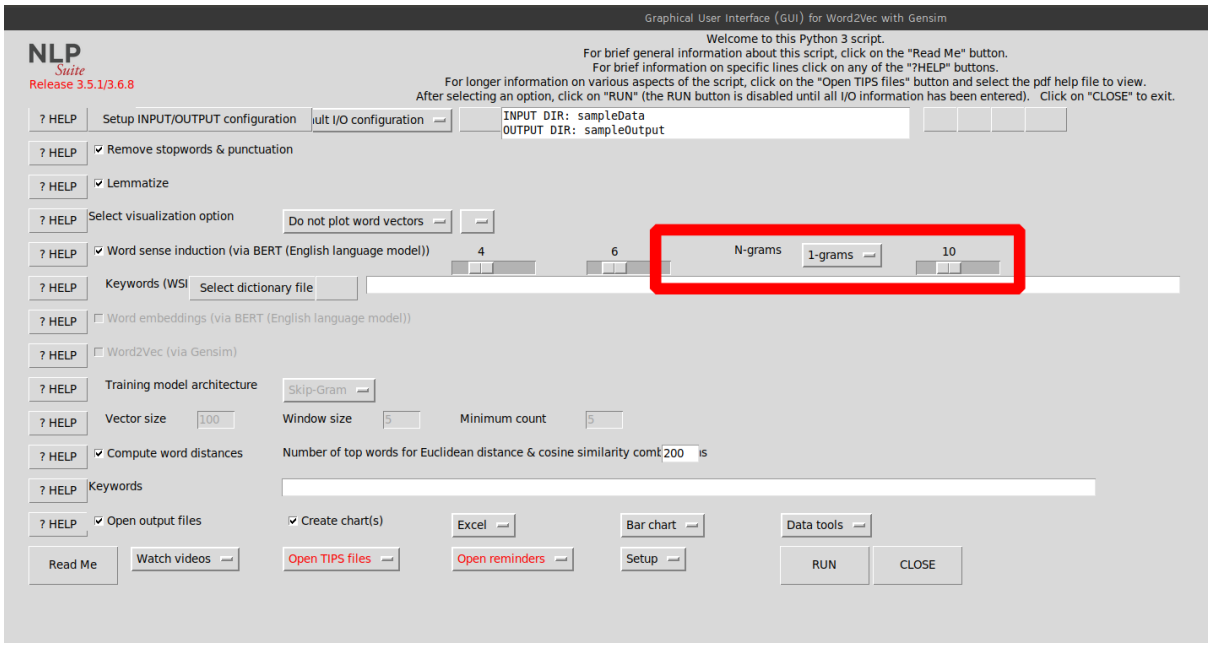
4. You have the option of specifying the words you want to extract senses for using the Keywords (WSI) option. You can either enter a comma separated list of words, or upload a file containing the list of words you want to use.



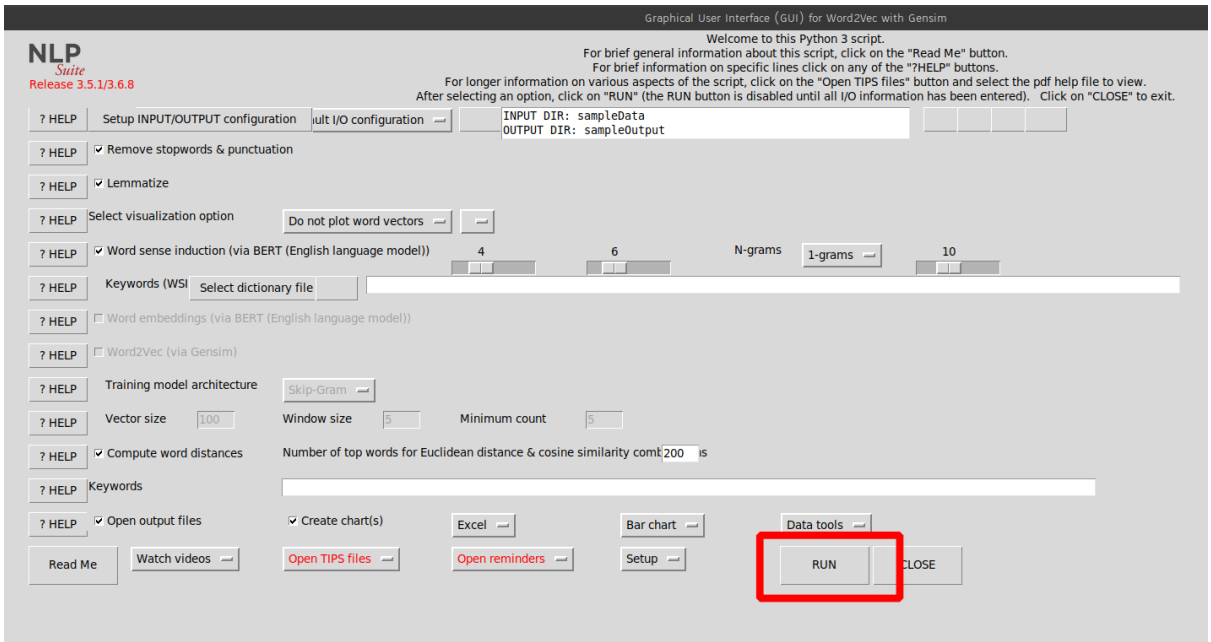
5. Set the lower bound and upper bound of k-means you wish to use. The procedure will run k-means on all values of *k* (where *k* represents the number of senses to be generated) between these bounds and then select the *k* that minimises the residual sum of squares of clusters produced through k-means.



6. To provide a summary of the content of the clusters generated, the most distinctive n-grams (measured using TF-IDF) of each cluster can be retrieved. You have the option of selecting what kind of n-gram (unigram, bigram, trigram etc.) you want to retrieve. You can also decide how many n-grams should be listed for each sense using the slider.

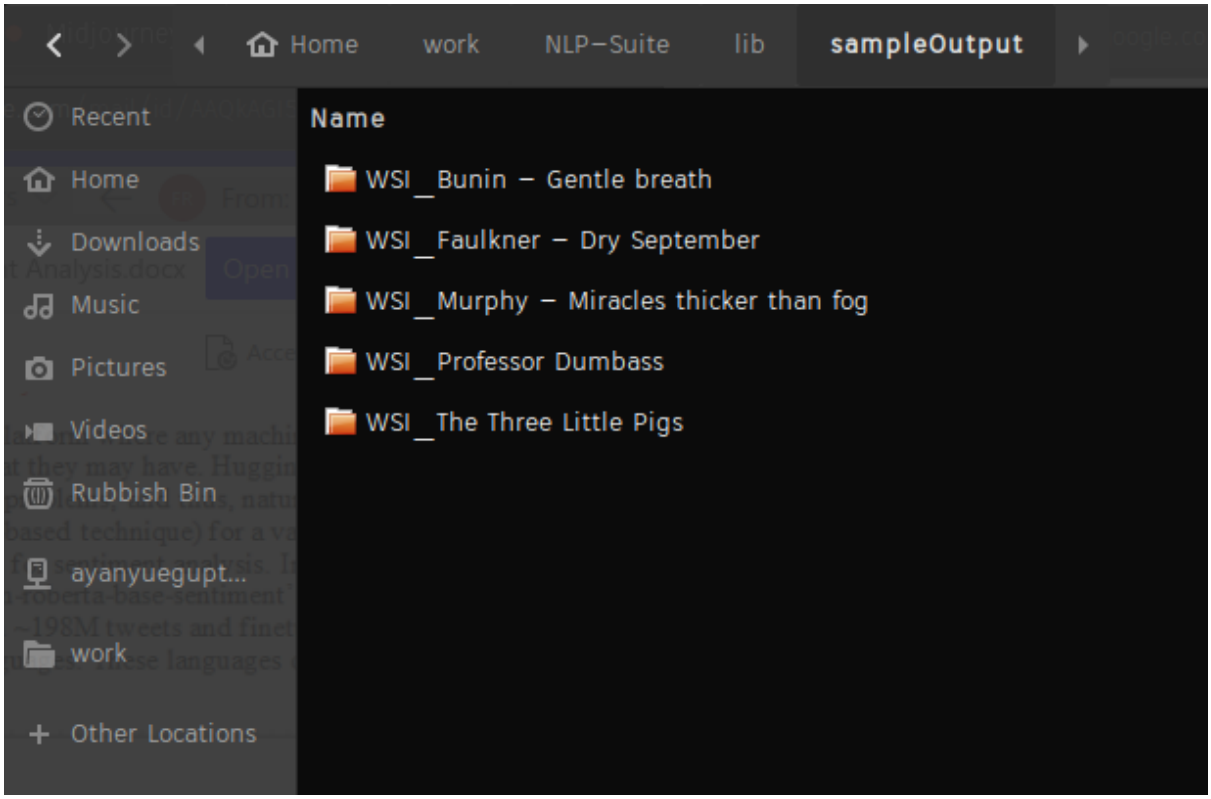


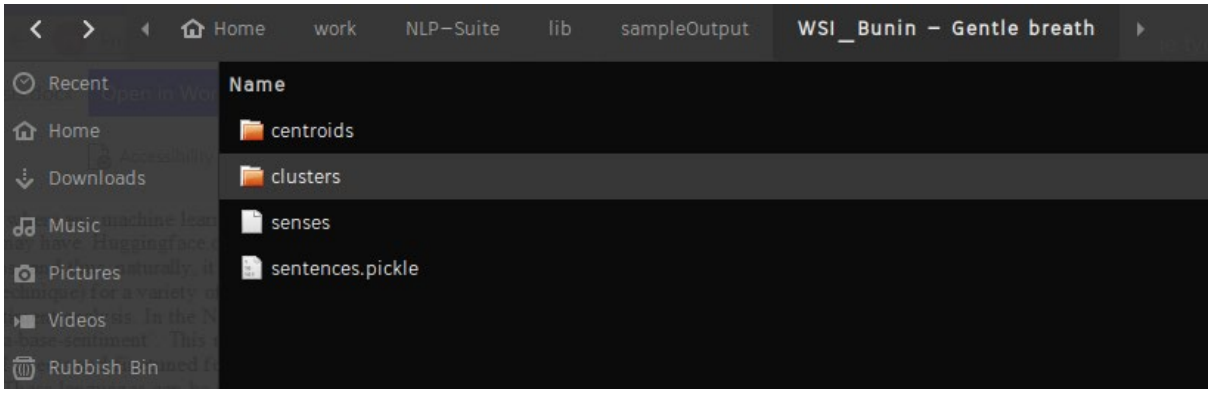
7. Once you have completed the above, press RUN.

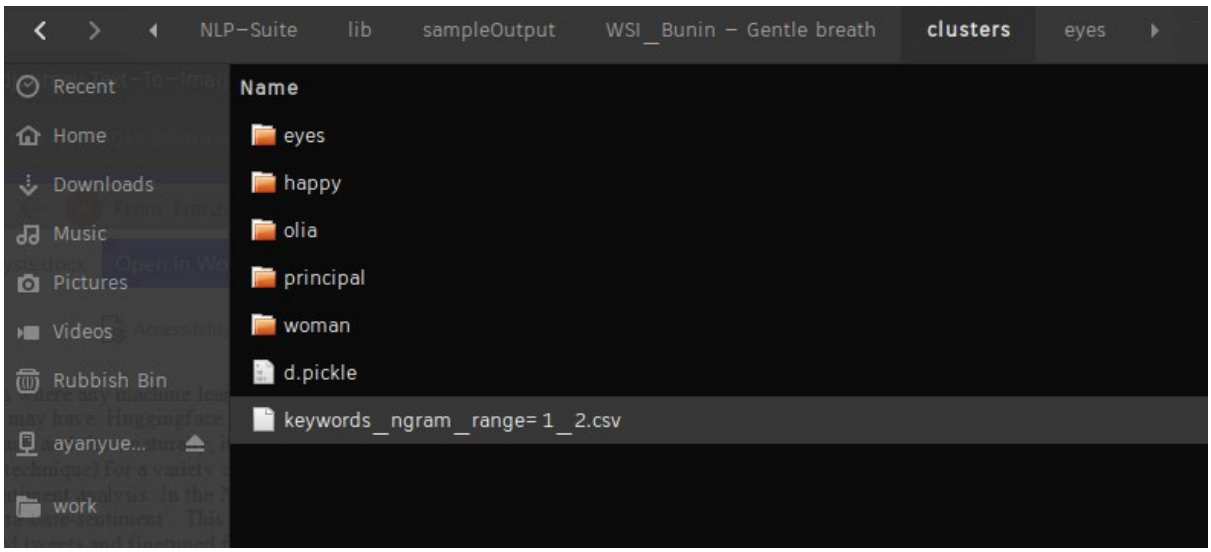


8. The output will be stored in directories entitled with the names of the documents you used prefixed by ‘WSI_’. These directories will be stored in the directory you selected for the output to be stored in step 2.

9. To examine the results of performing WSI on your document, select the relevant ‘WSI_’ directory and click the directory entitled ‘clusters’.
10. N-grams summarising the content of each sense cluster will be found in the ‘keywords_ngram_range=**(lower bound)_(upper bound)**.csv’ file. Other results for each individual word for which senses were extracted are stored in separate directories (in the example below the directories containing the WSI results for the words ‘eyes’, ‘happy’, ‘olia’, ‘principle’ and ‘woman’ can be seen):







11. In each directory containing the results for an individual word, a text file containing the occurrences of the word grouped by the sense they express and a pie chart visualising the proportion of occurrences that express each sense can be found.

References

Lucy L, Bamman D. 2021. “Characterizing English Variation across Social Media Communities with BERT.” *Trans Assoc Comput Linguist 9:538–556*. doi: 10.1162/tacl_a_00383