Contents

[Folder Structure 2](#_Toc497076266)

[Input Configuration 2](#_Toc497076267)

[Running the Code 2](#_Toc497076268)

# Folder Structure

All codes, including the commented code, are stored in the same folder. Typically, the code folder will have the following structure:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No** | **Name** | **Type** | **Purpose** | **Mandatory?** |
| 1 | \_\_pycache\_\_ | Folder | Temporary files | N |
| 2 | NLP | Folder | Data folder | Y (Data files) |
| 3 | big.txt | File | Sample file for testing module | N |
| 4 | commented.py | File | Maintaining commented code | N |
| 5 | dep\_parsing.py | File | Syntactic parsing of sentences | N (Currently inactive) |
| 6 | in\_file.cfg | File | Configuring input file | Y |
| 7 | lemmatization.py | File | Different functions for lemmatization | Y |
| 8 | main.py | File | Base code that uses all functions | Y |
| 9 | modeling.py | File | Functions used for modeling: classification, word2vec, topic modeling, etc. | Y |
| 10 | parser.cfg | File | Text file with Stanford parser configuration | N (Currently inactive) |
| 11 | pos\_tagger.cfg | File | Text file with Stanford tagger configuration | N (Currently inactive) |
| 12 | pos\_tagging.py | File | Functions used for part-of-speech tagging | Y |
| 13 | sample\_cleaned.txt | File | Processed sample (given) file | N |
| 14 | tokenization.py | File | Functions used to tokenize sentences | Y |
| 15 | treetagger.cfg | File | TreeTagger input folder configuration | Y |
| 16 | util.py | File | Utility functions | Y |
| 17 | visualizing.py | File | Visualizing output of text mining. Eg: TSNE | N (But implemented) |
| 18 | wd.cfg | File | Configuring the code directory | N (Deprecated) |

# Input Configuration

Text input to the code can be configured using ‘input.cfg’.

1. First line is either the relative path of the input with respect to the code directory or the absolute path of the input file.
2. Second line is the type of input file – currently supported formats are csv and text.
3. Third line is the name of the column that contains the text.
4. Fourth line is the name of the column that contains the labels.

# Running the Code

1. Update the relative path of input file and configuration as per the previous section.
2. Open terminal.
3. Change directory to code directory.
4. Type “python main.py”.