4045_NLP_Readme

4045_NLP

4045 Natural Language Processing

Installing Dependencies

Manual Installation

1. OpenJDK 8 - GNU General Public License 2.0
2. Python 3.7.4 - PSF Licence

Manual Download

1. Stanford CoreNLP Server - GNU General Public License 3.0
   1. unzip zipped file
   2. move `stanford-corenlp-full-2018-10-05` to *Desktop*
2. Stanford CoreNLP NER - GNU General Public License 3.0
   1. unzip zipped file
   2. enter directory `stanford-ner-2018-16`
   3. copy `stanford-ner.jar` to *Desktop*
   4. enter directory `classifiers`
   5. copy `english.all.3class.distsim.crf.ser.gz` to *Desktop*

Required Python Libraries

1. MatPlotLib - PSF Licence
2. NumPy - NumPy License
3. nltk - Apache License Version 2.0
4. Spacy - MIT Licence
5. StanfordNLP - GNU General Public License 2.0

Installation Steps

1. Once Python has been installed, input in *cmd* : `pip install -r requirements_win.txt`
2. From Desktop, copy `stanford-corenlp-full-2018-10-05` into project folder *server*
3. From Desktop, copy `stanford-ner.jar` and `english.all.3class.distsim.crf.ser.gz` into project folder *lib*
4. Place `reviewSamples20.json` and `reviewSelected100.json` into project folder *data*

Launch Project

1. cd to root of project folder (i.e. `contains main.py`, `requirements_win.py`)
2. Input in *cmd* : `python main.py`
3. There will a prompt to install `en_ewt` and `en_gum` models, enter `y` to install
4. Await till program ends
5. Outputs accessible in project folder *out*

Sample Out

1. Writing Style
    1. `a_reviews.txt` : list of reviews, 1 entire review paragraph per line
2. Sentence Segmentation
    1. `b_segmented_sentences_*.png` : graph of distribution
    2. `b_segmented_sentences_*.csv` : data used to plot graph
3. Tokenisation and Stemming
    1. `c_distribution_with*_stem.png` : graph of distribution
    2. `c_common_(before|after)_stem.csv` : csv of most frequent token, arranged from most to least frequent
4. POS Tagging
    1. `d_pos_tagged.json` : sentences are spliced into their token, appended with POS tags
5. Most Frequent Adjective
    1. `e_frequent_*.csv` : csv of most frequent adjectives, arranged from most to least frequent
    2. `e_indicative_*.csv` : csv of most indicative adjectives, arranged from most to least frequent
6. Noun-Adjective Pair Summariser
    1. `f_noun_adj_pair_*.json` : csv of most frequent noun-adjective pairs (after categorisation), arranged from most to least frequent
    2. `f_noun_adj_pair_*_old.json` : csv of most frequent noun-adjective pairs (before categorisation), arranged from most to least frequent
7. Application
    1. `g.neg_sents_results.txt`
        1. 1st line > total number of statements with negations
        2. 2nd+ line > statements with negation