README

Environment

1. python 3.6
2. python packages:

   NLTK, re, pandas, json.
3. java package:

   stanford-corenlp-full-2017-06-09/stanford-corenlp-3.8.0.jar,

   stanford-corenlp-full-2017-06-09/stanford-corenlp-3.8.0-models.jar

Code Explanation

As stated in report, the whole process includes data cleaning, target extraction

Data Cleaning

the codes are sotred in directory *data_process*

1. *clean_data.py* : this code is used to remove meaningless marks

   input: *raw_data.csv*, output: *step1_data.csv*
2. *remove_non_english_sents.py* : this code is used to remove reviews not writen in English

   input: *step1_data.csv*, output: *step2_data.csv*
3. *groupData_base_on_id.py*: group data based on course id.

   input: *step2_data.csv*. output: 36 course review files

All data used in above are stored in directory *data*, and 36 course review files are stored in the subdirectory *course* under *data*. These 36 files are named in form of *course_"course_id".csv*

Target Extraction

The codes are stored in directory *extract_target*.

1. *extract_target_list.py*: this code is to grow target and opinion list by propagation

   input: *course_"course_id".csv*, output: *course_"course_id"_target_list.txt*

2. *extract_target_phrase.py*: this code is to extract target word or phrase for each review.

   input: *course_"course_id".csv*, *course_"course_id"_target_list.txt*

   output: *course_"course_id"_transaction.csv*

3. *filter_target.py*: this coude is used to prun the targets.

   inuput: *course_"course_id"_transaction.csv*

   output: *course_"course_id"_target_filtered.txt*

   All output files produced by this part are stored under the directory *result*.

Result

Final result stored in the files names as *course_"course_id"_target_filtered.txt*.