Movie Review dataset

	Dataset	Text type	Length	Type
Sentiment polarity	1- Polarity_html	Unprocessed	27886	Original source
	2- review_polarity	Pos/Neg 1000/1000	2000	Sub data
	3- rt-polaritydata	Pos/Neg 5331/5331	10662	Sub data
Sentiment scale	4- scale_whole_review	Unprocessed	5006	Original source
	5- scale data	Rating	5006	Sub data
Subjectivity	6- subjectivity_html	Unprocessed	1620	Original source
	7- rotten_imdb	Sub/Obj 5000/5000	10000	Sub data

Info :

1_ Polarity_html :

_ The original (unprocessed, unlabeled) source files from which the processed, labeled, and (randomly) selected data(review_polarity) was derived.

2_ review_polarity :

_ txt_sentoken: 2000 text
- _ pos : 1000 txt
- _ neg : 1000 txt
- _ Each line = a single sentence
- _ terminology : pos/cv000_29590.txt :
  - _ Label : pos
  - _ extracted from the file 29590.html in polarity_html
- _Label Decision : " 8/10 ", " three-and-a-half out of five ", and " OUT OF **** : *** " , “ B : + / C- : -”

3_ rt-polaritydata :

* rt-polarity.pos : contains 5331 positive snippets
* rt-polarity.neg : contains 5331 negative snippets

_ Each line = a single sentence
_ parent source: subjectivity/subjective folder
_ Label Decision: in parent .html file, marked with ``fresh`` are positive, ``rotten`` are negative.

4_ scale_whole_review :

_The original : before passing through tokenization, sentence separation, and subjectivity extraction.
_ there are four reviewers sub-directories
_ at each:
_id.txt: each line = one paragraph of the review

5_ scale_data :

_ there are four author sub-directories
_ at each:
_id : file id's in polarity_html
_ label.3class : for the {0,1,2} three-category classification task
_ label.4class : for the {0,1,2,3} four-category classification task
_ rating(normalized ratings) : [0-1] with step size 0.1
_ subj : review text

example:
Steve+Rhodes :
_id : 11790
_ label.3class : 0
_ label.4class : 0
_ rating(normalized ratings) : 0.1
_ subj : this bit of lame physical humor is typical of disney's ...

6- subjectivity_html : (parent of rotten_imdb)

_ obj/2002,2003 : html files (all sentences from IMDb plot summaries are objective)
_ subj/2002 :plot summaries (all snippets from the RottenTomatoes are subjective)

7- rotten_imdb :

_ quote.tok.gt9.5000 : 5000 subjective sentences
_ plot.tok.gt9.5000 : 5000 objective sentences
_ each line = single sentence (at least 10 tokens)