

Parameters

Table 1: Parameters in ‘modeling.py’

	Argument	Description	Possible value / Notes
-w	-workdir	Working directory	NOT NULL
-d	-datadir	Data source	Will download sample files if no value passed
-l	-labelfile	Ground truth annotation file	
-i	-deid	Deidentification	T
-p	-parser	Using concept parser	ctakes
-q	-parserdir	Directory of cTAKES	
-s	-ctakessetup	Setting up cTAKES	Only for the first time user, need to pass the following four arguments too
-n	-utsname	UTS account	
-o	-utspw	UTS password	
-e	-ctakesminmem	Minimal cTAKES memory usage	2g
-g	-ctakesmaxmem	Maximal cTAKES memory usage	6g
-c	-convert	Convert cTAKES output XML files further processing	T
-m	-model	Whether performing modeling	T
-f	-feature	Type of feature extraction	bow, snomed, umls, sg, st or any combination, e.g. bow_sg
-v	-vec_represent	Vector representation methods	freq, tfidf, pv
-a	-algorithm	Machine learning algorithms	l1, l2, nb, svm, svmrbf, rf, adaboost, gbm
-b	-balancedclass	Whether using balanced class weight	balanced
-r	-repeat	Repeated cross validation	3
-k	-cv	n-fold cross validation (1 = no cross validation)	5

Table 2: Parameters in ‘predict.py’

	Argument	Description	Possible value / Notes
-d	-datadir	Data source	Will download sample files if no value passed
-p	-parser	Using concept parser	ctakes
-q	-parserdir	Directory of cTAKES	
-m	-model_file	The path directs to your model (.pkl)	

Footnote

The dataset used in this project was downloaded from iDASH repository (<https://idash-data.ucsd.edu>) supported by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant U54HL108460.