

Data Folder :

1. directionality\_dictionary\_all\_df.csv: The manually created dictionary

Notebook(s) for each part:

1. Data Preprocessing: data\_wrangling.ipynb
   1. input：topic\_tagged\_data\_set & topic\_untagged\_data\_set
   2. output:
      1. processed data set
      2. 1-gram & 2-gram vectorized feature matrix ( in pickle file)
2. Topic Classifier & Auxiliary Growth topic Classifier: topic\_classifier.ipynb
   1. input: processed data set
   2. output:
      1. tfidf\_vec.pickle: tfidf transformed feature matrix
      2. rf.pickle: random forest topic assign model
      3. tfidf\_vec\_is\_growth.pickle: tfidf transformed feature matrix for training the auxiliary growth topic classifier
      4. logi\_clf.pickle : logistic regression is\_growth classifier
      5. sample.csv: the result of the topic classifier and the original data
3. Directionality Analyzing Algorithms:

Baseline model: Directionality\_Analyzing\_Algorithms\_benchmark.ipynb

* 1. input: direction tagged data (in file “manual\_direction\_eng\_completed”)
  2. output: performance for different model showed in the notebook

1. Get feature Contribution (model preparation) : feature\_contribution\_train.ipynb
   1. this part of the notebook is used for getting the features which contributes the most to the topic assignment of the sentence
   2. this is part of the preparation for directionality prediction model
   3. output: excel with the most important features, in file “feature\_contribution\_train”
   4. eg. [('labor', 0.36359374997240984), ('labor market', 0.25873850511490115)], ‘labor’’ contribution for the result in 0.36 among all the 1-gram and 2-gram features.
2. Get word similarity to direction word(model preparation): word\_similarity\_train.ipynb
   1. this part of the notebook is used for getting the highlight direction word in the sentence, we get the direction word based on two algorithms, stem recognition and word similarity recognition
   2. output:
      1. excel with information of the direction word in the sentence and the corresponding direction word in the dictionary as well as their similarity.
      2. For each sentence, we return 7 lists (because there might be more than one direcitonality words in one sentence):
         1. list of indices of directionality words
         2. list of indices of corresponding dictionary words
         3. list of directionality words
         4. list of corresponding dictionary words
         5. list of similarities (between 0 and 1)
         6. list of number of negative words in front of the directionality words. We only consider two words preceding the directionality words.
         7. list of directionalities
3. Directionality Analyzing Algorithms：

Unsupervised model:Directionality\_Analyzing\_Algorithms\_unsupervised.ipynb

* 1. get the prediction of the directionality based on our unsupervised model.
  2. For each sentence we return 16 types of directionality score
  3. These scores are based on the combination of 2 distance types(absolute position and dependency parse tree shortese path(spacy)), 2 types of directionality word recognition(stem and similarity) and 4 types of aggregate methods(simple average, weighted by inverse of the distance, weighted by contribution, weighted by inverse of the distance and contribution)
  4. detailed result showed in the notebook

1. Backtesting strategy:Portfolio backtesting.ipynb
   1. input: prediction result and the financial index to compare with
   2. output: simulation and evaluation of the performance of the portfolio.