**1. all\_article\_extract.py**

Extract all english articles.

**2. generate\_english\_parallel\_trie.py**

Sentence tokenization of all articles with nltk and company recognition via substring match. Parallel is based on multiprocessing module

Original article to: sentence, sentence\_with\_location and lda input (word only)

**2\*. generate\_english\_parallel\_trie\_new.py**

Sentence tokenization of all articles with nltk and company recognition via substring match. Parallel is based on joblib module

**3. Source-LDA in C++**

**4. parse\_lda\_output.py**

Parse article weights from LDA output 'theta.txt'

Select weight top 20 article relative to topic(‘Green\_Economy’) and save. # lable1 means not in key\_file\_set.(what is the difference between total? Neg?)

**5. prolong.py**

Look down k sentences to prepare for (test) data of BERT.

**6. BERT in Colab**

**7. prolong\_add\_result.py**

Add predicted labels from BERT into original data frame.

**8. logistic\_regression\_prepare\_daily.py**

Prepare data for logistic regression prediction on daily data. (Article average + Daily weighted average + Moving average)

**9. logistic\_regression.py**

Logistic Regression for PD prediction.