This is a brief introduction of all files. At the same time, there are more detailed descriptions in each folder’s ‘readme’ file.

‘**origin**’ contains all MDA part in 10-K report from 1996 to 2006;

‘**token**’ contains all tokenized MDA part in 10-K report from 1996 to 2006;

‘**logvol**’ contains all stock log volatility data from 1996 to 2006;

‘**match**’ contains some intermediate processes aims to match different years together;

‘**word importance**’ contains mainly the result of the word list that having the highest relationship with stock volatility from different years.