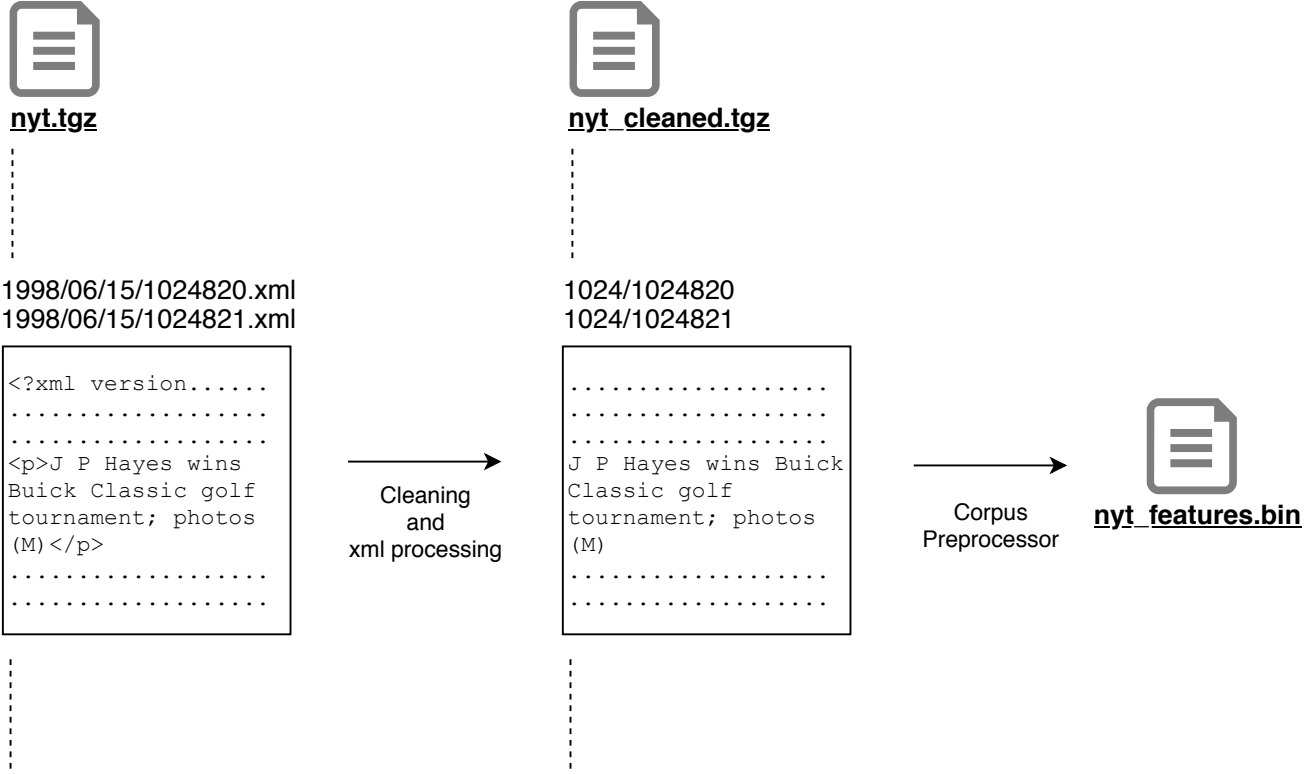
nyt.tgz
nyt_cleaned.tgz
1998/06/15/1024820.xml
1998/06/15/1024821.xml
<?xml version.....
.....
.....
J P Hayes wins
Buick Classic golf
tournament; photos
(M)
.....
.....
Cleaning
and
xml processing
1024/1024820
1024/1024821
.....
.....
.....
J P Hayes wins Buick
Classic golf
tournament; photos
(M)
.....
.....
Corpus
Preprocessor
nyt_features.bin