**Retrieving Titles and Subtitles – Fixing the Code 1**

**Details**

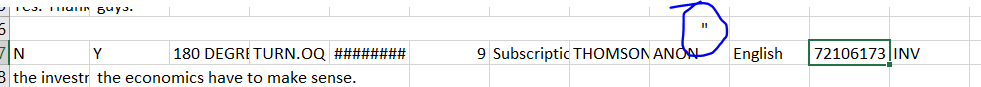
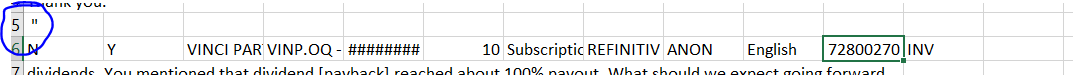
Total: 77748 entries

* Version 2: All entries, including entries with keyword matches but without gvkey matches. There are 77884 entries.
* Version 2\_DuplicateDropped: Dropped duplicates (for report, title, subtitle and date) in Stata. 136 duplicates were dropped and 77748 entries remain.
  + 3939 entries / 3473 unique entries with missing titles and subtitles
  + 0 with titles but missing subtitles
  + 8416 with titles but 'Â' as the subtitle
* Version 3 and 4: Added back titles and subtitles for entries, but with an incorrect *mergeclean.do*.
  + Many entries with missing titles and subtitles, without further manual matching
* Version 5: Added back titles and subtitles for entries without gvkey matches. Deleted the line of Stata code in *mergeclean.do* removing entries in the dataset with titles/dates/subtitles that didn’t match with the gvkey dataset. There are 77748 entries. After the ‘more inclusive’ merge, there are:
  + 88 entries / 78 unique entries with missing titles and subtitles
  + 0 with titles but missing subtitles
  + 8416 with titles but 'Â' as the subtitle

Fixing the line of code in *mergoclean.do* significantly reduced the number of missing entries, but did not eliminate all of them.

**Possible Reason for the Remaining Missing Entries**

After checking the missing entries, it seems that there is a recurring pattern of a “ appearing before the entry in the csv file, which could have led to some error with how the dates/titles/subtitles are extracted. However, we note that these conference calls still appear in keyword matches, so the error doesn’t seem to “ripple down the processing pipeline”.



* Direct solution: Merge directly with the .xls source files and not the intermediate files.
* Longer-term: Find out where the algorithm is not working and fix that part.