Intermediate Report

Megha Gupta(MT11024), megha1124@iiitd.ac.in

March 23, 2013

1 Progress

- Preprocessing and Data generation:
  Downloaded the Logfiles from the file server. These Logfiles (xml format) maintain the history of old OCR Text and corresponding corrected text. These logfile names were decoded into Newspaper name + Date of Newspaper. For example,
  Logfile name – AL19000105-changes.log
  AL19000105-changes.log – Amador Ledger, 1900-01-05
  Then each decoded name was translated into eight individual URLs.
  http://chroniclingamerica.loc.gov/lccn/sn93052980/1900-01-05/ed-1/seq-1/ocr.txt
  http://chroniclingamerica.loc.gov/lccn/sn93052980/1900-01-05/ed-1/seq-2/ocr.txt
  http://chroniclingamerica.loc.gov/lccn/sn93052980/1900-01-05/ed-1/seq-3/ocr.txt
  http://chroniclingamerica.loc.gov/lccn/sn93052980/1900-01-05/ed-1/seq-4/ocr.txt
  http://chroniclingamerica.loc.gov/lccn/sn93052980/1900-01-05/ed-1/seq-5/ocr.txt
  http://chroniclingamerica.loc.gov/lccn/sn93052980/1900-01-05/ed-1/seq-6/ocr.txt
  http://chroniclingamerica.loc.gov/lccn/sn93052980/1900-01-05/ed-1/seq-7/ocr.txt
  http://chroniclingamerica.loc.gov/lccn/sn93052980/1900-01-05/ed-1/seq-8/ocr.txt
  where seq-1 represents the sequence of the pages of the newspaper.
  Finally, downloaded the Raw OCR text from each of those URLs with the help of Python scripts to build the Raw OCR corpus.

- Analysis: In order to know the types of error corrections performed by users, I am analyzing the logfiles by counting the corrections made in each file. So that we can decide upon a threshold, above which we know files would contain sufficient data.
  Further to classify the errors to know whether they belong to spellcheck, omission of garbled text, Puctuation error or Capitalization error category, I am comparing the tokens from the old text and the new text. If there is no variation in both the tokens, then it is classified as, ”No Correction” else Correction. All this is done using Python scripts and the results are stored in csv files.

- Build Corrected OCRtext Corpus, Model and Query Set: The third phase that is still to be done is to build Corrected Corpus, IR Model, Query Set and using them to determine the effectiveness of retrieved results.