Project Proposal

Megha Gupta, MT11024

February 25, 2013

1 Project Participants

This is an individual project.

2 Summary of Proposed Research

This research is primarily focussed on retrieval effectiveness of OCR data versus Corrected data. Since OCR data gives lot of garbled text considering very poor print quality of the input which are the old historic American Newspapers. Here, we are trying to figure out what are the effects of noisy data on text retrieval. If there is then how big it is. Should the garbled text be corrected in order to get substantially good results. To know the difference between the results we need to find a heuristic measure that determines the difference between the retrieval effictiveness of OCR data versus Corrected Data.
This work can be used to determine whether the OCR text is sufficient enough to be used for the thorough text search and analysis and does it meet the user expectation without actually enhancing and enriching the text.

3 Novelty

There is not much reading material available on this subject. Some experiments on the related work has been performed but they are not quite exhaustive and conclusive. So, i plan to do some more research and come out with a conclusive solution.

4 Algorithm Description

Steps involved to accomplish the proposed goal are as follows:

- Extract both the Datasets including OCR Text and its corrected version
- Use Bag of Words model to represent documents.
- Formulate a query set

- Resultset generated from the Corrected Text proves to be the benchmark for our evaluation(F-Measure) for the results from OCR Text.

5 Datasets and Software

Dataset for OCR text corresponding to be used is extracted from the website [oC09] using the python scripts.

6 Evaluation

To evaluate the effectiveness of the results of the query applied on OCR text we create a benchmark consisting of results from the same query applied on Corrected Text.

7 Analysis

The analysis will be done on the basis of F-Measure/F-score which is the harmonic mean of Precision and Recall.
$F_\beta = (1 + \beta^2) \times \frac{PR}{\beta^2 P + R}$

8 Proposed Future Work

This project would carry out experiments for one model or Information Retrieval environment, rest of the models can be considered in future.

9 Bibiliography

References

[KT96] Allen Condit Kazem Taghva, Julie Borsack. *Evaluation of Model Based Retrieval Effectiveness with OCR Text*, volume 14. New York, US, 1996.

[KTE93] Allen Condit Kazem Taghva, Julie Borsack and Srinivas Erva. *Effects of Noisy Data on Text Retrieval*, volume 45. Las Vegas, US, 1993.

[oC09] Library of Congress. http://chroniclingamerica.loc.gov/, June 2009.