SummarizerScratch

<https://github.com/silverstar194/SummerizerScratch>

**Overview:**

This summarization tool will take a PDF, extract the text and provide a list of the most important sentences. It bases importance on frequency of words and position of sentence within the PDF. Loosely based on Summarization of Document using Java by Priyanka Sarraf (Banasthali University) and Yogesh Kumar Meena (Assistant Professor, MNIT). Also aimed to complete “How to Build an Automated Text Summarizer: An extraction-based summarizer” CS784 Spring 2013 David R. Cheriton School of Computer Science (although I am not enrolled in the course). See documents directory for more information on the research paper and course.

**Program Flow:**

1. PDF filepath and compression ratio provided by user
2. Using PDFBox text is extracted and broken into sentences
3. Each word is stemmed using Porter stemmer algorithm
4. Stemmed words occurrences are counted
5. Sentence keyword score is calculated based on number of keywords and frequency weight of the keyword
6. Sentence position score is calculated – with middle having highest weight and beginning and end lower weight
7. Position score and keyword score are each weighted with constants from Config.java
8. Highest Rated sentences (up to ratio limit) are selected and ordered based on position
9. Sentences are written to summary PDF using PDFBox

**Future Improvements:**

-Incorporate cue words

-Move away from linear weight of keywords

-Highlight original text within PDF

-Allow scores from adjacent sentences to compound