The new Government Security Classifications Policy will be used from April 2, 2014 [1], exacerbating the already difficult task of securing sensitive government information. Adapting will take time; automatic document categorization technology can help but categorization algorithms fail with documents that contain a mixture of data with different sensitivity levels. This problem can be approached with text segmentation, a technique consisting in breaking linearly a text into topically-consistent segments. Individually categorizing each segment can then be done accurately and the most restrictive policy (e.g., Top Secret) is then applied to the entire document, preventing potential data leaks.

We present a novel automatic segmentation algorithm that takes advantage of the B2S feature extraction [2], based on principles from image processing and especially on the Helmholtz Principle from the Gestalt Theory of human perception. Our algorithm is entirely unsupervised and flexible to allow segmentation at different scales, such as short paragraphs and large sections. It is also domain independent and even virtually language independent. Experimental results suggest it compares favourably with current state-of-the-art algorithms [3].

### Problem Statement

- Text documents typically contain a succession of topics, in the form of small paragraphs or larger chapters.
- Our goal is to identify **topically-coherent** segments in a document. In other words, what are the gaps that best separate the topics inside the document?

### Text Preprocessing

- Raw text documents must initially be converted into well-defined linguistic units that our algorithm can work with.
- We tokenize the text into sentences and words, then apply stemming (i.e., keep only the root of each word).

### B2S Feature Extraction

- Localized word repetitions suggest a locally strong lexical cohesion while vocabulary changes indicate topic shifts.
- Thus, we can identify topic boundaries using some measure of word repetitions, which can be quantified using the Helmholtz Principle:

  \[
  \text{meaning}(t, W_i) = -\frac{1}{m} \log \left( \frac{K}{m^m} \right) > 0, \quad \text{where:}
  \]

  - \( N = |W_i| \)  \( d \) and \( w \) are the lengths, in words, of the document and the window respectively.
  - \( K \) is the frequency of \( t \) in the document.
  - \( m \) is the frequency of \( t \) in the window.

- The frequency of a term \( t \) is **unusually high** in a window \( W_i \) if \( \text{meaning}(t, W_i) > 0 \), where:

  \[
  N = |W_i|, \quad d \text{ and } w \text{ are the lengths, in words, of the document and the window respectively.}
  \]

  \[
  K = \text{frequency of } t \text{ in the document.}
  \]

  \[
  m = \text{frequency of } t \text{ in the window.}
  \]

**“Meaningful features and interesting events appear from large deviations from randomness.”**

- A plot of the unsmoothed & Gaussian smoothed gap scores for a real document. Nine real segment boundaries are precisely identified, and one is slightly misplaced.

### Segmentation Algorithm

1. Set the window size for B2S feature extraction to the desired average segment size.
2. Initialize all gap scores to 0. Gap scores will capture and quantify word repetitions between the sentences before and after each gap.
3. Compute gap scores.
4. Compute the segment boundaries.
   - Smooth the gap score curve. The smallest gap scores are the best candidates for segment boundaries but gap scores are noisy. Smoothing filters, such as Gaussian or Savitzky-Golay kernels, will decrease the number of local minima until the desired number of segments is reached.

\[
\text{Gauss}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}
\]

\[
\text{SavGolay}(x) = (J'J)^{-1}J'x
\]

- Select the best gaps for placing the segment boundaries.

### Future Work

- We are currently running pilot trials with several HP Business Units to evaluate different applications of our algorithm.

### References

