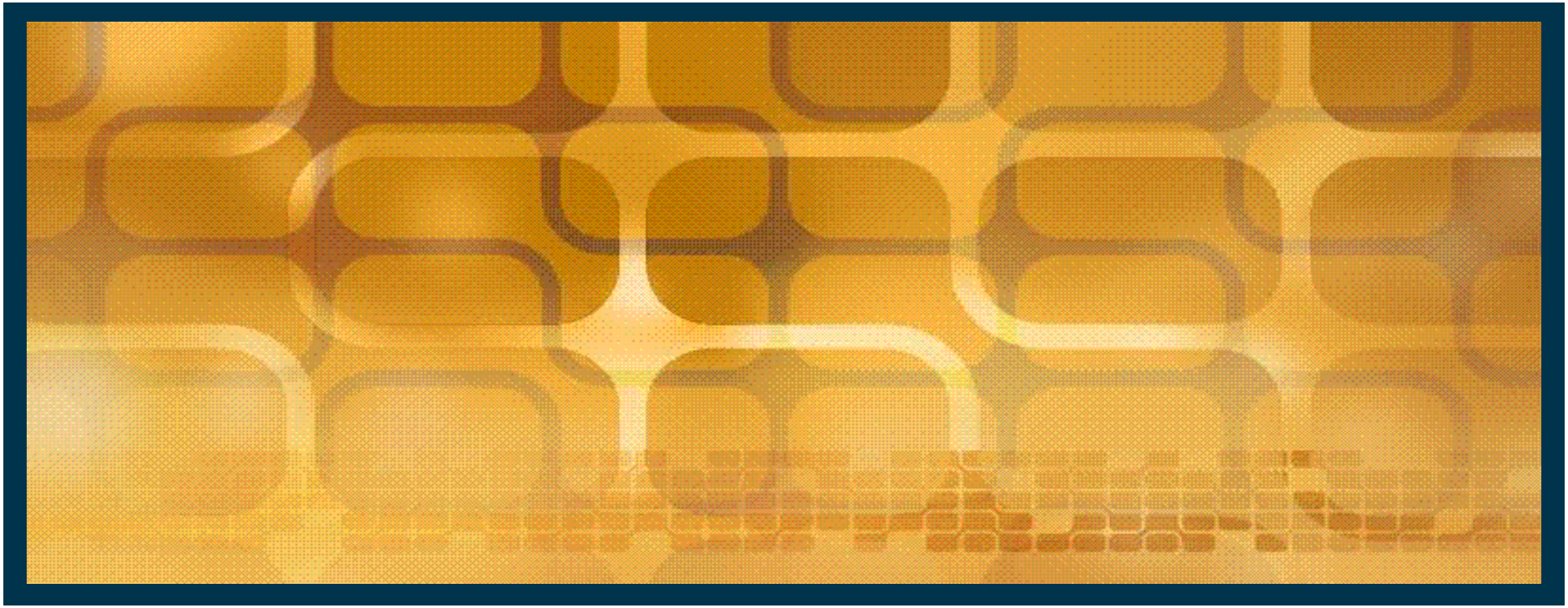
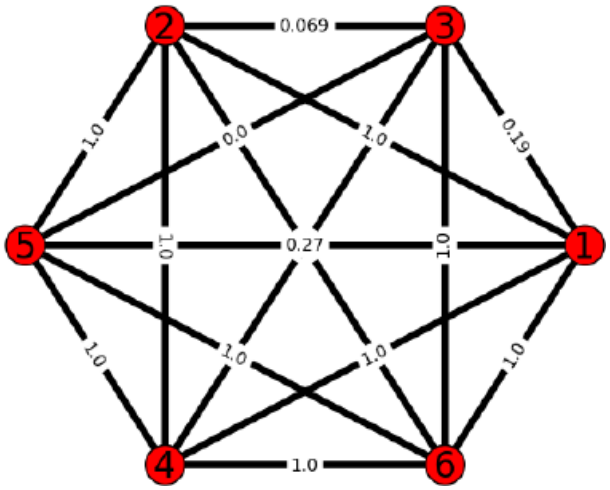


A Graph Based Ranking Strategy for Automated Text Summarization Summary

Charles Bocage

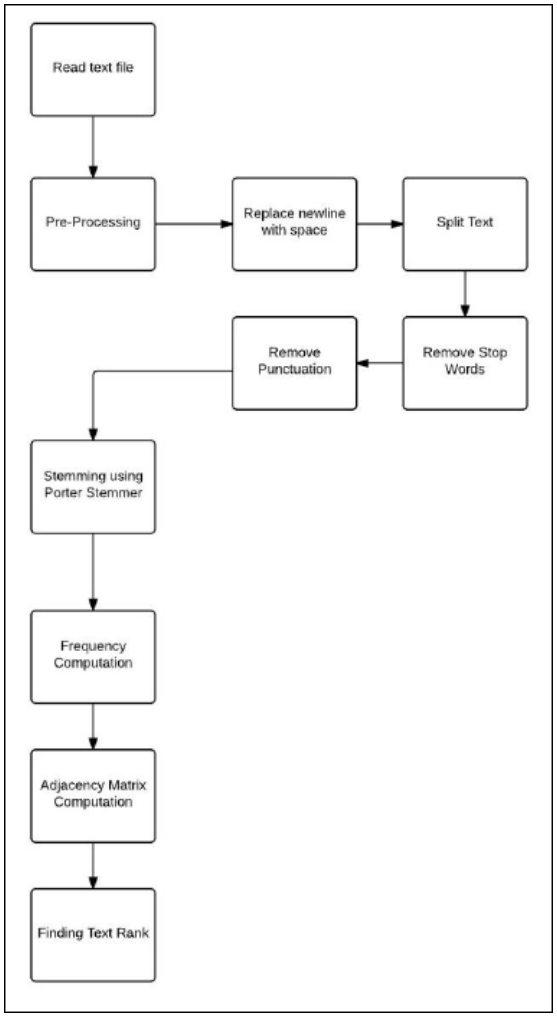
Algorithm Overview

- It is an extraction based algorithm
- It constructs a weighted graph out of the text where the sentences are the nodes.
  - Sentences are split up
  - Stop words are removed
  - Punctuation is removed
  - Words are stemmed
  - Word frequency is computed



- The weights are determined by a distortion measure
  - The semantic relation between two nodes
- The sentences are ranked based on the weights
- The sentences for the summary are chosen until a limit is reached

Algorithm Flow



A flowchart representing the proposed approach from
A Graph Based Ranking Strategy for Automated Text Summarization

Algorithm Parts in Detail

- Read Text File
  - The entire file is read in
- Pre-processing
  - Removal of non ASCII characters
- Replace New Line with Spaces
  - To make it easier to split the new lines are removed
- Split Text into Sentences
  - Uses a regular expression to split the text into sentences:

    (?<!\w\.\w.)(?<![A-Z][a-z]\.)(?<=\.|\?)

    This needs to be modified to handle more punctuation marks

Algorithm Parts in Detail

- Remove Stop Words
  - Stop words are common words that are not useful on their own
- Remove Punctuation
  - Punctuation is removed from each split up sentence
- Stemming Using the Porter Stemmer
  - All of the words are reduced to their stems
  - Kicks becomes kick and so forth
- Frequency Computation
  - Compute a key value pair dictionary where the words are the key and each occurrence is the value

Algorithm Parts in Detail

- Adjacency Matrix Computation
  - Each sentence is compared to each other to get their distortion measure

$$Distortion = \frac{Sum}{non\ common\ words}$$

- Find Text Rank
  - The text rank is computed by

$$TR(V_i) = 1 - d + d \sum_{V_j \in I} \frac{TR(Vj)}{Out(Vj)}$$

  - $TR(V_i)$ = Text Rank of the i^{th} sentence
  - d = damping factor usually 0.85
  - $TR(V_j)$ = Text Rank of the j^{th} sentence
  - $Out(V_j)$ = the degree of j^{th} sentence

References

- Agrawal, Sharma, Sinha and Bagai. A Graph Based Ranking Strategy for Automated Text Summarization. Retrieved from h*ttp://journals.du.ac.in/ugresearch/pdf/J16.pdf*

Questions and Comments