A DOCUMENT SUMMARIZER FOR NOVICES

REX RUBIN
WHY A DOCUMENT SUMMARIZER?

- Getting into a field of research is:
  - Daunting with the amount of information presented
  - Difficult to discern what is important and what isn’t
- How a summarizer will help:
  - Present the most relevant information and remove the excess
EXTRACTION VS ABSTRACTION

- **Extraction[1]**
  - Pulls sentences straight from the input
  - Does not make its own sentences

- **Abstraction[1]**
  - Creates sentences by joining several together
  - Works better for several documents at once
TEXTRANK

- Extraction based[2]
- Creates a web of sentences
- This web is used as an input for PageRank
  - PageRank will rank the sentences[3]
- Gives the summary as the output
HOW TO IMPROVE THIS MODEL?

- It is important to note the glossary should be of relevant terms compared to the original document.
- The way TextRank works, the glossary will allow for similar sentences to connect and score higher.
- This will help by giving more informative sentences.
- It is important to know that more informative does not mean easier to read.
MY TEXTRANK MODIFICATION

Original Document

Combined Document

TextRank

PageRank Assigns Scores

Glossary Removed

Top Sentences Chosen

Final Summary Made

Glossary

Sentences Counted

Graph Created
Will including a glossary of related terms in the original document bring about more informative sentences?
HYPOTHESIS

- Having a glossary included in the original document will bring out more informative sentences in the final summary.
EXPERIMENT OVERVIEW

- Two experimental groups:
  - Control Group (Y)
  - Test Group (X)
- Have the groups take a test on the original document
My summary was made using a document focused on cybersecurity and the glossary was filled with similar cybersecurity terms.
PARTICIPANTS

- Participants:
  - Union College students aged 18-22
  - Mixed group of CS students and non-CS students

- 2 Groups:
  - Control(Y) read the summary that was made through the original TextRank program
  - Test (X) read the summary that was made through my modified TextRank program
The test given to participants was based on the main points of the original document.

- Why the main points?
  - The main points should be in the summary

- Question types
  - 3 Multiples Choice
  - 3 Open Answer
AVERAGE SCORES OF QUESTIONS

Multiple Choice: 3 4 6
Open Answer: 1 2 5
Data on the left is Y and the right is X

<table>
<thead>
<tr>
<th>Question 1</th>
<th>Question 2</th>
<th>Question 3</th>
<th>Question 4</th>
<th>Question 5</th>
<th>Question 6</th>
<th>Total Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.94</td>
<td>0.06</td>
<td>0.22</td>
<td>0.89</td>
<td>0.39</td>
<td>0.56</td>
<td>3.06</td>
</tr>
<tr>
<td>0.94</td>
<td>0.19</td>
<td>0.33</td>
<td>0.44</td>
<td>0.50</td>
<td>0.89</td>
<td>3.22</td>
</tr>
</tbody>
</table>
DIFFERENCES IN RESULTS X - Y

Question 1
Question 2
Question 3
Question 4
Question 5
Question 6
Total Score

0.13
0.11
0.11
0.33
0.16

-0.45
Differences X-Y Outliers Removed

- Question 1: 0
- Question 2: 0.1160714
- Question 3: 0.375
- Question 4: -0.6
- Question 5: 0.0625
- Question 6: 0.571429
- Total Score: 0.625
WAS MY HYPOTHESIS CORRECT?

With these results, I can say my hypothesis is incorrect.
Differences in 4 and 6 were significant
