A DOCUMENT SUMMARIZER FOR NOVICES

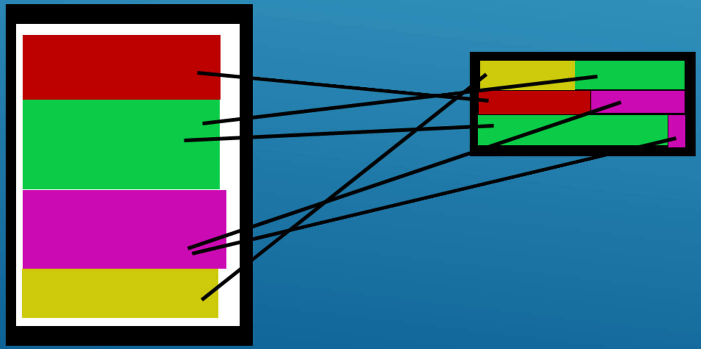
REX RUBIN

WHY A DOCUMENT SUMMARIZER?

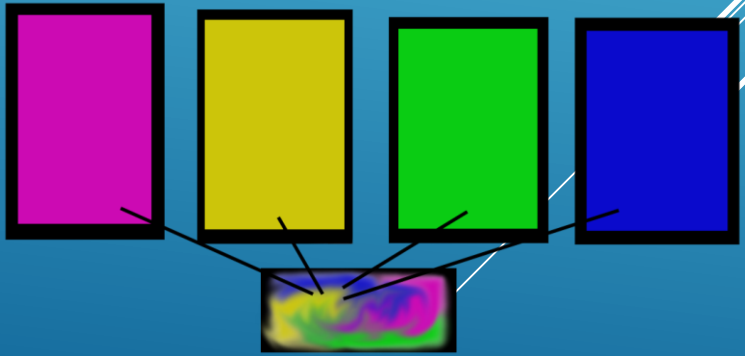
- Getting into a field of research is:
  - Daunting with the amount of information presented
  - Difficult to discern what is important and what isn’t
- How a summarizer will help:
  - Present the most relevant information and remove the excess

EXTRACTION VS ABSTRACTION

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

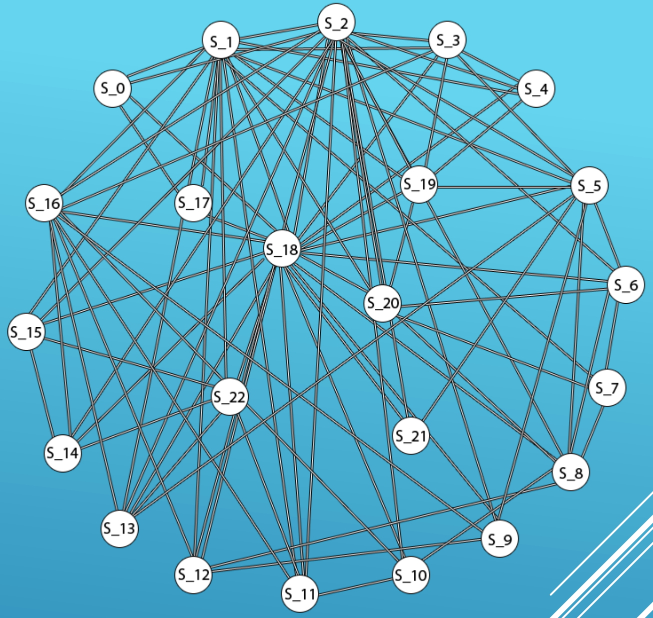


- Abstraction[1]
  - Creates sentences by joining several together
  - Works better for several documents at once



TEXTRANK

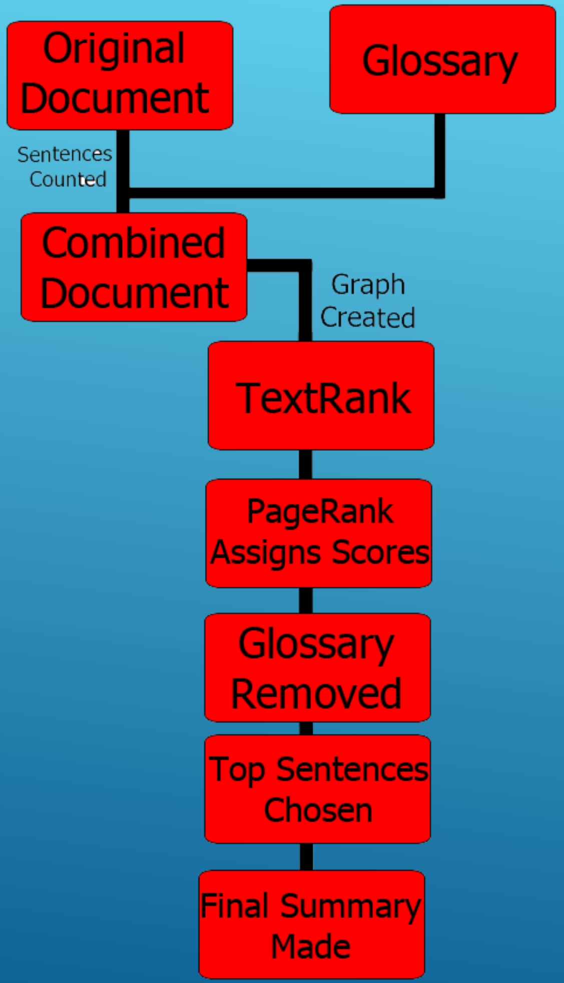
- Extraction based[2]
- Creates a web of sentences
- This web is used as an input for PageRank
  - PageRank will rank the sentences[3]
- Gives the summary as the output



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



RESEARCH QUESTION

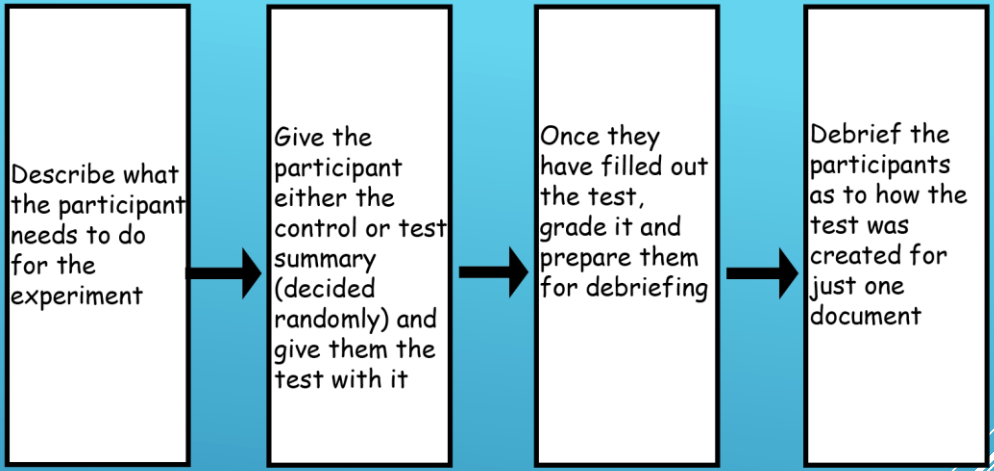
- Will including a glossary of related terms in the original document bring about more informative sentences?

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



Describe what the participant needs to do for the experiment

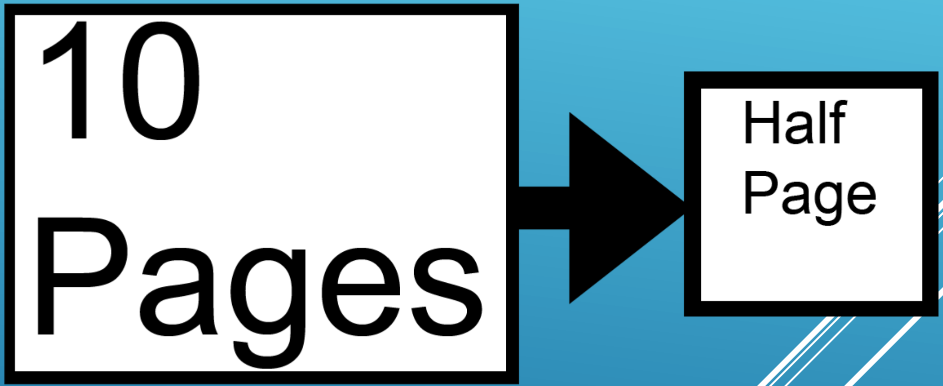
Give the participant either the control or test summary (decided randomly) and give them the test with it

Once they have filled out the test, grade it and prepare them for debriefing

Debrief the participants as to how the test was created for just one document

MY SUMMARY

- My summary was made using a document focused on cybersecurity and the glossary was filled with similar cybersecurity terms



10 Pages

Half Page

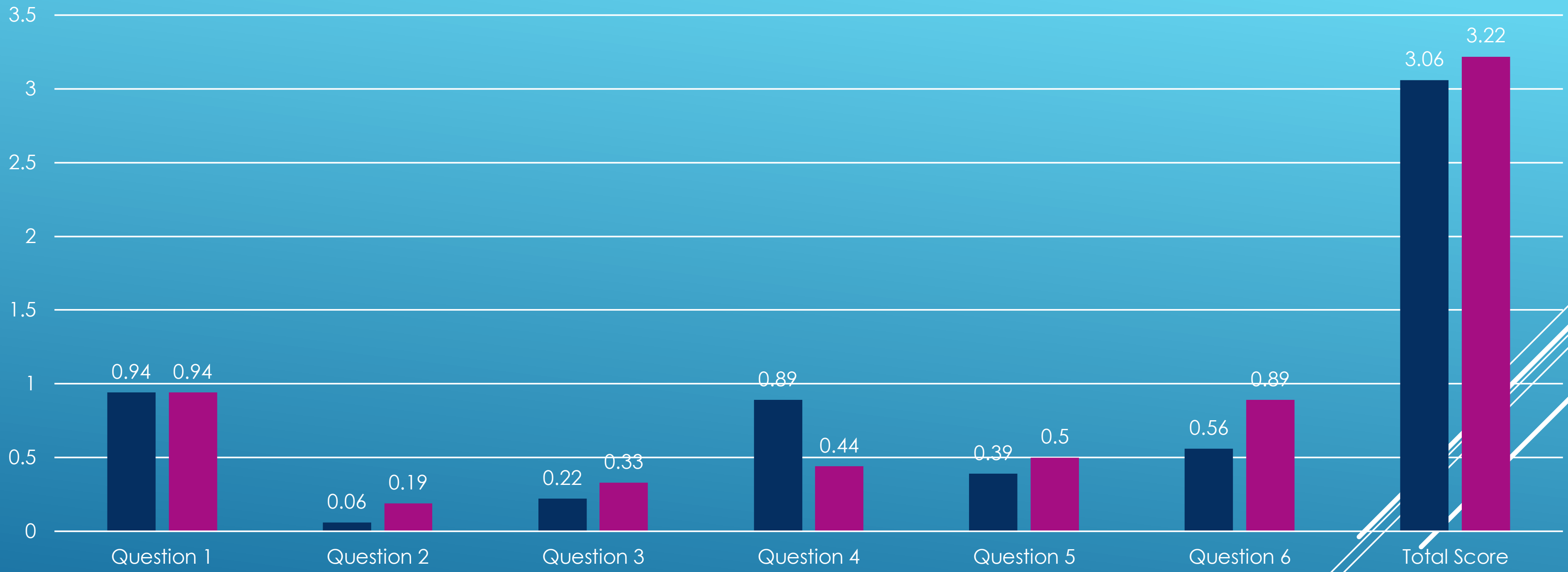
PARTICIPANTS

- 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

TEST GIVEN TO PARTICIPANTS

- 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

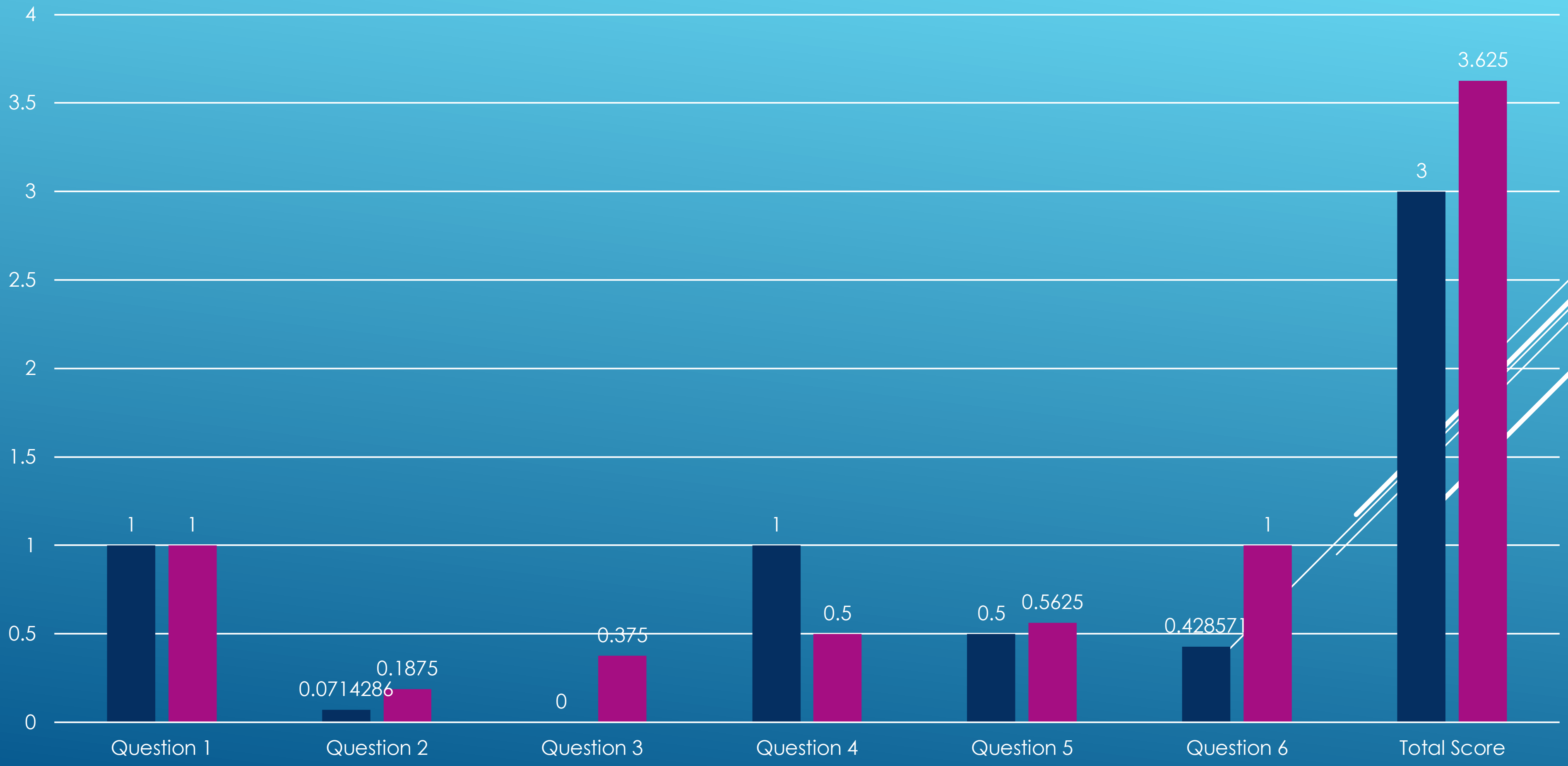
AVERAGE SCORES OF QUESTIONS



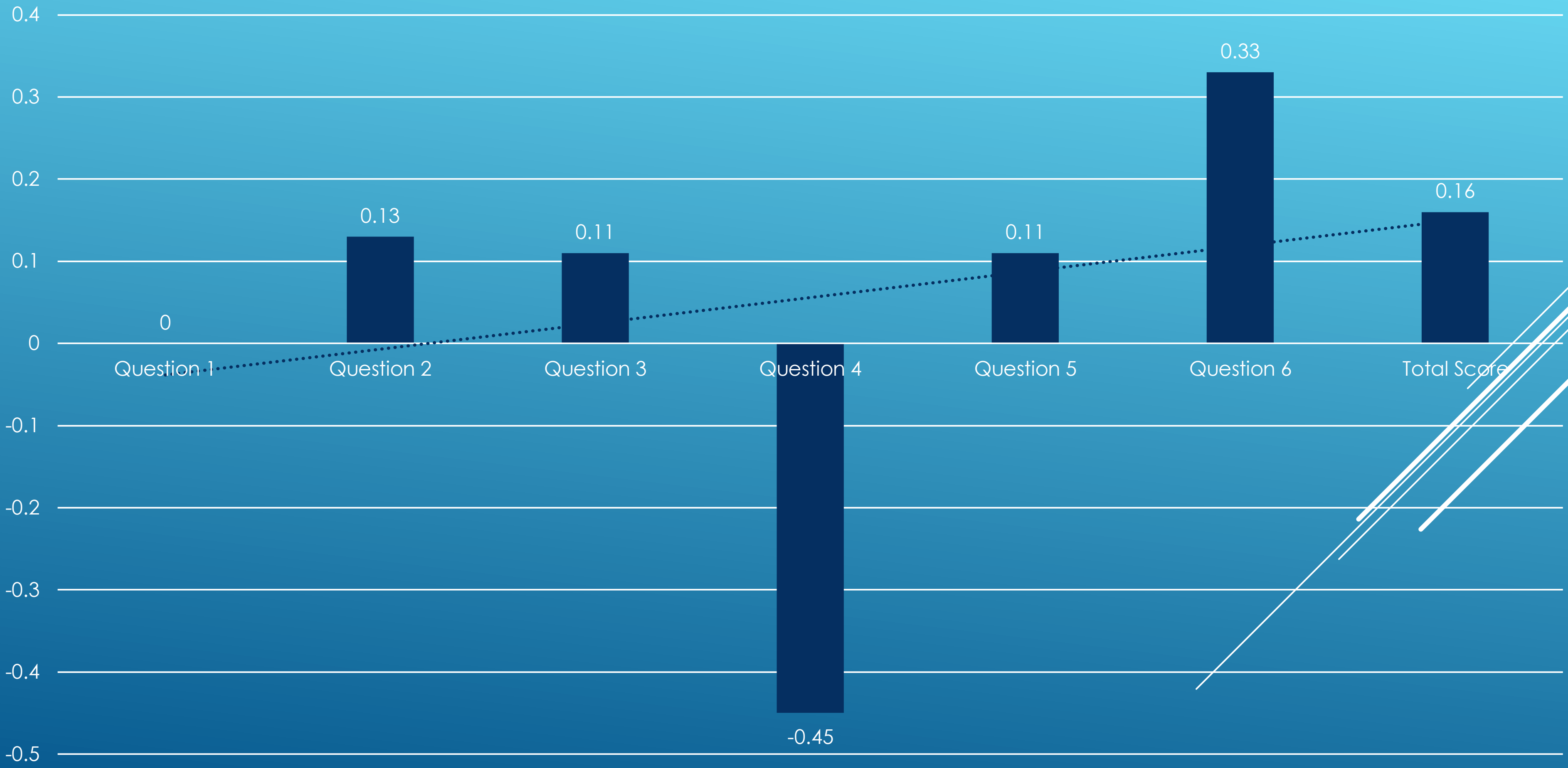
Multiple Choice:	Open Answer:
3	1
4	2
6	5

Data on the left is Y and the right is X

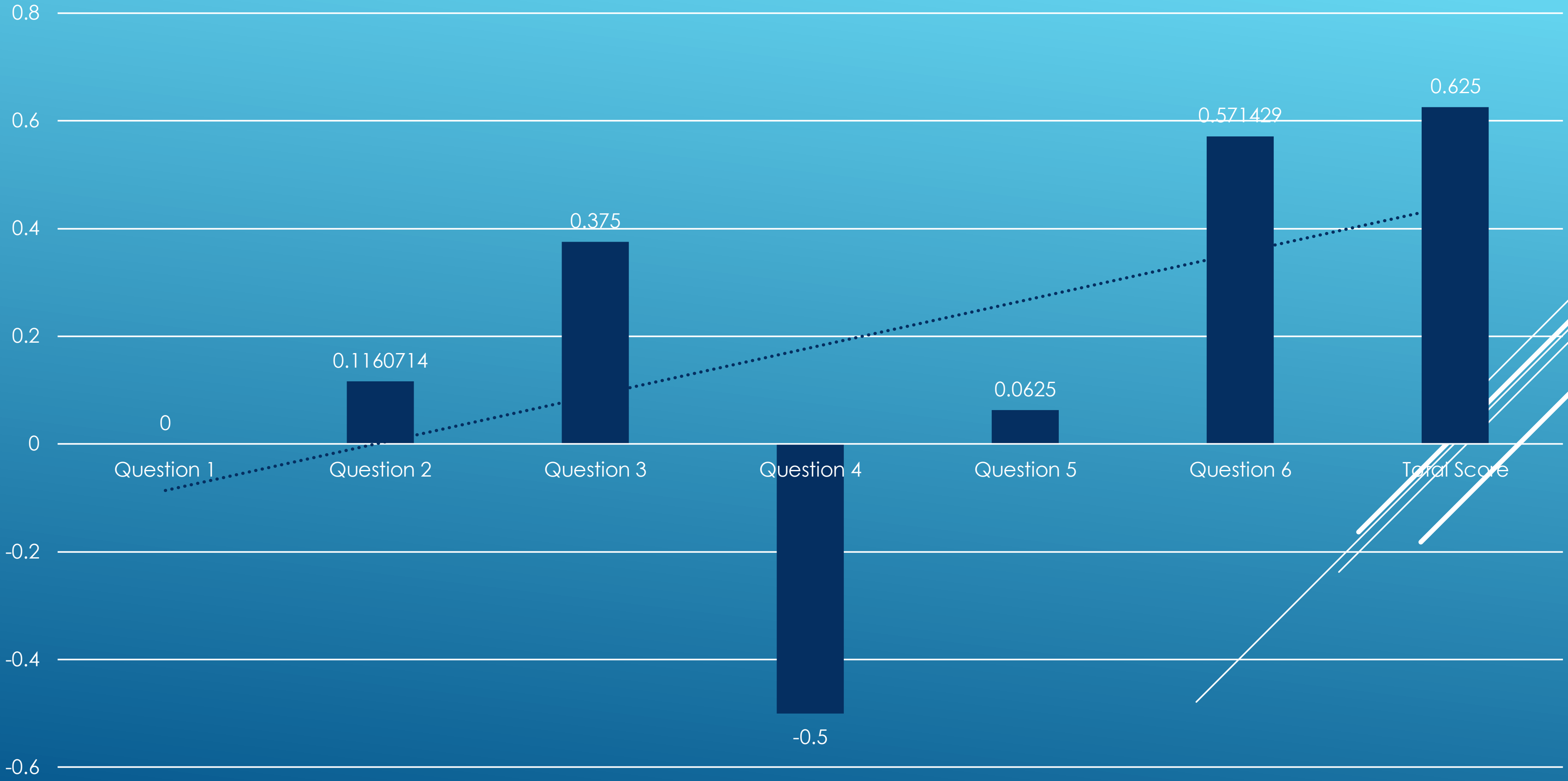
AVERAGE SCORES OF QUESTIONS OUTLIERS REMOVED



DIFFERENCES IN RESULTS X-Y



DIFFERENCES X-Y OUTLIERS REMOVED

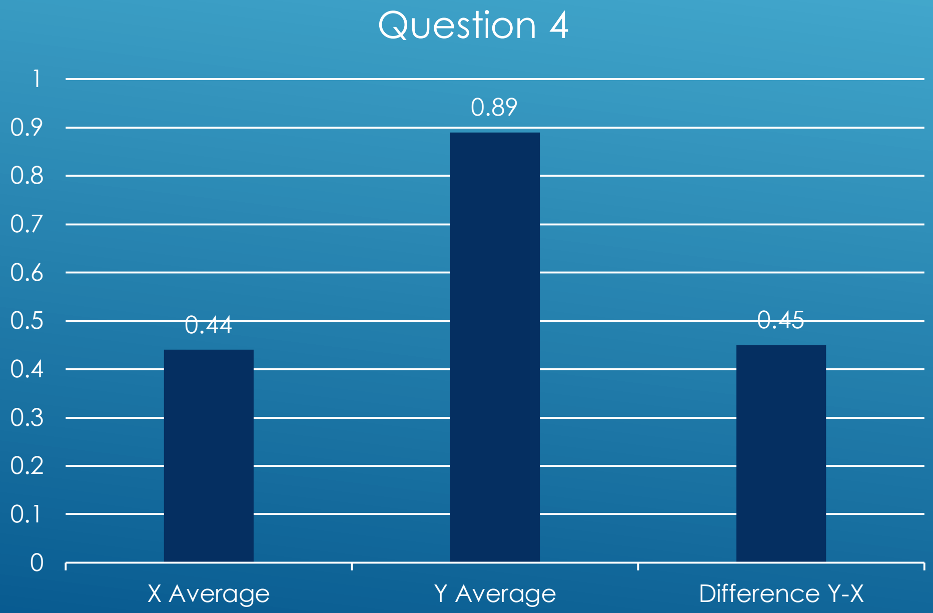


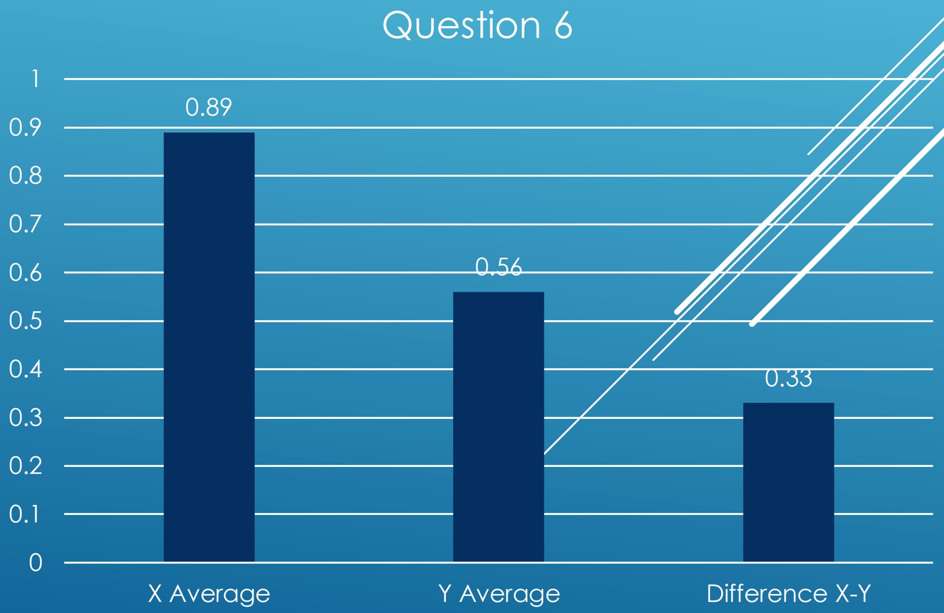
WAS MY HYPOTHESIS CORRECT?

With these results, I can say my hypothesis is incorrect

SOMETHING ELSE?

- Differences in 4 and 6 were significant





CITATIONS

[1]Jan Pedersen Kupiec, Julian and Francine Chen. A trainable document summarizer. ACM SIGIR conference on Research and development in information retrieval, (15):68–73, 1995

[2] Paul Tarau Rada Mihalcea. Textrank: Bringing order into texts. 2011.

[3] Herwig Unger Mario Kubek. Topic detection based on the pagerank's clustering property. 2011.