Summarizing Research Papers for Beginners

UNION COLLEGE

Rex Rubin

Aaron Cass, Advisor

Why a Document Summarizer?

It is difficult to begin in a new field of research, as most data presented to a novice is daunting and confusing. I am hoping to lessen that burden by making an easier to read document summarizer, specifically for research papers.

Document Summarizers

Automated document summarizers are lacking in a lot of fields, namely coherency. Since they are only lines of code, they cannot interpret importance the way humans can, and often return sentences that do not work with the rest of the chosen sentences. There are two different kinds of automated document summarizers, Extraction based and Abstraction based.

Extraction

Extraction lifts sentences from the original text and puts them in the final abstract[1]. It will connect the most important words to each other and return those words in an attempt to have a working abstract.

Abstraction

Abstraction based summarizers try to create their own sentences based on the text given[2]. There main way of making new sentences is by condensing sentences into a few. This method is much more Natural Language based than extraction.

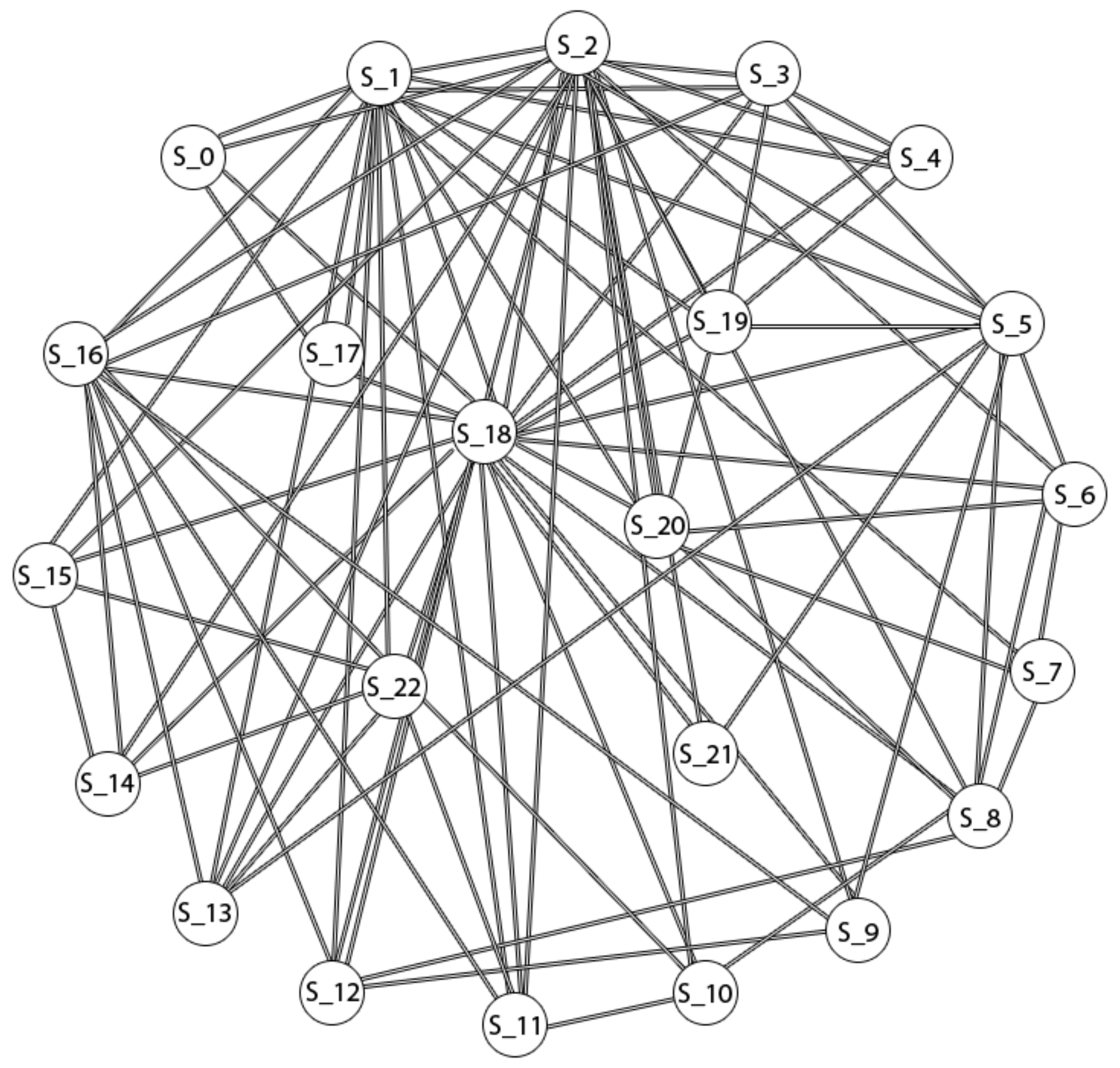
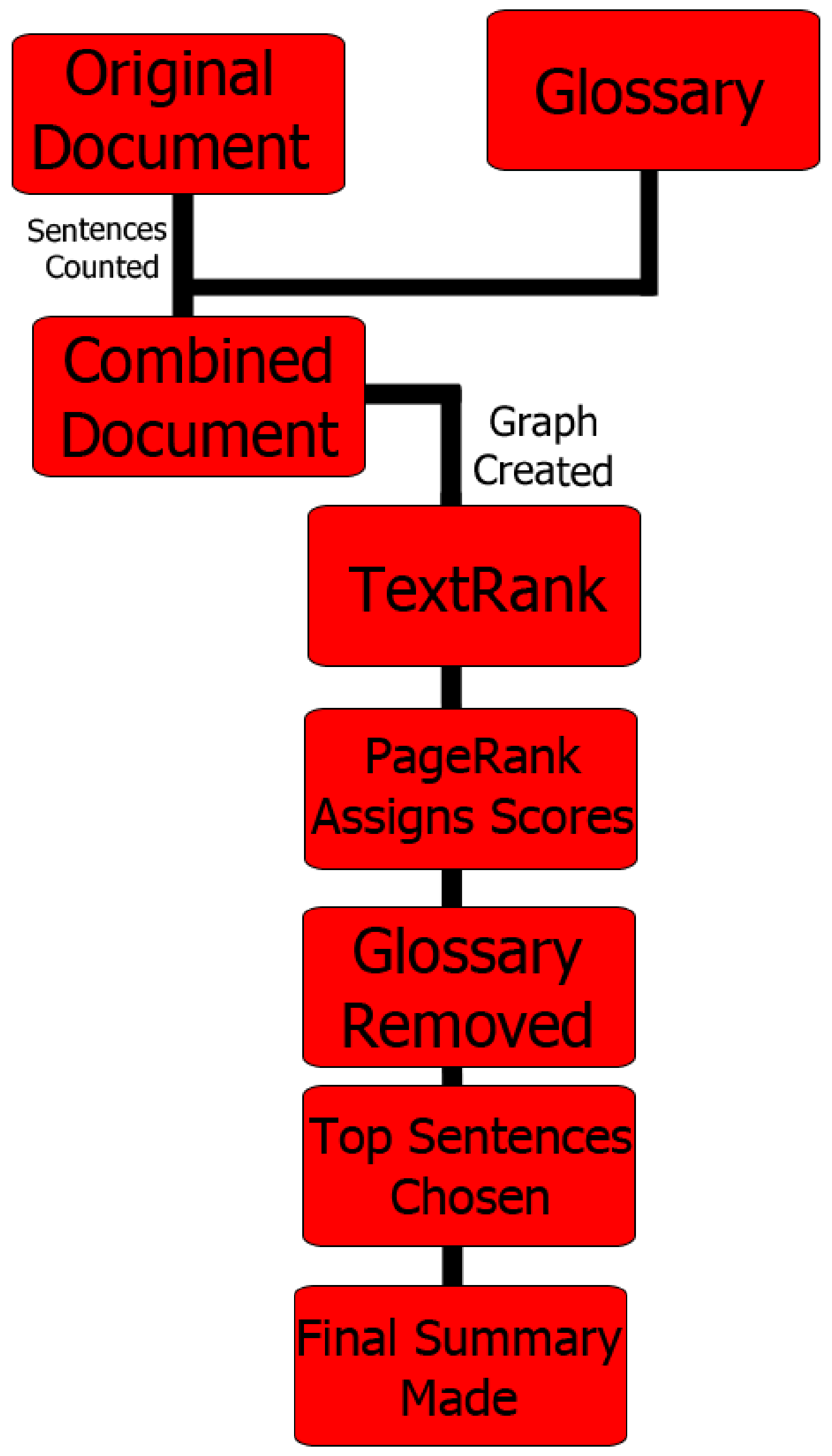


Figure 1. Sample TextRank Graph

TextRank and PageRank

TextRank is an already existing document summarizer that works differently than most other summarizers. It creates a graph of sentences where each sentence is a node[3], and links the words together based on their similarity. It will then take the top n most sentences and put them into the final summary. TextRank works by using PageRank once the graph is built. PageRank will check the probability of each node connecting to a neighbor by crossing over all of them semi-randomly. The nodes with the highest probability to be connected are the final chosen sentences used for the summary.



Original Document

Glossary

Sentences Counted

Combined Document

Graph Created

TextRank

PageRank Assigns Scores

Glossary Removed

Top Sentences Chosen

Final Summary Made

Figure 3. Planned modification to TextRank

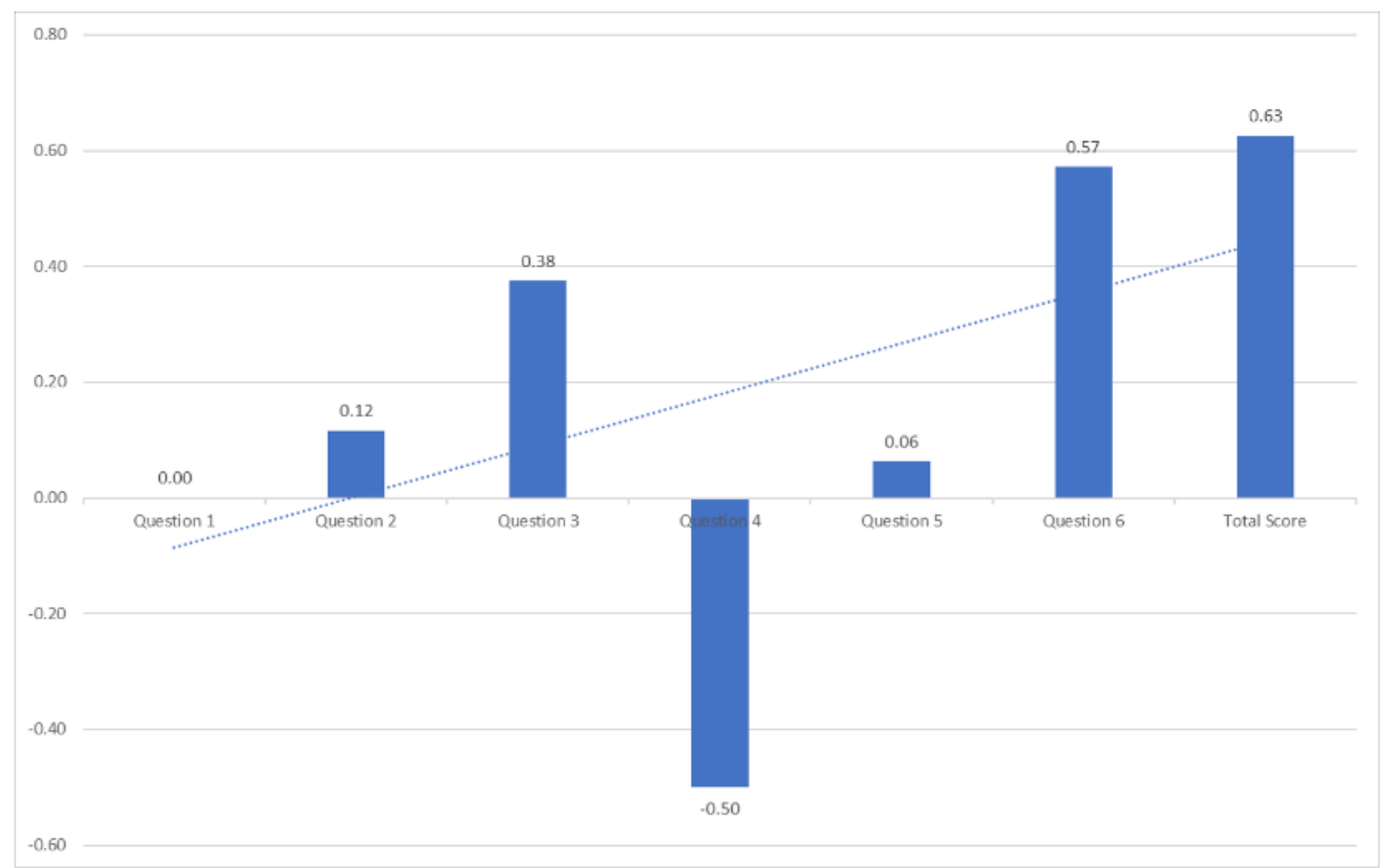


Figure 4. The differences of Test Group results to Control Group results

My Modification

I will be adding a glossary full of words related to CS research to the list of sentences that are put into TextRank in order to try and bring out more cohesive sentences that flow better with the original document. It will take the original document and count the sentences, then have the glossary added to the original document. This will then be turned into a graph to be run through TextRank, which will return the top sentences. Before the top n-most are chosen the sentences from the glossary will be removed so only the sentences from the original document can be in the final summary.

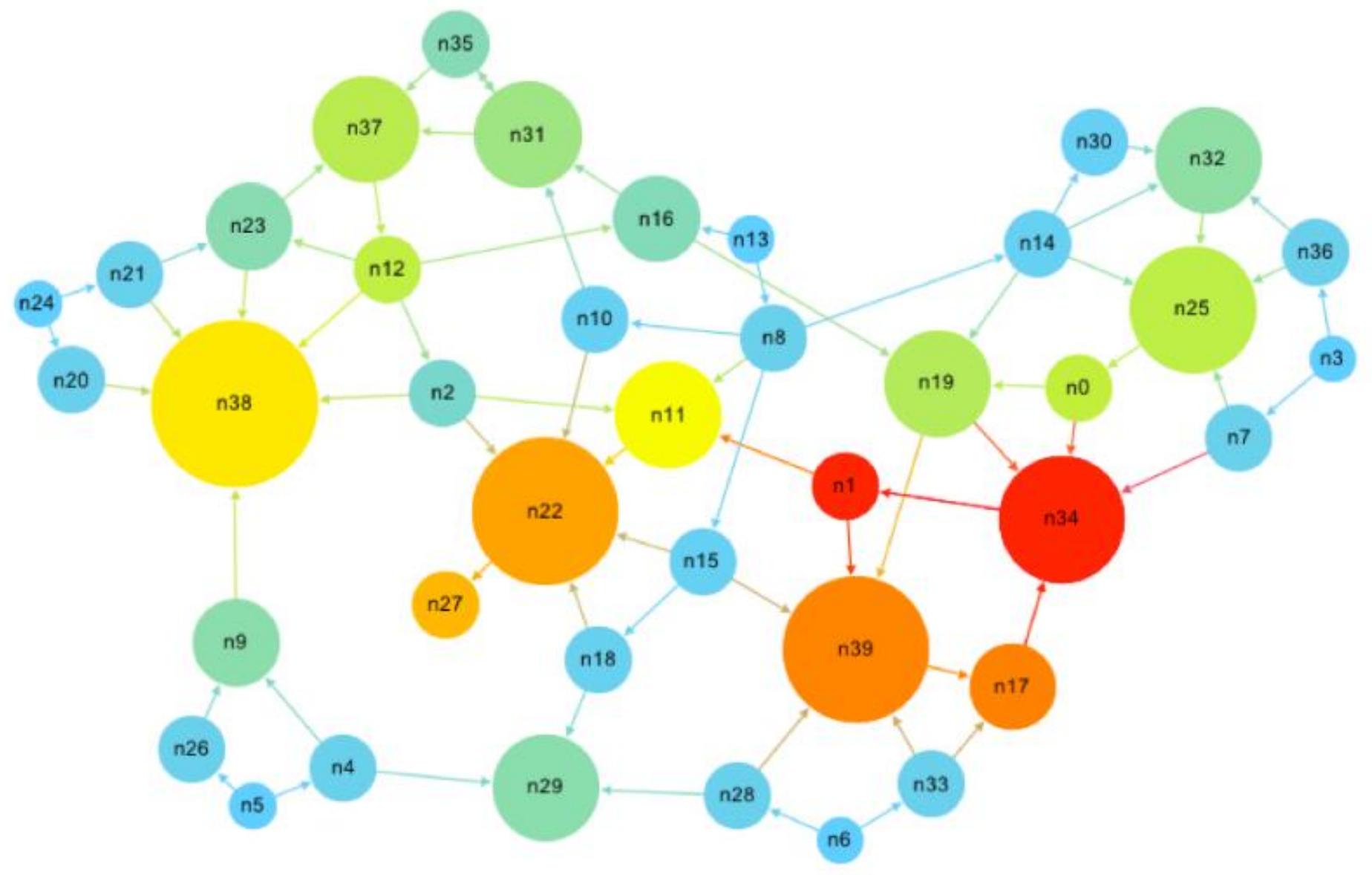


Figure 2. Sample PageRank Graph

Testing

I will be tested this with Union students to see what differences there were between the normal version of TextRank, and my modified version of it. I had a control group read the non-modified version of TextRank's abstract, and a test group reading the modified version's abstract. I Had the participants complete a test of the main points of the original document

Results

My document summarizer did not present all of the main points of the original article, but instead presented a portion of the main points.

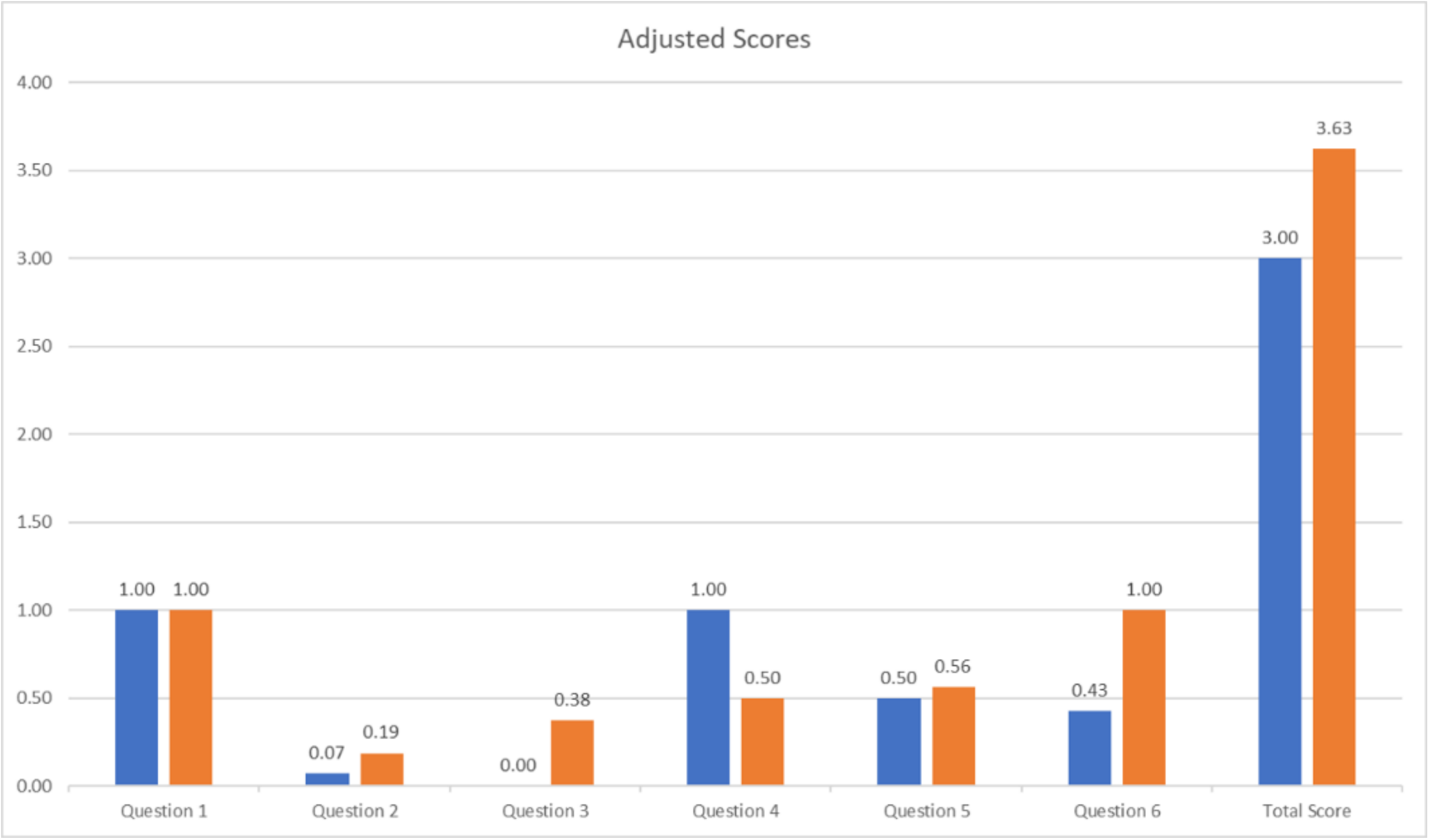


Figure 5. Averages of scores compared
(Blue: Control Group, Orange: test Group)

References

[1] Kupiec, Julian, et al. "A Trainable Document Summarizer." *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '95*, 1995

[2] Khan, Atif, and Naomie Salim. "A REVIEW ON ABSTRACTIVE SUMMARIZATION METHODS ." *Journal of Theoretical and Applied Information Technology*, vol. 59, 10 Jan. 2014.

[3] Mihalcea, Rada, and Paul Tarau. "TextRank: Bringing Order into Texts." *Department of Computer Science University of North Texas*, July 2004.