Topic database creation over the Rice Thresher

Zoe Katz (Head Fellow), Edgar Avalos-Gauna (Mentor), and Sarah Motteler (Researcher)

Introduction

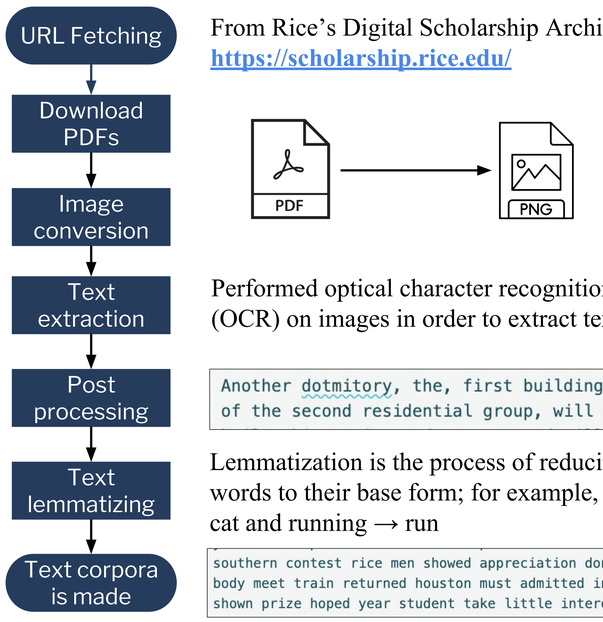
Topic modelling is the identification of important ideas in a large repository of unstructured text. Documents that contain similar topics can be tied together, and the choice of which topics appear in which document can tell us culturally significant subjects for the writers of a document at the time of its writing.

This project aims to create a large database of text from 100 years of editions of the Thresher, Rice University's student-run newspaper. The Thresher has been in publication since 1916, and continues publishing to this day, although the database intends to cover only from 1916-2016.

Textual databases are especially relevant to the field of natural language processing (NLP). Analysis of the topics in the Rice Thresher using NLP could better contribute to our understanding of shifting thoughts and beliefs across the entirety of the 20th century in Houston and beyond.

the Rice Thresher

Methodology



From Rice's Digital Scholarship Archive: https://scholarship.rice.edu/

Performed optical character recognition (OCR) on images in order to extract text

```
Another dotmitory, the, first building
of the second residential group, will be
```

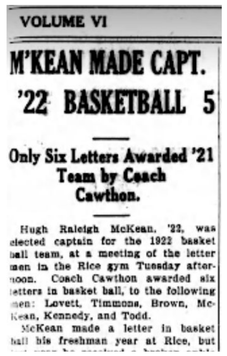
Lemmatization is the process of reducing words to their base form; for example, cats → cat and running → run

```
southern contest rice men showed appreciation done going
body meet train returned houston must admitted interest
shown prize hoped year student take little interest
```

Text Extraction Results

Comparison of Python OCR libraries based on the following factors:

- Ability to recognize words
  - Including spelling and punctuation
- Ability to discern columns and non-linearly formatted text
- Ability to extract numerical dates, roman numerals, and other non-alphabetical groups of characters



VOLUME VI

M'KEAN MADE CAPT. '22 BASKETBALL 5

Only Six Letters Awarded '21 Team by Coach Cawthon.

Hugh Raleigh McKean, '22, was elected captain for the 1922 basket ball team, at a meeting of the letter men in the Rice gym Tuesday afternoon. Coach Cawthon awarded six letters in basket ball, to the following men: Lovett, Timmons, Brown, McKean, Kennedy, and Todd.

McKean made a letter in basket ball his freshman year at Rice, but

PyTesseract (selected)

VOLUME VI

M'KEAN MADE CAPT.
22 BASKETBALL 5

Only Six Letters Awarded °21
Team by Coach
Cawthon.

Hugh Raleigh McKean, '22, was elected captain for the 1922 basket pall team, at a meeting of the letter men in the Rice gym Tuesday afternoon, Coach Cawthon awarded six jetters in basket ball, to the following men: Lovett, Timmons, Brown, McKean, Kennedy, and Todd.

PdfMiner

T^TWTI
THE

VOLUME V!

RICE INSTITUTE HOUSTON, TEXAS,
MARCH 4, 1921

NUMBER 23

Signs of Spring.

MR. CRAM LECTURES
O N ARCHTIECTURE
OF MCE !NST!TCTE

Desires Library,

Academic
Court end Chepe! as Next
Three Building*.

RICE TEAM DEFEATS LS.U.
TEXAS 24 TO 22 GAME
IN EXCITING GAME F N E - 35

TAKES HARD
FROM RICE
TO 23

Last Conference Game the
Hardest and Prettiest of
Season.

ExceHent Teamwork a nd
Good Guarding Feature
the Came.

M'KEAN HADE MPT.
'22 BASKETBALL 5

PyPDF2

T^TWTI
THE VOLUME V! RICE
INSTITUTE HOUSTON, TEXAS,
MARCH 4, 1921 NUMBER 23
M'KEAN HADE MPT.
'22 BASKETBALL 5

On!y Six Letter# Awarded '21
Teem by Ceach Cawthon.
Hugh Rateigh McKean. '22. was
aiected captain for the 1922 basket
inH team, at a meeting of the tetter
men in the Rice gym Tuesday after-
toon. Coach Cawthon awarded six
etters in basket baii, to the foilowing
nen: Lovett, Timmons, Brown. Mc-
Kean, Kennedy, and Todd.
McKean made a tetter in basket
hat

PyMuPDF

T^TWTI
THE
VOLUME V!
RICE INSTITUTE HOUSTON, TEXAS,
MARCH 4, 1921
NUMBER 23
M'KEAN HADE MPT.
'22 BASKETBALL 5
On!y Six Letter# Awarded '21
Teem by Ceach Cawthon. Hugh Rateigh
McKean. '22. was aiected captain for the
1922 basket inH team, at a meeting of
the tetter men in the Rice gym Tuesday
after-toon.
Coach Cawthon awarded six
etters in basket baii, to the foilowing
nen: Lovett, Timmons, Brown. Mc-
Kean, Kennedy, and Todd.
McKean made a tetter in basket hat

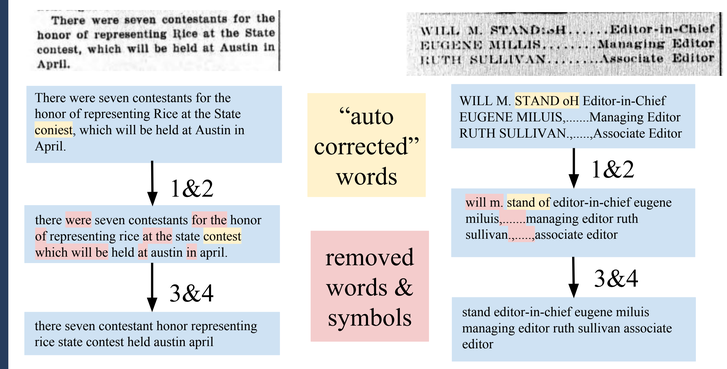
Limitations

Extracting text from the Threshers ended up being more time-consuming than expected. A 10 page Thresher took ~30 seconds to convert from a PDF to a folder of images. Extracting the text was even more costly, since one Thresher took about ~2 minutes to complete. Adding on another ~45 seconds for post-processing work, this means ~3.2 minutes needed to be allocated per Thresher, to get from the beginning to the end of the flow chart. There are 3088 Threshers from 1916-2016, meaning the estimated completion of the database will take a total of around 165 hours of running code.

Post Processing Results

Steps of post processing and lemmatization:

1. Autocorrect for words that the OCR incorrectly extracted
2. Standardizing font to lowercase
3. Lemmatization, in order to standardize the data
4. Removing extraneous non-alphabetical characters



There were seven contestants for the honor of representing Rice at the State contest, which will be held at Austin in April.

There were seven contestants for the honor of representing Rice at the State coniest, which will be held at Austin in April.

1&2

there were seven contestants for the honor of representing rice at the state contest which will be held at austin in april.

3&4

there seven contestant honor representing rice state contest held austin april

"auto corrected" words

removed words & symbols

WILL M. STAND:.oH.....Editor-in-Chief
EUGENE MILLIS.....Managing Editor
RUTH SULLIVAN.....Associate Editor

WILL M. STAND oH Editor-in-Chief
EUGENE MILUIS,.....Managing Editor
RUTH SULLIVAN,.....,Associate Editor

1&2

will m. stand of editor-in-chief eugene miluis,.....managing editor ruth sullivan,.....,associate editor

3&4

stand editor-in-chief eugene miluis managing editor ruth sullivan associate editor

Cons of post processing and lemmatization:

- Lemmatization can remove semantically relevant words like the name "Will"
- Autocorrect does not always successfully reconstruct the original word

Conclusion

Large text corpora are useful to analyze trends over variables like time and location. In data analysis and machine learning, fields like social media analysis, market research, and text mining make use of unstructured data. Large text corpora are not widely available, and are thus crucial to create. Therefore, building a textual corpora of Rice Threshers could be helpful for future generations of data scientists. Through the use of topic vectors and sentiment analysis, this database can aid in future analysis regarding the evolution of key events and ideas during the past 100 years and how they relate to Rice University, the Houston area, and the U.S. as a whole.

Acknowledgements

Thank you to the Fondren Fellows program, and my mentor Edgar Avalos-Gauna.

Fondren Fellows