# An Introduction to Article Separation

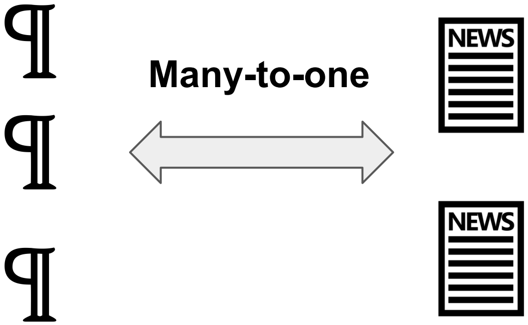
By Cole Juracek

# Agenda

- Background + problem statement
- Optical Character Recognition (OCR)
- Title segmentation
- Term frequency - inverse document frequency (TF-IDF)
- Hierarchical clustering
- Results

# Problem Statement

- *The Chronicle* is Duke’s independent, student-run newspaper founded in 1905
- Collection of historic *Chronicle* articles recently digitized
- **Goal**: Determine *which* text belongs to *which* article
    - ER: Entities are articles



# Optical Character Recognition (OCR)

- **Goal:** Convert image of text into raw text itself
- Performed with Tesseract
- Captures text at the paragraph level
- Bounding boxes denote confidence in character reading
    - Not a computer vision problem - performance assumed to be adequate

# The TRINITY CHRONICLE.

VOL. 1. No. 1. TRINITY COLLEGE, DURHAM, N. C., DECEMBER 19, 1905. PRICE 5 CENTS.

## HESPERIAN VS. COLUMBIAN.

### Sixteenth Annual Inter-Society Debate ---Won By the Hesperian.

A great debate!

When the historian shall go to write the record of the debates between the Columbian and Hesperian Literary Societies there will be no brighter page in all the annals than that of the debate for the year 1905-06, which was held in the Craven Memorial Hall, Saturday evening, December 16, 1905.

For the past half decade each one of these contests has been marked as being of a high order, the question being handled with the skill of trained speakers, with cogent logic, and the speakers presenting their arguments in smooth and easy-flowing language. The debate Saturday evening was up to, if not better, than the standard that had already been set in the years past. The subject debated was not an over-argued one that had lost more favorable circumstances.

The affirmative side was upheld by the Columbians, with C. E. Phillips, of Salisbury, and Hersey E. Spence, of South Mills, as representatives; the negative was defended by the Hesperians, with A. L. Wissburg, of Durham, and Holland Holton, of Durham, as speakers. On the first speeches each debater was allowed sixteen minutes, and on the rejoinder, the first man on each side had six, and the last one, seven. This amount gave each speaker ample time to set forth his argument, and it was also not enough to cause the audience to become wearied.

THE SPEAKERS.

The affirmative was opened by Mr. Phillips who spoke in part as follows:

He began by showing what a great part life insurance plays in the commercial and industrial life of the nation: how it increases national credit as well as individual integrity. He said that out of 80,000,000 of people more than 20,000,000 carry life more favorable circumstances. Then he furiously attacked the 20-year endowment and all deferred dividend policies, saying that they were instituted by the insurance officials in order to fill their treasuries quickly so they could begin their systems of graft.

In the climax, the speaker, in a forceful manner, brought forth a number of serious charges against the insurance companies, and, emphasizing the fact that state supervision had by no means met the requirements, he claimed that the Federal government should be given power to assume uniform control over American Life Insurance companies.

Following him came Mr. Wissburg in behalf of the negative. A synopsis of his speech follows:

The purpose of State supervision is three fold: 1. To see that the insurance laws of the State are obeyed. 2. To see that the policy-holders receive equitable treatment from the insurance corporations. 3. To see that the in-

## MR. D. A. TOMPKINS, OF CHARLOTTE.

### Prominent N. C. Business Man Speaks to Trinity Students.

The first lecture arranged for by the committee of public speaking for this year was delivered Friday evening in Craven Memorial Hall by Mr. D. A. Tompkins, of Charlotte, N. C. Owing to the very inclement weather there was only a small audience to hear his address. Especially were the students noticeable in their absence. The address was one of the best that has been heard here lately in that it was so very practical and plain.

The speaker is one of the State's best business men and every utterance of his was backed by the life which he has himself lived. He is one of the proprietors of the Charlotte Observer, a mill man of great repute, and has recently been appointed as one of the directors of the Equitable Life Assurance Society. He is also a member of the United States Industrial Commission

times even stop his practice to take a course under some special master to bring his education up to date. So with the locomotive engineer, the train hand, and others in responsible engineering or other industrial positions.

It rarely happens that a common school or a high school or even a college education properly equips a young man for a particular line of work. Therefore special schools and other facilities for special study and instruction are coming to be one of the most important factors in modern education.

Every education as received in the public schools and colleges is incomplete. Practical life requires in addition to general education some specific supplemental instruction to adopt a general education to the requirements of the particular work in hand. This can generally only be done in special day schools or night-schools which do not undertake to give a general education. There are many who do not ap-

# Title Segmentation

- Paragraphs contain content at varying levels of granularity
  - Meta
  - Separator
  - Ttitle
  - Text
  - Publicity (not part of article)
- Titles are either article names (separate entities) or sub-titles (same entity)
- **Goal:** Combine contiguous text delimited by titles

# Term Frequency - Inverse Document Frequency (TF-IDF)

- **Goal:** Cluster several article portions into a single article
- **TF-IDF:** Method for representing documents as vectors

$$TF(t,d) = N(t,d)$$

### *Term Frequency*

- Frequency of word in document
- High weight to frequently occurring words

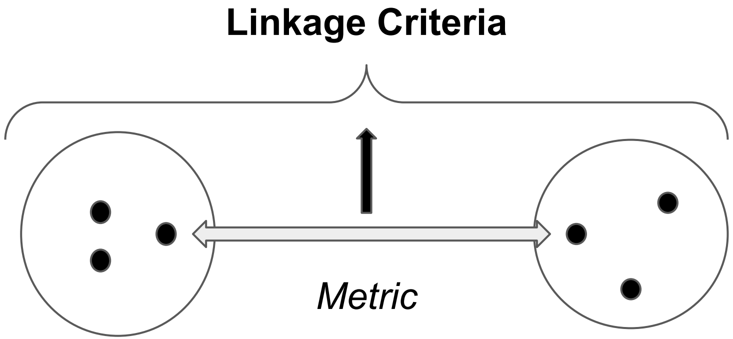
$$IDF(t,D) = log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)$$

### *Inverse Document Frequency*

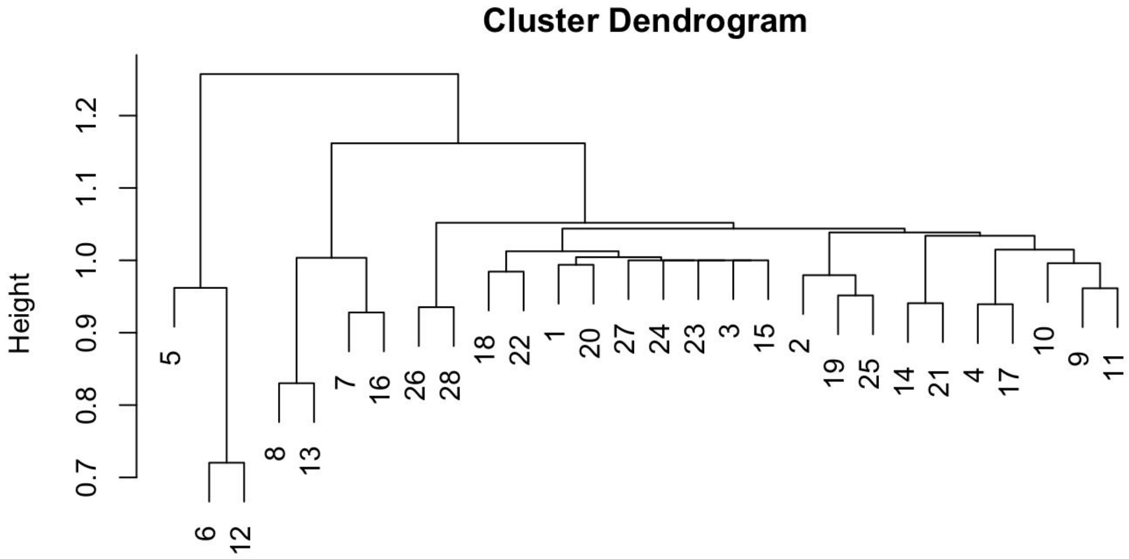
- Ratio of total documents to documents containing word
- High weight to “rare” words in a document

# Hierarchical Clustering

- Successively merge observations/clusters until one remains
- 2 components
  - Metric: Determine how similar (different) two observations are from each other
  - Linkage criteria: Determines how metric is used when comparing clusters with multiple observations



# Hierarchical Clustering

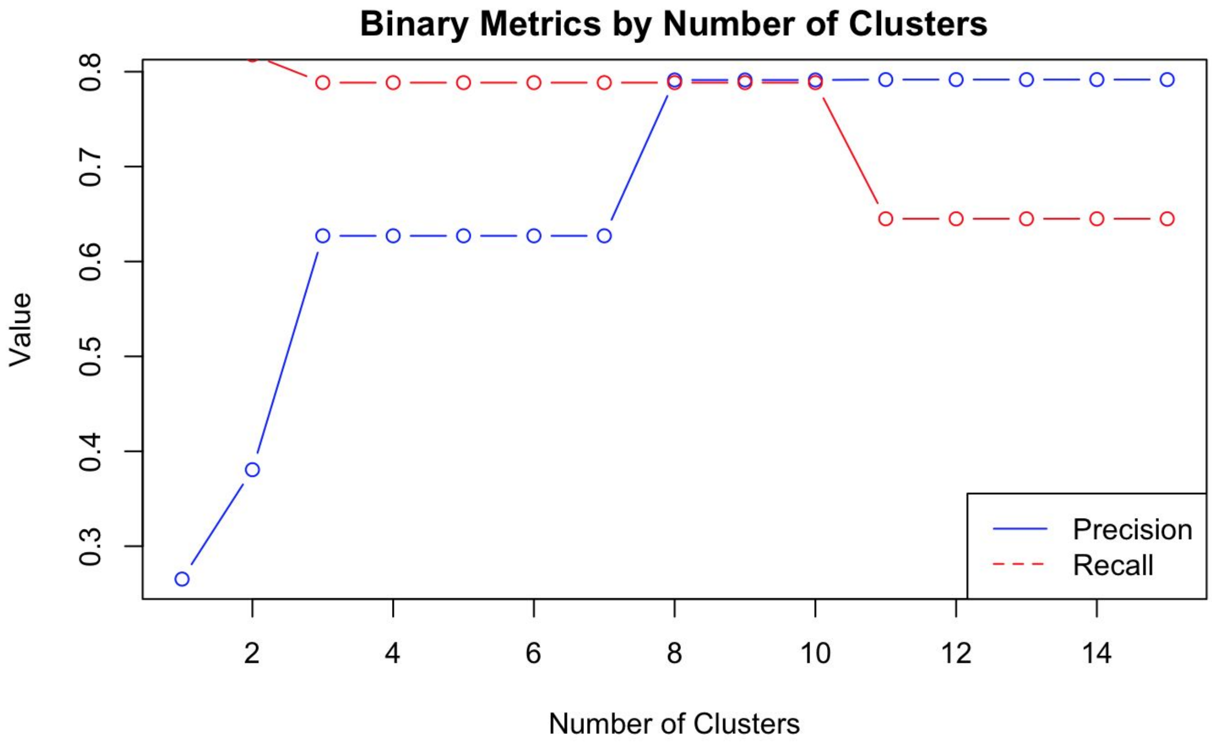


- Easily visualized as a *dendrogram*
- Vertical distance defines distance between observations
- Obtain different number of clusters via horizontal cut across dendrogram

# Workflow / Pipeline

1. Download *Chronicle* articles as JPEG files
2. Perform OCR with Tesseract to extract text paragraphs
3. Utilize title segmentation to identify (sub)-titles and combine appropriate paragraphs into article portions
4. Convert article portions into their respective TF-IDF vectors
5. Use hierarchical clustering on TF-IDF vectors to group article portions together into full articles

# Results


Binary Metrics by Number of Clusters
Value
0.3
0.4
0.5
0.6
0.7
0.8
Number of Clusters
2
4
6
8
10
12
14
Precision
Recall

# Conclusion

- Successful baseline implementation of article separation on historic *Chronicle* articles
- **Implication:** Recover individual articles in a system even *after* being digitally scanned into a system
- **Future work**
  - Consider taking further advantage of the structure of documents
  - More robust to identifying publicities, currently ignored

# Questions?