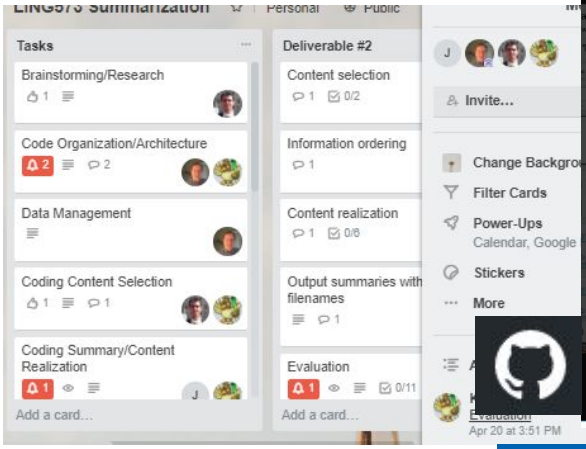
Summarization - Round 1

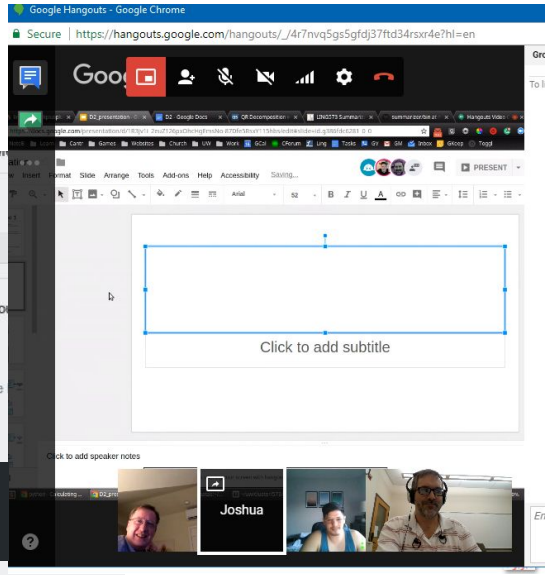
Team E2JK++

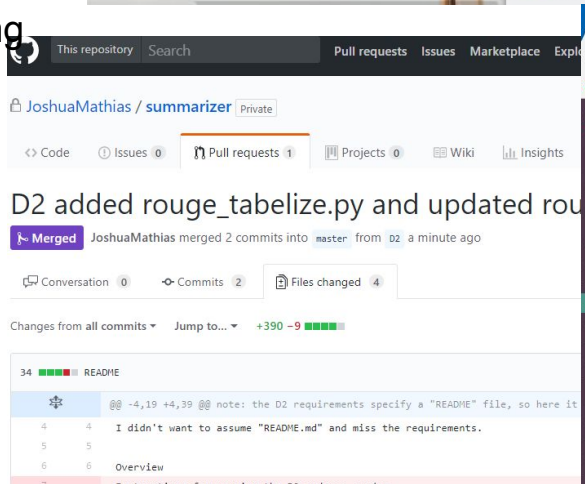
Eric Lindberg
John Greve
Josh Mathias
Kekoa Riggin

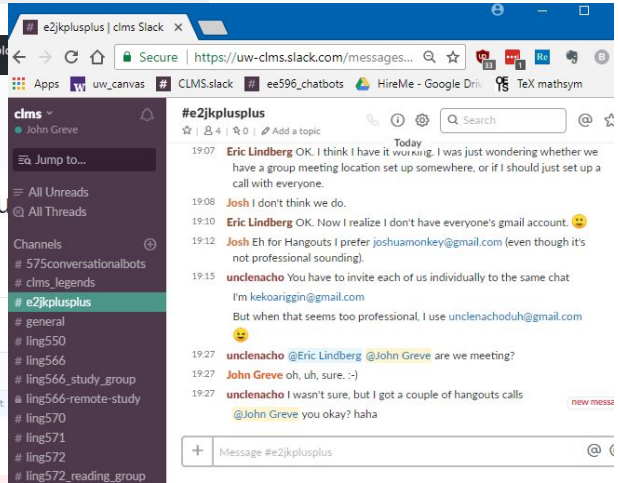
Communication

- Trello
  - Tasks, assignments
  - Requirements
- Slack
  - Chat
- Hangouts
  - Video calls
- Git
  - Pull requests
- Bi-weekly standups
  - In-person reporting, planning

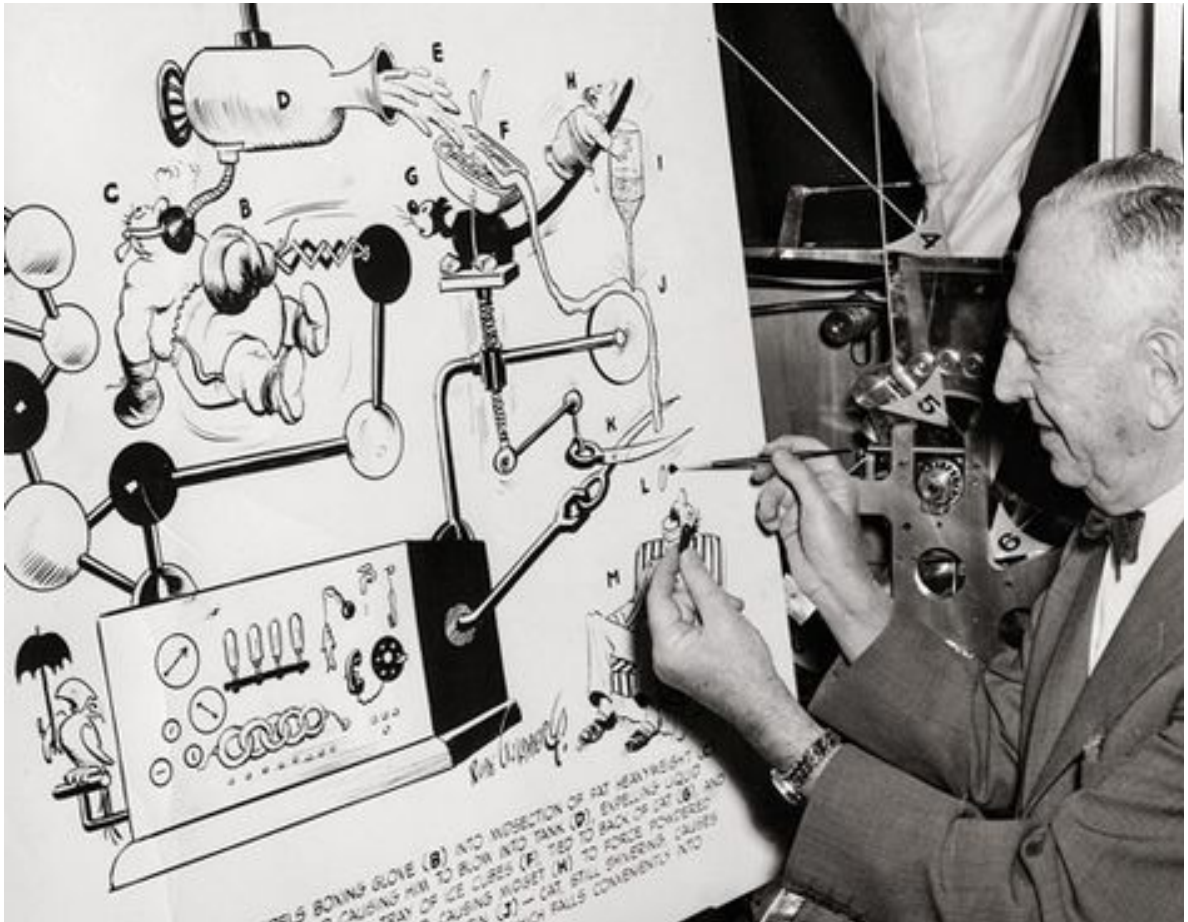








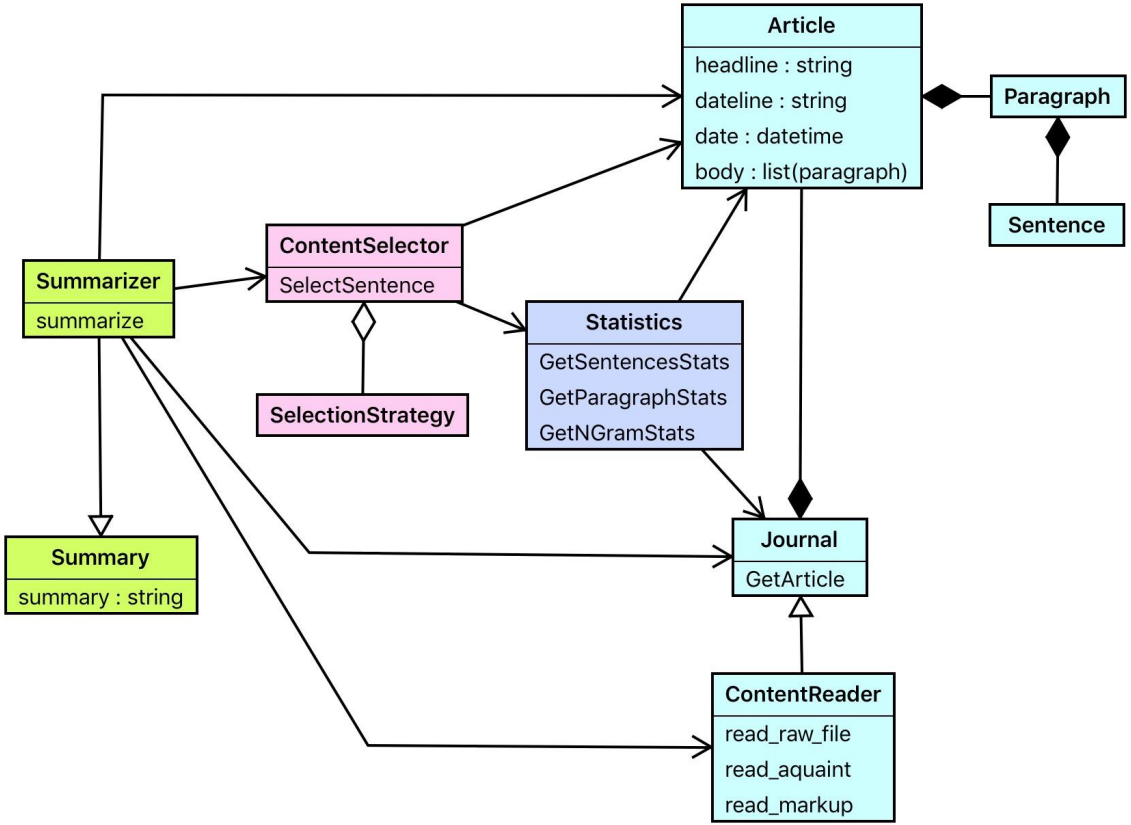
System Architecture



System Architecture

System Architecture Diagram
(the right one this time)

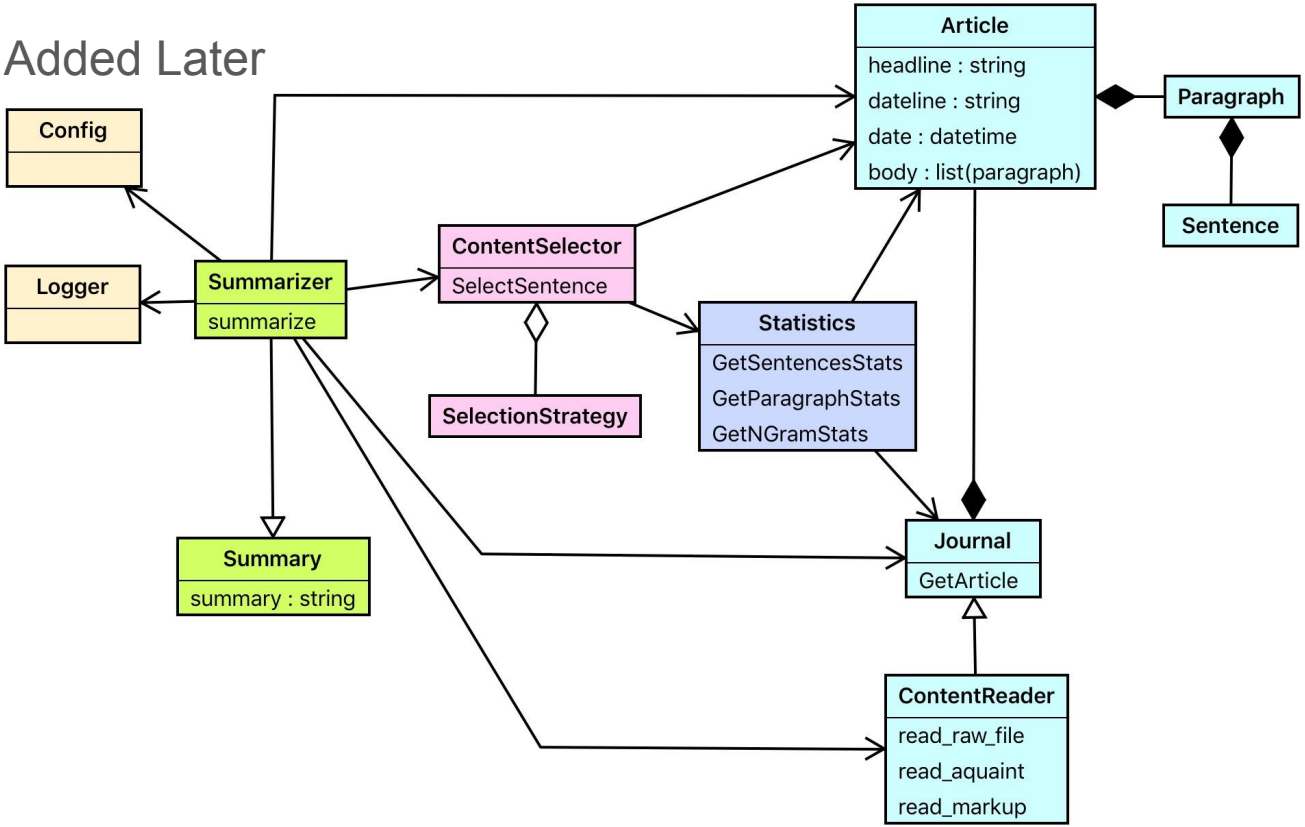
- Summarization
  Invokes reading of content and summary sentence selection
- Content Access
  Reads and iterates over articles for each document set
- Sentence Selection
  Extracts/Creates and Orders summary sentences based on selected articles.
- Statistics
  Provides details for sentence selection and for logging.



System Architecture

Important Components Added Later

- Configuration
- Logging



Content Selection

First Sentence Summarization

There are broadly two types of extractive summarization tasks depending on what the summarization program focuses on. The first is *generic summarization*, which focuses on obtaining a generic summary or abstract of the collection (whether documents, or sets of images, or videos, news stories etc.). The second is *query relevant summarization*, sometimes called *query-based summarization*, which summarizes objects specific to a query. Summarization systems are able to create both query relevant text summaries and generic machine-generated summaries depending on what the user needs.

An example of a summarization problem is document summarization, which attempts to automatically produce an abstract from a given document. Sometimes one might be interested in generating a summary from a single source document, while others can use multiple source documents (for example, a cluster of articles on the same topic). This problem is called multi-document summarization. A related application is summarizing news articles. Imagine a system, which automatically pulls together news articles on a given topic (from the web), and concisely represents the latest news as a summary.

Image collection summarization is another application example of automatic summarization. It consists in selecting a representative set of images from a larger set of images.[1] A summary in this context is useful to show the most representative images of results in an image collection exploration system. Video summarization is a related domain, where the system automatically creates a trailer of a long video. This also has applications in consumer or personal videos, where one might want to skip the boring or repetitive actions. Similarly, in surveillance videos, one would want to extract important and suspicious activity, while ignoring all the boring and redundant frames captured.

Pivoted QR Matrix Decomposition

There are broadly two types of extractive summarization tasks depending on what the summarization program focuses on. The first is *generic summarization*, which focuses on obtaining a generic summary or abstract of the collection (whether documents, or sets of images, or videos, news stories etc.). The second is *query relevant summarization*, sometimes called *query-based summarization*, which summarizes objects specific to a query. Summarization systems are able to create both query relevant text summaries and generic machine-generated summaries depending on what the user needs.

An example of a summarization problem is document summarization, which attempts to automatically produce an abstract from a given document. Sometimes one might be interested in generating a summary from a single source document, while others can use multiple source documents (for example, a cluster of articles on the same topic). This problem is called multi-document summarization. A related application is summarizing news articles. Imagine a system, which automatically pulls together news articles on a given topic (from the web), and concisely represents the latest news as a summary.

Image collection summarization is another application example of automatic summarization. It consists in selecting a representative set of images from a larger set of images.[1] A summary in this context is useful to show the most representative images of results in an image collection exploration system. Video summarization is a related domain, where the system automatically creates a trailer of a long video. This also has applications in consumer or personal videos, where one might want to skip the boring or repetitive actions. Similarly, in surveillance videos, one would want to extract important and suspicious activity, while ignoring all the boring and redundant frames captured.

Conroy and O'Leary [2001]

Information Ordering

First Sentence Summarization

Pivoted QR Matrix Decomposition

There are broadly two types of extractive summarization tasks depending on what the summarization program focuses on. The first is generic summarization, which focuses on obtaining a generic summary or abstract of the collection (whether documents, or sets of images, or videos, news stories etc.). The second is query relevant summarization, sometimes called query-based summarization, which summarizes objects specific to a query. Summarization systems are able to create both query relevant text summaries and generic machine-generated summaries depending on what the user needs.

An example of a summarization problem is document summarization, which attempts to automatically produce an abstract from a given document. Sometimes one might be interested in generating a summary from a single source document, while others can use multiple source documents (for example, a cluster of articles on the same topic). This problem is called multi-document summarization. A related application is summarizing news articles. Imagine a system, which automatically pulls together news articles on a given topic (from the web), and concisely represents the latest news as a summary.

Image collection summarization is another application example of automatic summarization. It consists in selecting a representative set of images from a larger set of images.[1] A summary in this context is useful to show the most representative images of results in an image collection exploration system. Video summarization is a related domain, where the system automatically creates a trailer of a long video. This also has applications in consumer or personal videos, where one might want to skip the boring or repetitive actions. Similarly, in surveillance videos, one would want to extract important and suspicious activity, while ignoring all the boring and redundant frames captured.

There are broadly two types of extractive summarization tasks depending on what the summarization program focuses on. The first is generic summarization, which focuses on obtaining a generic summary or abstract of the collection (whether documents, or sets of images, or videos, news stories etc.). The second is query relevant summarization, sometimes called query-based summarization, which summarizes objects specific to a query. Summarization systems are able to create both query relevant text summaries and generic machine-generated summaries depending on what the user needs.

An example of a summarization problem is document summarization, which attempts to automatically produce an abstract from a given document. Sometimes one might be interested in generating a summary from a single source document, while others can use multiple source documents (for example, a cluster of articles on the same topic). This problem is called multi-document summarization. A related application is summarizing news articles. Imagine a system, which automatically pulls together news articles on a given topic (from the web), and concisely represents the latest news as a summary.

Image collection summarization is another application example of automatic summarization. It consists in selecting a representative set of images from a larger set of images.[1] A summary in this context is useful to show the most representative images of results in an image collection exploration system. Video summarization is a related domain, where the system automatically creates a trailer of a long video. This also has applications in consumer or personal videos, where one might want to skip the boring or repetitive actions. Similarly, in surveillance videos, one would want to extract important and suspicious activity, while ignoring all the boring and redundant frames captured.

Content Realization

Sentences are written as-is to file with no modification.

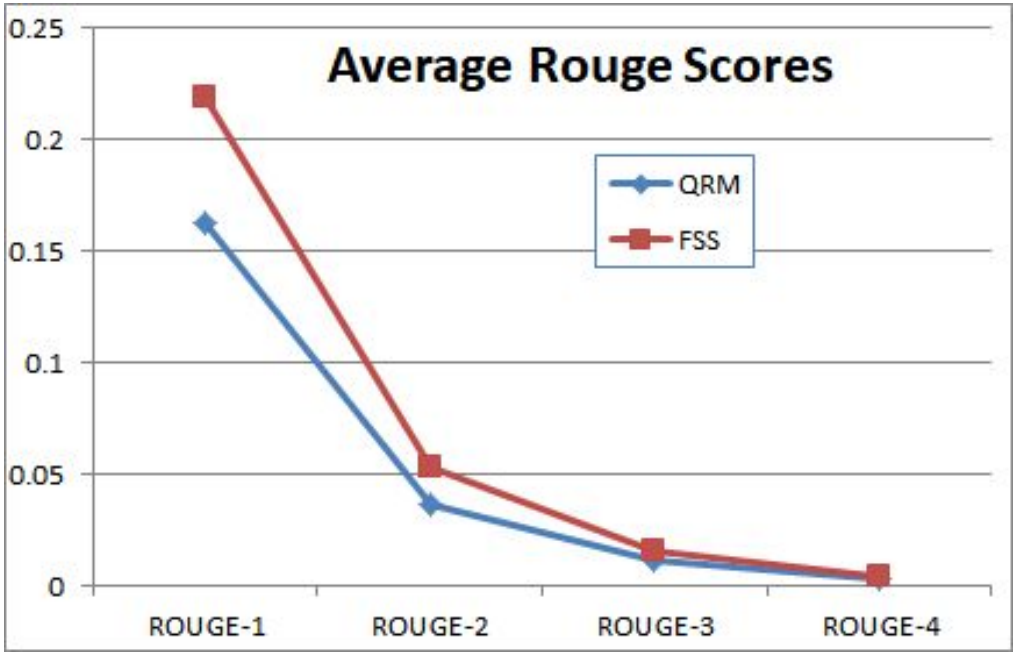
Issues and Successes

- **Know Your Data!**
  AQUAINT/AQUAINT2 formatting issues up to the last minute
  Still have issues to resolve... this is in many ways an engineering effort to finish an end-to-end pass.
  Far more coding / infrastructure than actual linguistics effort.
- **Unit Tests and Functional Tests Helpful**
- **Data Results - ongoing profiling.**

```
runtime anomaly excerpt:
 17 D1003A-A :  Giant Panda
 18 docset=DocumentSet(): topic_id="D1003A" #docs=10
 19    docset.id      ="D1003A-A"
 20    docset.topic_id="D1003A"
 21 D1004A-A :  Papua Tsunami
 22 docset=DocumentSet(): topic_id="D1004A" #docs=10
 23 WARNING: empty ArticleContent(): headline="Wall of water crashes " (#2 docset=DocumentSet(): topic_id="D1004A" #docs=10)
 24 WARNING: empty ArticleContent(): headline="At least 70 dead after" (#3 docset=DocumentSet(): topic_id="D1004A" #docs=10)
 25 WARNING: empty ArticleContent(): headline="Wall of water crashes " (#4 docset=DocumentSet(): topic_id="D1004A" #docs=10)
 26 WARNING: empty ArticleContent(): headline="URGENT"                (#5 docset=DocumentSet(): topic_id="D1004A" #docs=10)
 27 WARNING: empty ArticleContent(): headline="PORT MORESBY: are foun" (#6 docset=DocumentSet(): topic_id="D1004A" #docs=10)
 28 WARNING: empty ArticleContent(): headline="URGENT"                (#7 docset=DocumentSet(): topic_id="D1004A" #docs=10)
 29    docset.id      ="D1004A-A
```

Issues and Successes (cntd)

- **Initial baseline established.**
- **Empty articles undoubtedly hurt extractive summary ROUGE scores.**



Thank you!

Related Reading

Left Brain Right Brain Multi-Document Summarization
John M. Conroy, Judith D. Schlesinger [2004]

Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition
John Conroy and Dian P. O'Leary [2001]

Beautiful Soup Documentation https://www.crummy.com/software/BeautifulSoup/bs4/doc/