# LING 575: Summarization Deliverable #1 Report

**Ben Cote**
University of Washington
bpc23@uw.edu

**Mohamed Elkamhawy**
University of Washington
mohame@uw.edu

**Karl Haraldsson**
University of Washington
kharalds@uw.edu

**Alyssa Vecht**
University of Washington
avecht@uw.edu

**Josh Warzecha**
University of Washington
jmwar73@uw.edu

## Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam vel erat vel mi efficitur faucibus. Nulla consectetur aliquet dolor, sit amet vulputate quam lobortis lacinia. Integer in velit dolor. In pellentesque, ligula sed ornare viverra, enim mi commodo neque, eu ultrices odio sapien nec diam. Fusce tempor aliquam nunc, nec placerat odio venenatis sit amet. Vivamus eget egestas libero, eget porta arcu. Phasellus imperdiet lobortis facilisis. Etiam sit amet nisi quis risus gravida luctus quis nec magna. Etiam velit purus, tempor ac ex eu, vulputate ultrices turpis. Donec blandit tempus placerat. Nullam commodo felis sit amet risus pulvinar fringilla. Praesent convallis, magna non vestibulum tincidunt, tortor purus fermentum elit, at scelerisque augue leo vitae turpis. Sed a iaculis lacus. Vivamus vulputate convallis lorem.

## 1 Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam vel erat vel mi efficitur faucibus. Nulla consectetur aliquet dolor, sit amet vulputate quam lobortis lacinia. Integer in velit dolor. In pellentesque, ligula sed ornare viverra, enim mi commodo neque, eu ultrices odio sapien nec diam. Fusce tempor aliquam nunc, nec placerat odio venenatis sit amet. Vivamus eget egestas libero, eget porta arcu.

## 2 System Overview

Phasellus imperdiet lobortis facilisis. Etiam sit amet nisi quis risus gravida luctus quis nec magna. Etiam velit purus, tempor ac ex eu, vulputate ultrices turpis. Donec blandit tempus placerat. Nullam commodo felis sit amet risus pulvinar fringilla. Praesent convallis, magna non vestibulum tincidunt, tortor purus fermentum elit, at scelerisque augue leo vitae turpis. Sed a iaculis lacus. Vivamus vulputate convallis lorem.

## 3 Approach

### 3.1 Preprocessing

The System ingests the raw XML files and processes them for downstream summarization. Ingestion relies on a document index file, which identifies the unique identifier (the document ID) for each relevant document in the AQUAINT, AQUAINT2, and 2009 TAC corpora. Those XML files are thereafter loaded into memory. Once in memory, each document is segmented into sentences using the nltk library. Subsequently, each sentence is tokenized (again using the nltk library. The system then generates processed data files for each document. These files include the document ID along with a handful of metadata. There is a blank line after the metadata. After that blank line, each tokenized sentence is presented on a new line. Paragraphs are separated by a blank line. These files, which are minimally processed, are the basic input for the System's content selection components.

## 4 Results

Phasellus imperdiet lobortis facilisis. Etiam sit amet nisi quis risus gravida luctus quis nec magna. Etiam velit purus, tempor ac ex eu, vulputate ultrices turpis. Donec blandit tempus placerat. Nullam commodo felis sit amet risus pulvinar fringilla. Praesent convallis, magna non vestibulum tincidunt, tortor purus fermentum elit, at scelerisque augue leo vitae turpis. Sed a iaculis lacus. Vivamus vulputate convallis lorem.

## 5 Discussion

Phasellus imperdiet lobortis facilisis. Etiam sit amet nisi quis risus gravida luctus quis nec magna.

Etiam velit purus, tempor ac ex eu, vulputate ultrices turpis. Donec blandit tempus placerat. Nullam commodo felis sit amet risus pulvinar fringilla. Praesent convallis, magna non vestibulum tincidunt, tortor purus fermentum elit, at scelerisque augue leo vitae turpis. Sed a iaculis lacus. Vivamus vulputate convallis lorem.

## 6 Conclusion

Phasellus imperdiet lobortis facilisis. Etiam sit amet nisi quis risus gravida luctus quis nec magna. Etiam velit purus, tempor ac ex eu, vulputate ultrices turpis. Donec blandit tempus placerat. Nullam commodo felis sit amet risus pulvinar fringilla. Praesent convallis, magna non vestibulum tincidunt, tortor purus fermentum elit, at scelerisque augue leo vitae turpis. Sed a iaculis lacus. Vivamus vulputate convallis lorem.

## A Appendix: Workload Distribution

- **Ben Cote**: Prepared/formatted D2-PDF; literature review for D3 content selection
- **Mohamed Elkamhawy**: Setup conda environment; created intial system architecture; literature review for D3 content selection
- **Karl Haraldsson**: Wrote the Preprocessing section for D2-PDF; reviewed and revised document-processing code
- **Alyssa Vecht**: Developed the ingestion and preprocessing mechanisms for the system
- **Josh Warzecha**: Wrote D2-presentation-PDF

## B Appendix: Code Repository & Additional Resources

Our team's repository can be found here on GitHub or directly via this URL: `https://github.com/summarization-team/summary`

## References


David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA. Association for Computational Linguistics.

Karen Spärck Jones. 2007. Automatic summarising: The state of the art. *Information Processing Management*.

Zhou Tong and Haiyi Zhang. 2016. A text mining research based on lda topic modelling. In *Computer Science Information Technology*, volume 6, pages 201–210.