



# Project 1 Deliverable 2

LING 575: Summarization  
Team 2



# Team 2

Ben Cote

Mohamed Elkamhawy

Karl Haraldsson

Alyssa Vecht

Josh Warzecha



# Project 1 Overview

## Multi-document summarization

- News articles
- Various categories and topics

## End-to-end system

- Document ingestion
- Content selection
- Information ordering
- Content realization

## Evaluation of output summaries

## D2: Preprocessing





# XML Processing

## Goal:

Read in XML files (in varying formats) containing a list of DocSets, each associated with a set of articles in the AQUAINT, AQUAINT-2, and 2009 TAC corpora.

## Implementation:

DocumentProcessor Class provides functionalities to parse XML documents and process them into a structured format suitable for further NLP tasks.

Uses Python xml package (ElementTree module).



# Sentence Segmentation and Parsing

## Goal:

Process the articles by breaking paragraphs into tokenized sentences.

## Implementation:

Lists of paragraphs are extracted from the XML document.

Sentence tokenization is accomplished using the NLTK package (tokenize module).



# Data File Generation

## Goal:

Return the processed data files.

## Implementation:

DocumentProcessor Class contains a method, `process_documents`, that reads the input XML file, parses its content, and processes it into a structured format (DocSet).

Processed documents are saved in the specified output directory.

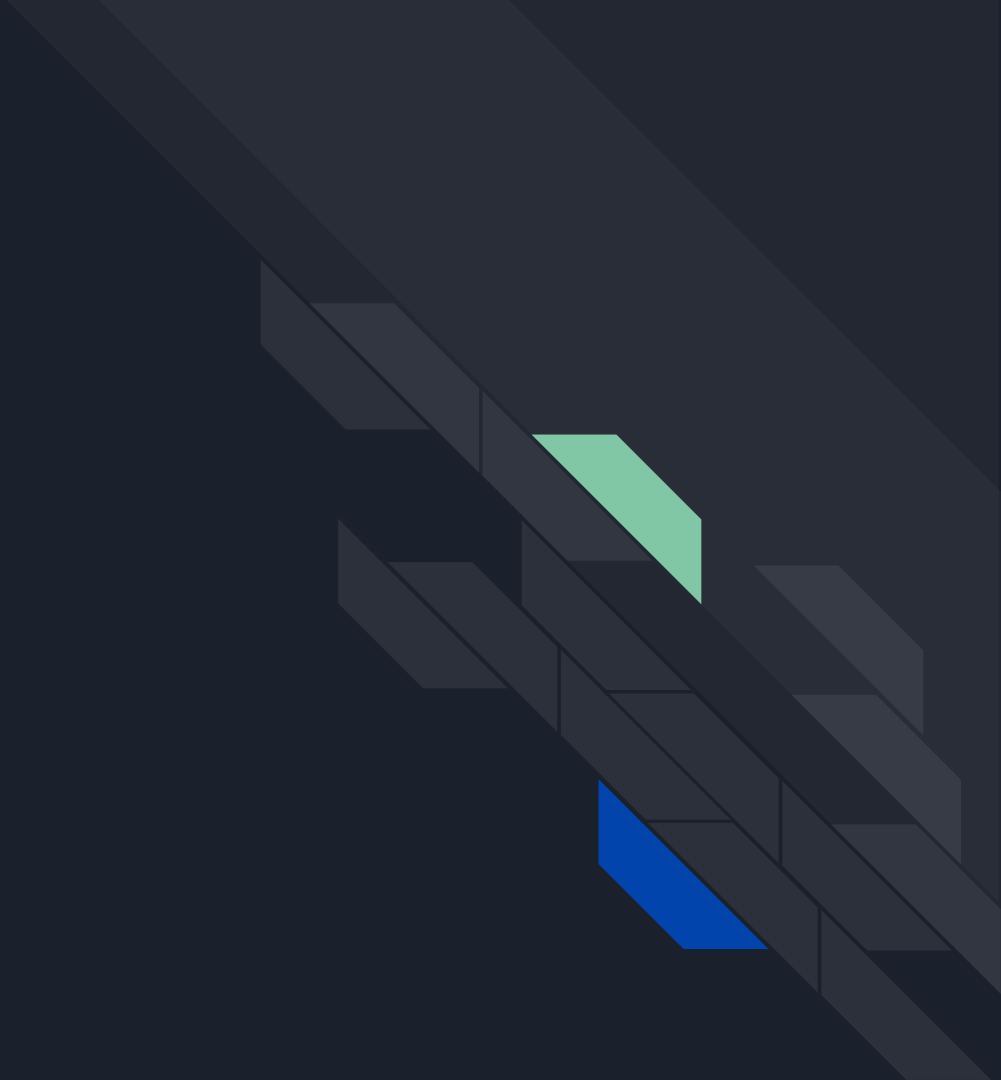
A list of processed document IDs is returned.

# Code Demo





Next Steps





# Initial Plans for D3

## Content Selection

- LDA/Topic modeling with Gensim
- TF-IDF

## Other Tasks

- Naive implementations planned for initial system iteration.
- Literature review in progress to guide more refined implementations.