

D2

Pre-Processing for Summarization

Anna Batra, Sam Briggs, Junyin Chen, Hilly Steinmetz

Table of contents

01

Introduction

Meet the Team!

02

Pre-processing

Pre-processing +
Tokenization Methods

03

Unit Tests

Checking our Code

04

Demo

Switch to Patas!



01

Introduction

Meet the team!

Anna Batra (she/her)

- UW 22' BA Ling, Math & Data Science minors
- TA-ed a bunch for DS&A
- I haven't learned much about about this subject

Fun Fact!

I like to sing in choir!



Sam Briggs (he/him)

- UW 22' BA Math + Ling
- I don't like writing summaries, so hope the computer will do it for me

Fun Fact!

I play both basketball and soccer and have an intramural basketball game tonight!



Junyin Chen (he/him)

- B.A in Linguistics and Japanese at UW

Fun Fact!

[Insert fun fact here...]



Hilly Steinmetz (he/him)

- B.A. in Linguistics from the University of Chicago
- Interested in the ways our discourse conventions convey importance

Fun Fact!

Love to stay at the East Asian Library at UW. Unfortunately the renovation is taking ages to complete.





02

Pre-processing

Pre-processing + Tokenization Methods



Pre-processing timeline

Get Data Path

Figure out which
corpus data in
docSetA belongs

Tokenize

Tokenize the sentences and
print results to output files

01 ——— 02 ——— 03 ——— 04

DatasetXML

Get the input XML
containing
docSetAs

Read Data XML

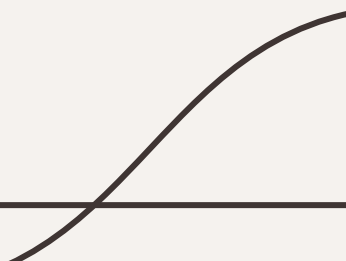
Parse the headline,
time, and
paragraphs based
on parsed data path

Tokenization Methods

spaCy

- Word tokenization on paragraphs
- Transformers

NLTK

- Sentence and word tokenization
 - Rule-based
- 



03

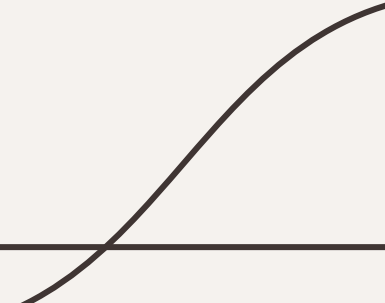
Unit Tests

Checking our Code

Our Tests

- Are all scripts and folders in the expected places
- Are all the data in new file formats in correct locations
- Can it get the root of the XML file tree?
- Can it can read AQUAINT, AQUAINT2, and TAC files properly
- Does it tokenize a document correctly?
- Does each file Test each kind of file outputs correctly?

List of packages

- lxml.etree (for processing AQUAINT, AQUAINT2, TAC files)
 - xml.etree.ElementTree (for processing docSetA file lists)
 - spaCy 2.0 (for word tokenization on paragraphs)
 - English model “en_core_web_sm”
 - NLTK (for sentence and word tokenization)
 - English model “tokenizers/punkt/english.pickle”
- 

Thanks

Do you have any questions?
(Demo upcoming!)

CREDITS: This presentation template was created
by **Slidesgo**, including icons by **Flaticon**, and
infographics & images by **Freepik**



04 Demo

Switch to Patas!