

Text Summarization

Week 3

Project Overview

- Week 3: Project definition and planning
- Week 4: Data cleaning and generating Word Cloud visualizations
- Week 5: Preprocessing text using RegEx
- Week 6: Tokenizing text using NLTK
- Week 7-8: Build text summarizer model
- Week 9: Prepare for presentation

Different Text Summarization Strategies

- 1) **Extractive Summarization:** These methods rely on extracting several parts, such as phrases and sentences, from a piece of text and stack them together to create a summary. Therefore, identifying the right sentences for summarization is of utmost importance in an extractive method
- 2) **Abstractive Summarization:** These methods use advanced NLP techniques to generate an entirely new summary. Some parts of this summary may not even appear in the original text.
- 3) **Hybrid** - use **extractive** summarization techniques to select most important phrases; use **abstractive** techniques to produce fluent output

Datasets we will be using:

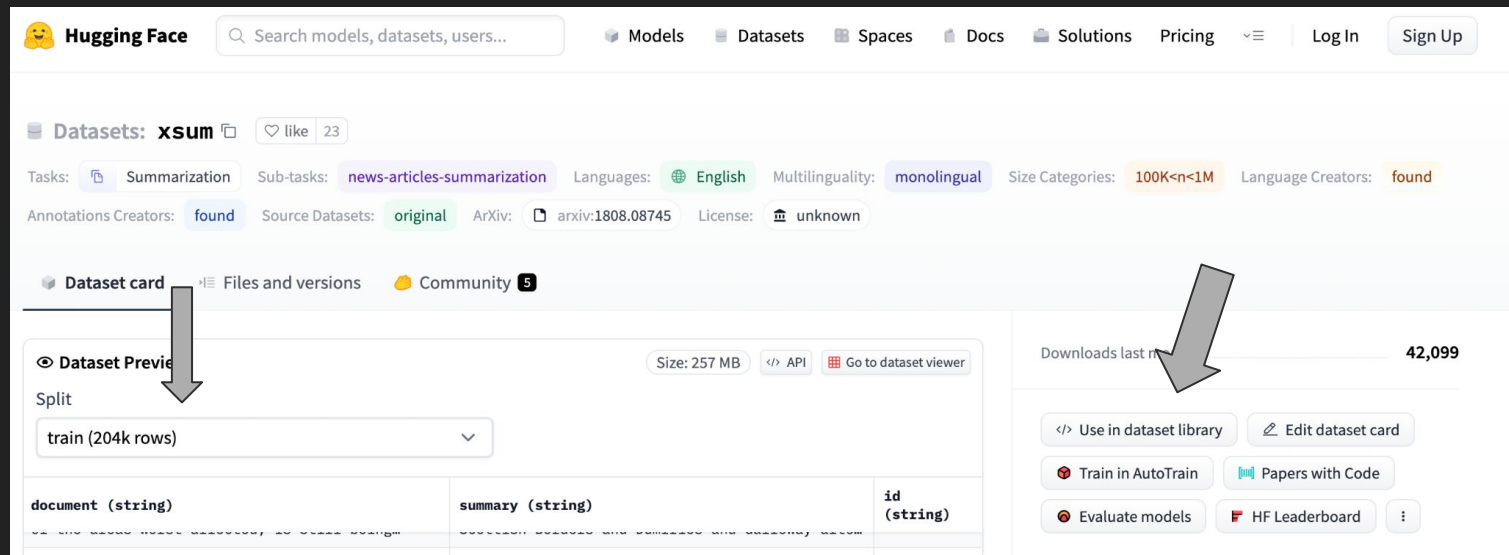
1) Huggingface Datasets

- a) US Congressional and California State Bills
 - i) <https://huggingface.co/datasets/billsum>
- b) BBC Articles
 - i) <https://huggingface.co/datasets/xsum>

2) <https://metatext.io/datasets-list/summarization-task>

3) Or any dataset of your choice

Loading a HuggingFace Dataset



The screenshot shows the HuggingFace interface for the 'xsum' dataset. The top navigation bar includes the HuggingFace logo, a search bar, and links for Models, Datasets, Spaces, Docs, Solutions, Pricing, Log In, and Sign Up. The dataset page for 'xsum' is displayed, showing its task (Summarization), sub-task (news-articles-summarization), language (English), and other metadata. A grey arrow points to the 'Dataset card' tab, and another grey arrow points to the 'Use in dataset library' button in the right-hand panel.

Hugging Face Search models, datasets, users... Models Datasets Spaces Docs Solutions Pricing Log In Sign Up

Datasets: xsum like 23

Tasks: Summarization Sub-tasks: news-articles-summarization Languages: English Multilinguality: monolingual Size Categories: 100K<n<1M Language Creators: found

Annotations Creators: found Source Datasets: original ArXiv: arxiv:1808.08745 License: unknown

Dataset card Files and versions Community 5

Dataset Preview Size: 257 MB </> API Go to dataset viewer

Split

train (204k rows)

document (string)	summary (string)	id (string)
On the basis of the summary, the text is a...	Between 1992 and 1993, the summary is...	

Downloads last month 42,099

</> Use in dataset library Edit dataset card

Train in AutoTrain Papers with Code

Evaluate models HF Leaderboard

</> How to load this dataset directly with the datasets library

```
from datasets import load_dataset

dataset = load_dataset("xsum")
```

Copy

```
In [ ]: pip install datasets
```

```
In [ ]: from datasets import load_dataset
```

```
# choose what split you want from the dataset using the split argument
dataset = load_dataset("billsum", split = "test")
```

```
In [4]: import pandas as pd
dataset.set_format("pandas")
df = dataset[0:]
df
```

Out[4]:

	text	summary	title
0	SECTION 1. ENVIRONMENTAL INFRASTRUCTURE.\n\n ...	Amends the Water Resources Development Act of ...	To make technical corrections to the Water Res...
1	That this Act may be cited as the "Federal Fo...	Federal Forage Fee Act of 1993 - Subjects graz...	Federal Forage Fee Act of 1993
2	SECTION 1. SHORT TITLE.\n\n This Act may be...	. Merchant Marine of World War II Congression...	Merchant Marine of World War II Congressional ...
3	SECTION 1. SHORT TITLE.\n\n This Act may be...	Small Business Modernization Act of 2004 - Ame...	To amend the Internal Revenue Code of 1986 to ...
4	SECTION 1. SHORT TITLE.\n\n This Act may be...	Fair Access to Investment Research Act of 2016...	Fair Access to Investment Research Act of 2016
...
3264	SECTION 1. PLACEMENT PROGRAMS FOR FEDERAL EMPL...	Public Servant Priority Placement Act of 1995 ...	Public Servant Priority Placement Act of 1995
3265	SECTION 1. SHORT TITLE.\n\n This Act may be...	Sportsmanship in Hunting Act of 2008 - Amends ...	A bill to amend title 18, United States Code, ...
3266	SECTION 1. SHORT TITLE.\n\n This Act may be...	Helping College Students Cross the Finish Line...	Helping College Students Cross the Finish Line...
3267	SECTION 1. SHORT TITLE.\n\n This Act may be...	Makes proceeds from such conveyances available...	Texas National Forests Improvement Act of 2000
3268	SECTION 1. SHORT TITLE.\n\n This Act may be...	Federal Power Asset Privatization Act of 1995 ...	Federal Power Asset Privatization Act of 1995

3269 rows x 3 columns

Structure of Project

- Working in partners or as one group?
- Partners:
 - Assign each of you a partner
 - Each of the pairs works on analyzing a different dataset/method, approaching the project in their own way
 - Come together and share our findings
- One Group:
 - Have to meet up more often
 - All work on the same dataset
 - Work on the project as a whole and do each step during these meetings

Potential Social Ideas?