# NLP Resources

Prepared for NLP Task Force, Batch #4

This README file and datasets mentioned here are at <https://drive.google.com/drive/folders/1CWudkqGZJMk3k_bjErilpaY5lRWycCeZ?usp=sharing>

## 1. Devopedia Dataset

This is a dump of content on Devopedia. This content is shared only for the purpose of the Task Force. Please don’t share or forward to anyone outside the Task Force.

The file is a JSON file at **Devopedia/devopediaArticles.v2525.json**. Only published articles are part of the file. This version has 291 published articles.

Each article content is available in two fields:

* introtext: original text in Markdown format
* cleantext: text after removing most Markdown syntax and links

A sample Python Notebook to read this file is shared at **Devopedia/Devopedia-Article-Reader.ipynb**

## 2. Devopedia References Dataset

Devopedia articles often include references. Some models may require the content of these references for training purposes. Hence, we have scraped the content of these references from their URLs. Note that not all content was obtained successfully. Old URLs might have returned 404 errors. In some cases, servers might have prevented automated scraping or some other error occurred.

References at source are mainly in HTML, PDF or plain text. HTML and PDF formats are automatically converted into text format. Sometimes conversions fail, which implies that the text files will be missing but the original files will be present in the folder. Sometimes conversions give wrong output.

The dataset is in folder **articleRefs.v2525**. Content for 7510 out of 7704 crawled URLs are in this dataset. Content files are actually compressed into one of three files: **articleRefs.htm.tar.gz**, **articleRefs.pdf.tar.gz** and **articleRefs.txt.tar.gz**

The file **articleRefs.v2525/meta.v2525.jl** is a suitable starting point. This is a JSON Lines file. For each URL, the file records the crawl status. Any status other than 200 should be ignored. The field fname records the filename in which the content of that URL is stored. This file is usually **\*.htm**, **\*.pdf** or **\*.txt**. If conversion is successful, file *abcd.htm* or *adcd.pdf* will have an equivalent *abcd.txt*. If conversion fails, the equivalent text or HTML file will be absent.

A sample Python Notebook to read the scraped references is in file **Devopedia/Devopedia-Ref-Reader.ipynb**

## 3. Wikipedia Dataset

Some models may benefit from fine tuning with content that’s relevant to the work of Devopedia. For this purpose, we have crawled and selected about 17,800 Wikipedia articles.

These files are in the folder **WikipediaArticles/2020Nov09.084212**. The folder in fact contains many more than 17k articles. Articles suitable for Devopedia are noted in file **WikipediaArticles/2020Nov09.084212/all-accepted.csv**, which should be a suitable starting point.

Multiple Wikipedia articles are combined together into JSON files. JSON files are zipped into **\*.json.bz2** files. These filenames are referenced in the **all-accepted.csv** file. Other CSV files in the folder can be ignored. They have been combined into **all-accepted.csv**.

Each Wikipedia article is available in the JSON files in two formats: [Wikitext](https://en.wikipedia.org/wiki/Help:Wikitext) and HTML. You may use whichever is convenient.