Practice II

Text normalization

Objectives

1. Collect information from different sources
2. Normalize the collected text

Collect information from different sources

- Get news using RSS feeds from La Jornada and Expansión platforms using the following URLs:
  - https://www.jornada.com.mx/v7.0/cgi/rss.php
  - https://expansion.mx/canales-rss
- The data collection should be done once a day during 5 days at *agreed time*
- News can be repeated from one day to the next, so you must avoid collecting it again
- For each news article extract:
  - Title (<title>)
  - Content summary (<description>)
  - Section
  - URL (<link>)
  - Date of publication (<pubDate>)
- Section of interest are:
  - Sports
  - Economy
  - Science and technology
  - Culture

Collect information from different sources

- With the information collected you must generate a corpus in csv format with the following format
- We will call this corpus the *raw data corpus*

Source	Title	Content	Section	URL	Date
La Jornada	Title 1	Content 1	Sports	https://...	04/03/2024
La Jornada	Title 2	Content 2	Economy	https://...	05/03/2024
…	…	…	…	…	…
Expansión	Title n	Content n	Culture	https://...	04/03/2024
Expansión	Title m	Content m	Science and ..	https://...	08/03/2024

Normalize the collected text

- Apply the following normalization processes to the title and content of the raw data corpus
  - Tokenization
  - Remove stop words from the grammatical categories: articles, prepositions, conjunctions and pronouns
  - Lemmatization
- For Stop Words you can use a defined list of words (stop words dictionary) or identify the grammatical category of the word
- The normalized version of the corpus should be saved in a csv, with the same format as the previous one, called *normalized data corpus*

Evidence

- Upload as evidence to the Teams platform the source code of your program and the generated corpora (*raw data and normalized data*)