Project – ISIS 4221

Stage I: Data capture (5%/25% grade)

Due date: 06-04-2021

The course project will focus on shaping the public discussion around COVID. Different modeling strategies using natural language processing tools should be performed on public documents published on the internet in three different languages (English, Spanish, and one of your selection). In this first stage you will have to:

I. Identify information sources in three different languages (English and Spanish are mandatory).
   a. Social Networks: twitter, reddit, tumblr, mastodon, care2, quora, etc.
   b. News: New York Times, BBC, CNN, etc
   c. Academic papers: arXiv, core (core.ac.uk)
II. Develop and document automatic extraction strategies.
   a. The extraction process under no circumstances can affect the availability of the sources.
   b. Develop strategies to ensure that you do not have duplicate documents.
   c. At the end of the process, you will have to export the consolidated dataset for each language into a plain text file. Each row will correspond to one document. If you use a schema like json or xml, you must include an explanation of it.
   d. Please include for each document the publication date, the source, the internal id/url, author, and all the additional information that provides the source (e.g. geolocalization for tweets).
III. The information extraction process will start from the moment of publication of Stage I and until May 14.
   a. At least 500,000 DIFFERENT documents are expected for each language.

Report: About the previous points, you must prepare a detailed report to deliver on April 6. You have to upload the datasets with the information extracted to date, and the scripts used. Your report should describe the sources selected, and the extraction processes status.