Project – ISIS 4221

Due date: 04-12-2023

At this stage of the project, two Colombian indigenous languages have already been selected, and datasets for the translation task have been identified. Your participation in the final project for the course entails constructing various translation models and comparatively evaluating them.

Please adhere to the following set of steps:

1. **Definition and Training of Translation Models:**
    - Justify the selection of each model, including its hyperparameters and architecture if applicable.
    - Consider exploring variations of the transformer architecture.
    - Evaluate the fine-tuning of pre-trained models in other languages.
    - Incorporate linguistic information specific to the language.
2. **Definition of Training, Validation, and Testing Data**
    - The preprocessing steps must be detailed; it is crucial to ensure proper alignment between languages.
    - Carefully consider the partitions to test interesting hypotheses derived from the data. For instance, contemplate the significance of incorporating information from dictionaries versus textual documents such as the Bible and the constitution.
3. **Validation and Comparison of Results at the Following Levels:**
    - a. Models.
    - b. Languages.
    - c. Datasets.
4. **Additionally, conduct a comparative analysis of your work with existing literature.**

Exchange of information between groups

You carry out collaborative processes with other groups at the data level exclusively. You must describe in detail how you carried out the integration processes. **No code or models can be exchanged.**

Final project submission outcomes

1. **Code and datasets**: Use jupyter notebooks and be sure that the notebook is executed and contains the results before submitting. All classes, methods, functions and free-code MUST contain

docstrings with a detailed explanation. You must submit the datasets organized, compressed, and they should be uploaded to Bloque Neon directly or via a link to a repository (i.e. Dropbox, OneDrive, etc).

Important Note: Ensure that the teacher/monitor can replicate the results by executing the Python scripts attached by each group in the submission, utilizing the course GPUs.

2. **Paper**: Together with the notebooks, you must submit a written report as an academic paper. The paper must include all the sections outlined in:

https://www.uregina.ca/student/ssc/assets/docs/pdf/Research-Paper-Writing-in-Engineering.pdf

You should use the Springer LaTeX template for academic papers in engineering. You should use the Springer LaTeX template for academic papers in engineering:

https://www.overleaf.com/gallery/tagged/springer-nature

I recommend checking this paper:

https://aclanthology.org/2023.americasnlp-1.9.pdf

3. **Presentation:** All members of the group are required to prepare a presentation showcasing the project results, record a video, and publish it on YouTube.