CSCE 5290/4290:

Natural Language Processing Project

# **Names:**

Oriana Borges-Salinas, Zachary Chenausky, Anas Mohammed Nayeem, Dagar Rehan

# **Project Title:**

Research Paper Translation and Summarization

# **Github:** [**https://github.com/oriAleph/Research-Translation-Summarization**](https://github.com/oriAleph/Research-Translation-Summarization)

# **Project Proposal:**

The project's primary goal is to translate and summarize research papers. The project will be divided into two sections. The first section focuses on research paper translation from one language to another. The second section focuses on summarizing the document to provide the user with the most important information. Four deep-learning models will be trained using the same training set and epochs for both portions of the project. There will be four models in total, two for translation and two for summarization. Furthermore, comparisons will be made to show which models performed well and which did not, and ideas for improving the models will be tested.

TL;DR: Train and compare four deep-learning models for the translation and summarization of research papers.

# **Description:**

## **Motivation**

The goal of the project is to develop software that will enable someone who doesn't speak the original language to comprehend the significant ideas of a research article. This comprises translating the paper's content from the original tongue into English and summarizing it for easier user comprehension. All the models will also be compared to one another to see which combination performed the best and how to make the models better.

## **Significance**

This project has many practical applications and is a key component of natural language processing. A wide range of businesses and organizations use translation and summarization, two of the most explored topics in computer science. With this study, we will concentrate on Czech and English research syntax and semantics because various languages have distinct ways of describing information. Due to the project's NLP components, we apply what we learned in class to use by designing a tool for cross-cultural research comprehension. This serves as an excellent learning tool for our team.

## **Objectives**

* Train two models that can translate from the native language to English
* Train two models that can summarize English text and give key points
* Compare which combination of models gives the best results
* Try to improve the models to give better results

## **Features**

* Natural Language Toolkit (NLTK)
* Hugging Face
* spaCy
* TextBlob
* English datasets
* Czech datasets
* REGEX
* GPT Corpus
* Collections
* TensorFlow
* Keras
* Numpy

## **Task Visualization**

|  |  |
| --- | --- |
| **Name** | **Task** |
| Dagar | Summarization - T5-small model |
| Zachary | Translation - RNN model with Attention |
| Anas | Summarization - Abstractive roBERTa model |
| Oriana | Czech to English NMT with a Transformer and Keras |