UNIVERSITY OF SOUTHERN MAINE

Text Mining and Analytics, Spring 2024, Course Project Proposal

Project Name (refer to lab's name if working on CLEF labs): SimpleText

Team leader (point of contact): Nick Largey

Team members: Nick Largey, Finn Michaud, Ben Gaudreau and Gabi Akers

Project description: Given a scientific text, use a machine learning or deep learning model to provide a simplification so that the text can be understood by as wide an audience as possible.

Labor division: Explicitly mention what each team member will be doing (referring to task number if working on CLEF labs)

Collectively, we will be working on the CLEFlab2024 SimpleText lab. Nick will be taking on task 2, Complexity Spotting, in which the participant is tasked with analyzing a given text and isolating what parts include complex subjects then providing explanations and simplifications for those subjects. Along with this he will be working with all the team members to gain a better understanding of all of these foundational problems for text simplification. Finn and Ben will be taking on task 1, Content Selection, in which they will retrieve parts of a text in order to provide a simplified summary of the given text. Gabi will be working on task 3, Text Simplification, which will take in a scientific text and re-write it in simpler terms.

Many of the techniques we will use for data pre-processing and EDA will be the same, which will lead to collaboration, but since our tasks are different enough, we will need to produce our own models in the end.

GitLab/GitHub Link: The team leader should create a git repository and share it among the team members and make it available for the instructor. Throughout all phases of the project, this repository will be monitored for team participation.

https://github.com/NickLargey/SimpleText

Dataset Description and Evaluation Measures: Just brief indication of your understanding of your proposed approaches evaluation

The dataset we will be receiving will be a group of scientific articles, they have not been released yet, so much of what our approach will be decided in our EDA period.