Capstone_README

#SchoolNotes

NL2Esgish2 README

Convert ISS Stoxx formal query language: Esgish2 to natural language and vice versa.

Description

NL2Esgish2 is a NLP-based tool which converts human-readable text input into the query language Esgish2, with the intent of streamlining non-technical user's interactions with the ISS Stoxx database.

Features

The scope of this project is an API program capable of conversion to and from Esgish2. This means there are two main functions that will be syncretically developed.

1. Esgish2 -> Natural Language conversion. (FOCUS OF SPRINT 1)
   This is the initial focus of the project due to its lack of reliance on a pre-generated set of data. Due to the formatting of Esgish2, this conversion does not rely on vast quantiles of data being utilized to train a model. In a perfect world we would utilize a model regardless, however our coordinator at ISS Stoxx has only provided us with a datasheet containing queries in Esgish2 without associated natural language.

   This means it will be far simpler if we first implement a simple parser for Esgish2 -> NL as opposed to trying to force a model when we have no relevant data.

Machine Learning Model Esgish2 -> Natural Language conversion (FOCUS OF SPRINT 2)
After Sprint 1 we were asked by our representative within ISS Stoxx to attempt a more model based conversion from Esgish2 to English. Because of this we shifted technologies to

2. Natural Language -> Esgish2
   Once the Esgish2 -> NL converter has been developed, we can now utilize it in tandem with the data given to us regarding well formatted Esgish2 queries. From here we are now armed with usable data for modelling.

   Our goal now is to train a NLP model that can consistently and efficiently convert natural language to well-formatted Esgish2 queries.

Technologies and Tools

Our old plan revolved around investigating and utilizing a variety of different Python libraries:

- NLTK
- spaCy
- Transformers
  Which then shifted into utilizing the ANTLR language to assist in creating a program to digest the Lexer and Grammar information given to us about Esgish2.

When asked to shift to a Model-based system instead of static query translation we began utilizing Google Colab along with T5 Flan. After difficulties with multiple group members being unable to interface with Google Colab, we shifted to using Jupyter Notebooks with a Mistral-7b-Instruct LLM to translate the original queries to English, then continued using T5 Flan to train our model.

Plan

1. Work to better understand Esgish2 as a query language
2. Begin developing Esgish2 -> NL tool
3. Ensure Esgish2 -> NL tool works sufficiently

After finishing the top 3 we moved to a Model Based system, leading to the redoing of the first 3 steps instead with training a model.

4. Utilize the Esgish2 -> NL tool on the dataset we have been given to create the training set.
5. Split the dataset and begin creating the model for the NL -> Esgish2 tool
6. Train the model to a high success rate
7. Request more data if needed
8. Implement a short demo/display if extra time