Name: Chris Binoi Verghese

StudentID: 21092999

Email: cbverghe@uwaterloo.ca

Name: Matthew Badal Badalian

StudentID: 20777980

Email: mabadalb@uwaterloo.ca

Project Title: A Large Language Model to Answer Religious Questions Based on Sacred Text in Real-Time Using Classification and Word Embedding

Work done:

Github: https://github.com/mbadalbadalian/Introduction-to-Machine-Learning--Graduate

Read one of the datasets i.e. the bible and creates a dictionary mapped from book to chapter to verse, and then wrote this dictionary into a json file which can be read from instead of consistently creating a dictionary each time.

Dataset obtained from: https://github.com/aaronjohnsabu1999/bible-databases/blob/master/XML/ESVBible\_Database.xml

Program Name: PrepareBibleData.py

Then there is a tokenizer file, which reads this dictionary, and then breaks down the verses into lemmatized tokens of nouns and verbs which can be used for classifications. We used the spacy library for tokenizing while implementing the language model: “en\_core\_wb\_sm”

After doing some research, we are considering using BERT or LST for classification and are reading up on them. We are also considering using an RNN.

Extra:

That is our team’s update. Now here’s a badly written is a Haiku description of it for a quick laugh.

Bible's sacred text,

Mapped from book to verse in style,

JSON's grace unveiled.

Tokenizer awaits,

Lemmatizes nouns and verbs,

Verse's wisdom shines.

BERT and LSTM,

Classify divine knowledge,

Text's meaning unfolds.

In haikus we learn,

From ancient sacred scripture,

Knowledge now digitized.