Project Progress Report

UIUC: Text Information Systems

Group Name: Team Coco

Theme: Free Topics

Specific Topic: ~~AllenNLP~~ Question-and-Answer Chatbot

Captain/Project Leader (and sole team member): Joel Haas, joelah2@illinois.edu

Project Task

Configure a question-and-answer chatbot ~~using the Allen Institute for Artificial Intelligence open-source library AllenNLP~~. Will now use native python instead of AllenNLP due to inability to successfully install AllenNLP python library.

Project Update

I was not able to successfully install the AllenNLP python library. I spent 3 hours researching the AllenNLP capability. I then tried to install the AllenNLP library so I could start learning the library. However, after 2 hours troubleshooting and researching the installation challenges that others have also had trying to install the library, I decided I needed to move on. So now I will configure a question-and-answer chatbot using native python libraries rather than the AllenNLP python library.

Programming Language:	Python
Tools:	Scikit-Learn, NLTK, Pandas
Systems:	Personal laptop
Datasets:	Deepmind NarrativeQA Reading Comprehension https://github.com/deepmind/narrativeqa
Expected Outcome:	NLP pipeline for a Chatbot
Evaluation Methodology:	Compare the algorithm's predicted answers against the idea answers given in the annotated question and answer dataset

Completed Tasks by Workload (13 hours):

3 hours - AllenNLP research
2 hours - Troubleshooting installation of allennlp python library
2 hours - Research labeled question / answer datasets
6 hours - Cleaning and preparing datasets

Pending Tasks by Workload (15 hours):

7 hours - Configure and test Q&A NLP pipeline using TF-IDF and cosine similarity
3 hours - Evaluate performance and iterate
5 hours - Documentation and reporting requirements

Challenges

Preparing (wrangling) the datasets proved more challenging than expected. I have made quite a bit of progress so far, but still have more formatting remaining to get the data (the questions and the summary documents that are used for searching for the answers to the questions) into the required format for the TF-IDF and cosine similarity algorithms in my NLP pipeline.