Initial Project Plan

Group Name: DFA: Dumbledore’s F*cking Army
Group Members: Nikitha Murikinati, Mihika Bairathi, Sunjana Kulkarni, Elias Joseph

Relationship between Asking and Answering Components:

We want our asking and answering components to run separately, because we do not want our answering capabilities to be constrained to the questions we can ask Thus, while we will make sure we can answer the questions we ask, we think it’s important that the answering component of out project is not dependant on the asking component.

Tools and Development Data:

The project codebase will be in Python3. Other tools we plan to use are the NLTK module and Tensorflow if needed. The development data that is provided will be used to manually run and test our model on. We will compare how our model is faring on the answer system with the answers found in the text. Fluency checking of the questions and answers returned by our model on the given data will also be done.

Sharing code/data and coordinating with team:

We plan on using Github to share all of our code and data within the team. Each of the members will create their own branch of the code and then commit it to a master branch. Communication within the team will be done using a Facebook Messenger group chat and Gmail if needed.

Technical Approaches:

For our technical approach, we are planning on using a lot of the existing methods utilized in question answering/question generation systems.

On the question generation side, we would plan on using a vector space model to store information. In our model, we would be using some type of weighting system to weight proper nouns higher, since some of the best questions will be those that utilize proper nouns as their focal point (i.e. in an article about Pittsburgh, some good question would be “Where is Pittsburgh”, or “How large is Pittsburgh”, with Pittsburgh, the noun, being the focus). On the other hand, we will be weighting determiners, conjunctions, and prepositions like “the”, “and”,

and “on” less so they don’t become the central focus of the question. Furthermore, we would utilize seeding tuples to find patterns in the text to base questions on, as well as entity recognition to determine interesting entities to ask questions about. Our model will also focus on the introduction paragraph of the wikipedia article in order to find big ideas and utilize those to create questions. Most likely, the answers to most questions asked about the article will be relating in part to the introduction paragraph, which is why it is an ideal place to do information extraction.

On the question answering side, we can utilize keyword analysis to target words in the text and figure out what types of nouns they are (person, place, or thing). This categorizing will help us narrow down potential answers for questions like “Where [...]”, “Who [...]”. It would also help us in question generation: we could pair the question words, “Who”, “Where”, etc. with the appropriate noun when generating questions.