David Schulz

CS 4980 001

4/26/2022

**Data Collection and Processing**

Since this project requires two different types of data, the knowledge database and the corpus of possible questions, data had to be collected from different types of sources. The knowledge database was captured from the following websites:

* [www.minecraft.fandom.com/wiki/Minecraft\_Wiki](http://www.minecraft.fandom.com/wiki/Minecraft_Wiki)
* [www.game.guide/minecraft](http://www.game.guide/minecraft)

The corpus of possible questions was captured from the following websites:

* [www.quizlet.com](http://www.quizlet.com)
* [www.allthetests.com](http://www.allthetests.com)
* [www.gamefaqs.gamespot.com/pc/606524-minecraft/answers](http://www.gamefaqs.gamespot.com/pc/606524-minecraft/answers)

None of the data from the sources, both the question corpus and knowledge base, has one specific author because everything has been created and edited by many people over the years. However, during my initial search for the data, I stumbled upon a GitHub repository that did the web scraping for me, created by someone named Corentin Dumont. Their repository also contains the skeleton of an ontology, which will be nice for future work, but I won’t be using it for the scope of this project.

Corentin’s repository has the knowledge database and question corpus in two separate folders. The knowledge database folder contains three folders, one for each of the original website sources. There are three instead of two because the data was scraped back when the Minecraft Fandom Wiki was split between two sites: gamepedia.com and wikia.com. In each folder, there are more folders that split the information into topic categories from the game (blocks, items, mobs, building, etc.). Those folders may also have more folders that split the information further (natural, nether, ores, mechanisms, etc. in the blocks folder). The lowest level of the folder tree contains text files for each specific entity or topic. The formatting of the text data is the same between all 3 source folders. Each text file contains raw sentences of information from their corresponding sources. Each line in a file is a single continuous idea, so the sentences on one line are related, but sentences on separate lines are not as closely related.

As for the question corpus, there are three files in the folder, but only the csv file is relevant to this project. Its columns are different topics in Minecraft (recipe, blocks, mobs, items, etc.) and the rows are questions about those topics, so some columns contain more questions than others, which means that columns in many of the rows are empty.

Before reading the data, the knowledge database text files that contain information about the same topic from different sources will be put together into one file, since the distinction between sources is irrelevant. Both types of data will be read with the pandas read\_csv() function, even though the knowledge database is made up of text files. Having everything in DataFrames makes it easier and more efficient to work with. After the data is read, the knowledge database will be pre-processed through tokenizing, stemming, and stop word removal so that the relevant data can be put into an index term dictionary, where each keyword maps to a list of all documents that contain the keyword. The question corpus will be pre-processed through tokenizing, POS tagging, stemming, and stop word removal. The POS tagging will be used for answer extraction later on in the process, while the stemming and stop word removal is used to extract keywords from each question.

As a result of training, the final model should be able to read a question input by the user, pre-process it, use keyword extraction to match it with documents from the knowledge database dictionary, perform file ranking to determine which document has the best answer, and extract the correct answer using POS tagging.