Project Update - The Wikipedia Game

Mark Kong, Benjamin Rafetto, and Claire Stolz

November 25, 2018

1 Problem Descrition - Wikipedia Game

The Wikipedia Game is a path-finding challenge where a player is presented with two Wikipedia pages and is asked to find a path from the first page to the second by clicking links to travel from article to article. The player's score is determined by the length of the path and the time it took to find the path. Due to the interconnectedness of various topics and a moderately large branching factor, there is a wide variety of possible paths. Our goal is to solve this game in as efficient a manner as possible.

2 Progress

So far we've focused on reviewing and implementing various tools that will allow us to work with Wikipedia data and build efficient heuristics for the Wikipedia game.

2.1 Data Acquisition

We've downloaded the entire English Wikipedia database, as well as a simple subset we can use for prototyping algorithms before working with the larger dataset. We've also identified tools for interacting with those datasets, including packages such as WikiUtils for parsing sql.gz files and ordinary xml parsers for xml.bz2 dumps.

2.2 Querying Wikipedia

We've tested various tools for interacting with Wikipedia's API, including the python packages wikipedia and pywikibot, and have settled on pywikibot for live querying, as it offers some advanced functionality and builds in more respectful querying practices to avoid calling Wikipedia's servers too often.

2.3 Natural Language Processing

We've downloaded Google's pre-trained word2vec word embeddings, and used the gensim python library to work with it. Our initial plans include implementing a greedy search based on calculating the nearest distance to our goal page using the pre-determined embeddings before evaluating other approaches.

2.4 Project Goals

We've also refined our goal for the project to focus more on the irreversibility inherent in the rules of the Wikipedia game. Technically no backtracking is allowed, so for our initial approach we've emphasized identifying useful heuristics to efficiently navigate in the correct direction towards our target page.

3 First Attempts

We have successfully combined Wikipedia scraping and word2vec to implement a greedy search that has so far yielded promising results for simple topics. By implementing a priority queue based on the average distance between the words in an article title and the goal node we have been able to successfully find paths between a small number of sample topics, although the lengths of our paths have varied widely from as little as ten to as many as forty steps.

4 Problems

The extremely large size of the Wikipedia database (16 GB compressed) makes working with it challenging in a variety of ways. A particular one is how to use clustering to determine groups and distances between topics given the scale of the dataset. We will also need to consider how to deal with pages that are combinations of words or phrases, such as "Natural Environment" rather than "Environment", or topics that are not identifiable English words, such as "Vostok 1".