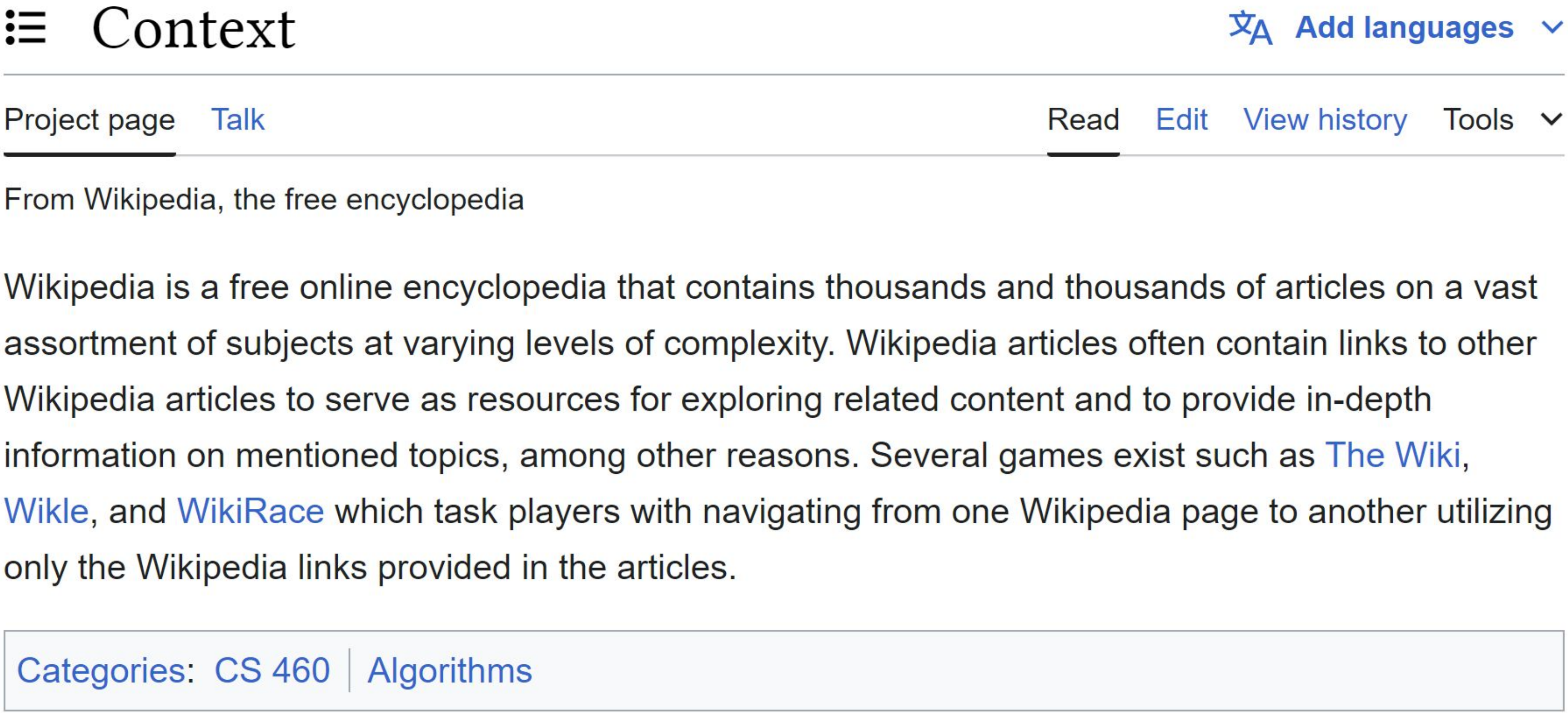
WIKIPEDIA Traverse

By Chaz Gabelman



Context

From Wikipedia, the free encyclopedia

Wikipedia is a free online encyclopedia that contains thousands and thousands of articles on a vast assortment of subjects at varying levels of complexity. Wikipedia articles often contain links to other Wikipedia articles to serve as resources for exploring related content and to provide in-depth information on mentioned topics, among other reasons. Several games exist such as The Wiki, Wikle, and WikiRace which task players with navigating from one Wikipedia page to another utilizing only the Wikipedia links provided in the articles.

Categories: CS 460 | Algorithms

OBJECTIVE

I wanted to create a program that will find an efficient path from one Wikipedia page to another utilizing principles of graph traversal algorithms.

To do this, I needed an algorithm that utilizes a heuristic to guide it toward the correct article. The heuristic I have decided to use for this problem is semantic similarity, in other words, looking at how similar two words are in meaning.

Essentially, I am repurposing a greedy graph traversal algorithm and applying it to Wikipedia articles to create a program that will traverse from one article to another efficiently via the semantic comparison of the included links with the target article.

Wikipedia
Search
Alcatraz Federal Penitentiary
United States Penitentiary, Alcatraz
1.08

This, but using code

ALGORITHM

1. Determine if current URL is the target URL
2. If not, go to current URL
3. Collect valid /wiki/ links (links that go to other Wikipedia articles)
4. Parse links for article titles
5. Run a semantic comparison on each article title with the target article title, keeping track of the highest semantic similarity
6. Go to the page with the highest semantic similarity
7. Repeat steps 1-6 until you have reached the target

CODE - Github link

Runtime Complexity: $O(n^2)$, with n being the total number of articles on Wikipedia. In the worst case, each article has a link to every other article and thus we are forced to semantically analyze every link n times.

Space Complexity: $O(n)$. As we traverse, we add visited page URLs to an array to prevent looping as well as to document the path taken. In the worst case, we would have to traverse every article on Wikipedia and thus would have to add n URLs to our array.