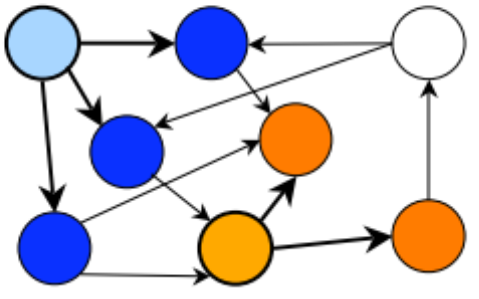
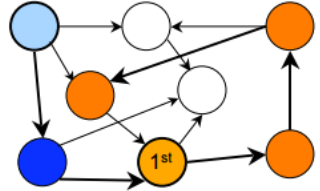
**Solution 2C: Breadth First Crawling vs Depth first Crawling**

All crawlers work in a recursive or loop fashion, but there are two different ways to handle it. Links can be crawled in a depth-first or breadth-first manner. I have concluded the following points from the results of the program regarding the two approaches to web crawling:

* **Breadth-first** crawling checks each link on a page before proceeding to the next page. Thus, it crawls each link on the first page and then crawls each link on the first page’s first link, and so on, until each level of links has been exhausted.
* This approach finds pages along shortest paths.
* If we start with “good” pages, this keeps us close; Breadth-First Search Crawling Yields High-Quality Pages: this we can say because the context of the key phrase being searched is maintained for a longer time in the BFS approach
* The procedure generally is time consuming



* **Depth-first** crawling follows each possible path to its conclusion before another path is tried. It works by finding the first link on the first page. It then crawls the page associated with that link, finding the first link on the new page, and so on, until the end of the path has been reached. The process continues until all the branches of all the links have been exhausted.
* This approach finds pages across the web and not necessarily along the shortest paths
* We tend to wander away from the topic to be searched (“Lost in cyberspace”). The major disadvantage of Depth First crawling is how to predetermine the end condition and total no of web pages to crawl to produce a good corpus.
* The procedure generally takes shorter time



Choosing whether to use depth or breadth-first crawling often depends on the crawling application and its needs. But in my opinion I believe as far as the as our goal remains to produce a good corpus, breadth first approach is better.

**Results from the code**: As you can see in depth first approach we deviate from the topic ‘solar’ as we proceed further into depth of the web (As highlighted in red).

First 5 links of Breadth first crawling

http://en.wikipedia.org/wiki/Sustainable\_energy

https://en.wikipedia.org/wiki/Energy\_conservation

https://en.wikipedia.org/wiki/Cogeneration

https://en.wikipedia.org/wiki/Efficient\_energy\_use

<https://en.wikipedia.org/wiki/Geothermal_heat_pump>

First 5 links of Depth first crawling

http://en.wikipedia.org/wiki/Sustainable\_energy

https://en.wikipedia.org/wiki/Energy\_conservation

https://en.wikipedia.org/wiki/Cogeneration

https://en.wikipedia.org/wiki/Efficient\_energy\_use

https://en.wikipedia.org/wiki/Compact\_fluorescent\_lamp