Shri Vile Parle Kelavani Mandal’s
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA: 3.18)

JUNAID GIRKAR | 60004190057 | BE COMPS A2 | WEB INTELLIGENCE

EXPERIMENT - 1

AIM: Implementation of Page rank estimation

THEORY:

PageRank (PR) is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages.
According to Google:
PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

It is not the only algorithm used by Google to order search engine results, but it is the first algorithm that was used by the company, and it is the best-known.
The above centrality measure is not implemented for multi-graphs.

The formula for calculating PageRank is:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$

Where the PageRank value for a page u is dependent on the PageRank values for each page v contained in the set Bu (the set containing all pages linking to page u), divided by the number L(v) of links from page v. The algorithm involves a damping factor for the calculation of the PageRank.

CODE:

```
import numpy as np
import networkx as nx
import requests
from bs4 import BeautifulSoup
```

```
def compute_pagerank(url, num_iterations=100, d=0.85):
    response = requests.get(url)
    soup = BeautifulSoup(response.content, "html.parser")
    links = []

    # First round
    for link in soup.find_all("a"):
        href = link.get("href")
        if href and href.startswith("http"):
            links.append(href)
    G = nx.DiGraph()
    G.add_nodes_from(links)

    # Here we go again
    for i, u in enumerate(links):
        response = requests.get(u)
        soup = BeautifulSoup(response.content, "html.parser")
        for link in soup.find_all("a"):
            href = link.get("href")
            if href and href.startswith("http") and href in links:
                G.add_edge(u, href)

    pr = nx.pagerank(G, alpha=d, max_iter=num_iterations)
    return pr

pr = compute_pagerank("https://www.team-bhp.com/")

for url, score in pr.items():
    print(url, score)
```

OUTPUT:

```
https://www.team-bhp.com/ 0.00782992767585498
https://www.team-bhp.com/forum/ 0.07463188563968107
```

```
https://www.team-bhp.com/hot-threads 0.06399958870407477
https://www.team-bhp.com/news 0.0395026543146141
https://www.cars24.com/buy-used-car?utm_source=teambhp&utm_medium=head
er_nav_bar&utm_campaign=desktop 0.04748246777845908
https://oriparts.com/?utm_source=team-bhp&utm_medium=link&utm_campaign=
main&back_url_id=https%3A%2F%2Fboodmo.com%2f Catalog%2
Part-p-%7Bitem_id%7D%2F&back_url_pn=https%3A%2F%2Fboodmo.com%2Fsearc
h%2F%7Bpn%7D%2F 0.04748246777845908
```

CONCLUSION: PageRank algorithm is one of the main algorithms that links different websites over the internet, making the work of search engines highly efficient. Without such algorithms, the internet won't be the way it is. In this experiment, we have implemented the PageRank algorithm using python.