Subrato — 60 — DI6AD

BDA - Experiment 7

> Aim: To study and implement page rank algorithm using Pyspark.

> Theory:
- Page Rank is an algorithm used by Google to rank web pages in search engine results.
- Web pages are considered as nodes in a directed graph, and hyperlinks b/w pages represent edges.
- Page Rank of a web page depends on the number and quality of links pointing to it.

- Formula for Page Rank:

$$PR(A) = \frac{1-d}{N} + d \sum \frac{[PR(T_i)]}{L(T_i)}$$

N: Total number of pages
d: damping factor
$PR(T_i)$: Page Rank score of page T_i
$L(T_i)$: Number of outbound links of page T_i

- Each page distributes a fraction 'd' of its page rank to pages it links to. '1-d' is distributed evenly among all pages.

> Conclusion: Page Rank Algorithm has been implemented using PySpark.