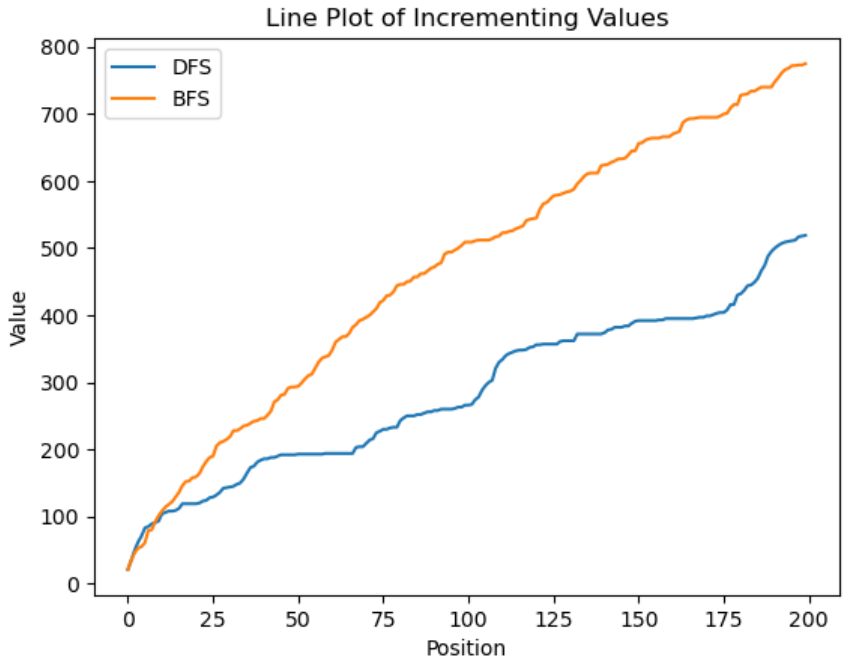
Session 1 Questions

(a) Explain why, having explored the same number of nodes, the order of the two graphs (gB and gD) differs.

The order of graph B is 780 while the order of graph D is 525.

Both graphs share: the same number of crawled nodes and the initial starting node. However, the key difference lies in the crawler's strategy employed. In Graph B, a breadth-first search approach is utilized, where the crawler systematically explores all neighbors at the same level before delving deeper. Conversely, Graph D adopts a depth-first search technique, where the crawler traverses as far as possible along each branch before backtracking.

When analyzing the number of nodes retrieved at each iteration using the Spotify function sp.artist_related_artists(artist_id)["artists"], we observe that it consistently returns 20 friends. To examine the node increment in the graphs using both strategies, we created a plot.



The plot reveals that the BFS strategy demonstrates a linear increase in the number of nodes. In contrast, the DFS approach deviates from linearity.

This disparity arises because, at certain depths, when using DFS the majority of nodes retrieved by sp.artist_related_artists(artist_id)["artists"] are already present in the graph, and thus not added. In these cases, we don't add as many new nodes compared to BFS, but the crawled nodes increment by one. On the other hand, BFS effectively avoids excessive node repetition as it expands across different branches.

(b) Justify which of the two graphs should have a higher order

As explained before, the graph with higher order should be the graph B.

(c) Explain what size the two graphs should have.

The size of the two graphs, in terms of the number of edges, should be 4000 each. This is because for each iteration of the crawler, we retrieve 20 related artists, resulting in 20 new edges being added to the graph. Regardless of the crawling strategy used (BFS or DFS), the number of related artists retrieved remains the same.

It's important to note that even though we may not add 20 new nodes at each depth of the algorithm due to some nodes already being present in the graph, we still add the corresponding edges. In the case of a directed graph, the edge connecting the newly crawled node to the previously seen node is added in the opposite direction. This ensures that all relevant connections are represented in the graph.
Therefore, considering the consistent retrieval of 20 related artists per iteration and the addition of corresponding edges, both graphs should have a size of 4000 edges each.

2. Indicate the minimum, maximum, and median of the in-degree and outdegree of the two graphs (gB and gD).

	Graph B	Graph D
Minimum in-degree	0	0
Maximum in-degree	34	58
Median in-degree	3	4

3. Indicate the number of songs in the dataset D and the number of different artists and albums that appear in it.

Number of Songs	Number of artists	Number of different albums
3643	382	2050

(a) Explain why the number of artists is between 200 and 400, considering the input graphs.

Both the graph created using Breadth first search and depth first search contain 200 nodes with an out degree bigger than 0. Therefore 3 scenarios can happen:

1. All nodes of the two graphs refer to the same artists meaning that the total number of artists is equal to 200.
2. All nodes of the two graphs are different. Meaning that the total number of artists is equal to 400.
3. There are some nodes in both graphs that repeat, in this case the total number of artists is between 200 and 400.

In our case the number of artists is 382 which means that 18 nodes are the same in both graphs.

(b) Justify why the number of songs you obtained is correct, considering the input graphs.

The pandas dataset comprises a total of 3643 songs. The Spotify API provides information for approximately 10 songs per artist, considering the 382 artists in the dataset. This implies that the maximum number of songs we could have is 3820. However, we currently have 3643 songs because there are instances where different artists collaborate on the same song. As a result, our dataset contains 177 songs where multiple artists from our graphs are credited as creators.

(c) Justify why the number of retrieved albums is correct.

Out of the data collected, we have successfully retrieved a total of 2050 albums. This number is lower than the total count of songs in our dataset. The reason for this disparity is that albums generally contain multiple songs, resulting in a higher count of individual songs compared to the count of albums.