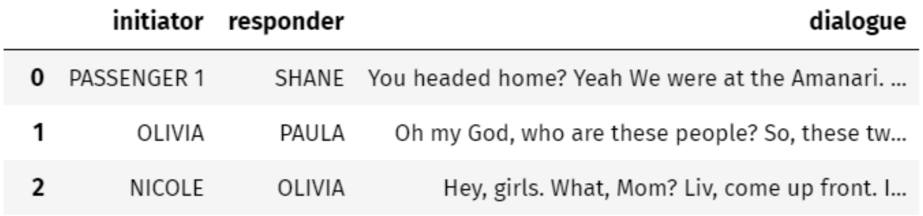
Joram Mutenge
Progress report

Partitioning the project

I have divided the project into three parts. Part 1 involves data collection and creation of the dataset. Part 2 involves calculating centralities of the nodes in the network and creating visualizations of the network. Part 3 involves performing natural language processing (NLP) on the dialogue between the nodes.

About the dataset

To test the feasibility of the project performed preliminary analysis on a complete dataset for episode 1 of *The White Lotus*. Below is a sample dataset.



	initiator	responder	dialogue
0	PASSENGER 1	SHANE	You headed home? Yeah We were at the Amanari. ...
1	OLIVIA	PAULA	Oh my God, who are these people? So, these tw...
2	NICOLE	OLIVIA	Hey, girls. What, Mom? Liv, come up front. I...

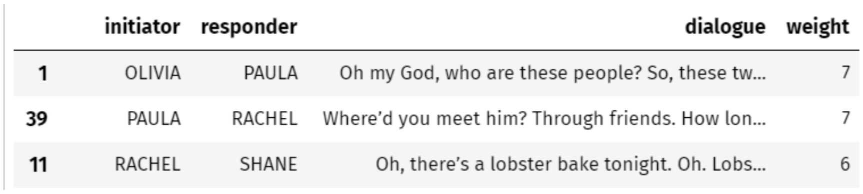
The column `initiator` represents the nodes (characters) who start the conversation while `responder` represents the nodes who respond back. The `dialogue` stores all the conversations between the initiator and the responder throughout the episode. For instance, the dialogue for Olivia and Paula will be all the words exchanged between them from the beginning to the end of the episode.

Data cleaning

Since I will be attempting to do some natural language processing on the data in the dialogue column, I had to do something data cleaning. For example, given the nodes Armond and Lani with multiple instances of their conversations, I had to collect all those instances and combine them into a single instance represented by a row. This ensured that all their dialogue was in a single cell. Also, to make this work I had to treat `‘initiator = Armond’` and `‘resonder = Lani’` the same as `‘initiator = Lani’` and `‘resonder = Armond’`. That is because it doesn't matter who starts the conversation, for it still involves those two people.

Data analysis

A new column `weight` was feature engineered to represent the frequency of the node pairing in the data set. Below is the sample dataset.

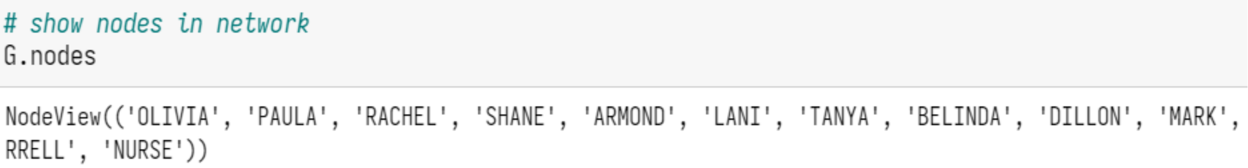


	initiator	responder	dialogue	weight
1	OLIVIA	PAULA	Oh my God, who are these people? So, these tw...	7
39	PAULA	RACHEL	Where’d you meet him? Through friends. How lon...	7
11	RACHEL	SHANE	Oh, there’s a lobster bake tonight. Oh. Lobs...	6

We can see that the pairing Olivia and Paula have a weight of 7 while the pairing Rachel and Olivia have a weight of 6. This means that in episode 1 of *The White Lotus*, the interactions between Olivia and Paula outnumbered those between Rachel and Shane.

Network creation

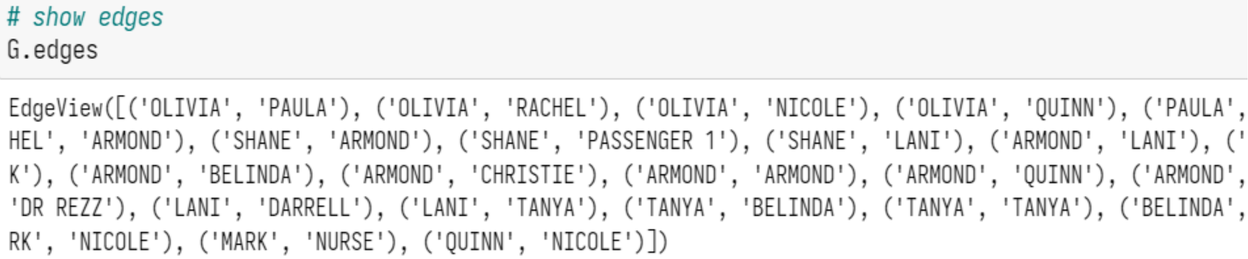
I used the Python package NetworkX to create the nodes and edges of the network. Below is an extract of the nodes in the network.



```
# show nodes in network
G.nodes
```

```
NodeView(('OLIVIA', 'PAULA', 'RACHEL', 'SHANE', 'ARMOND', 'LANI', 'TANYA', 'BELINDA', 'DILLON', 'MARK',
RRELL', 'NURSE'))
```

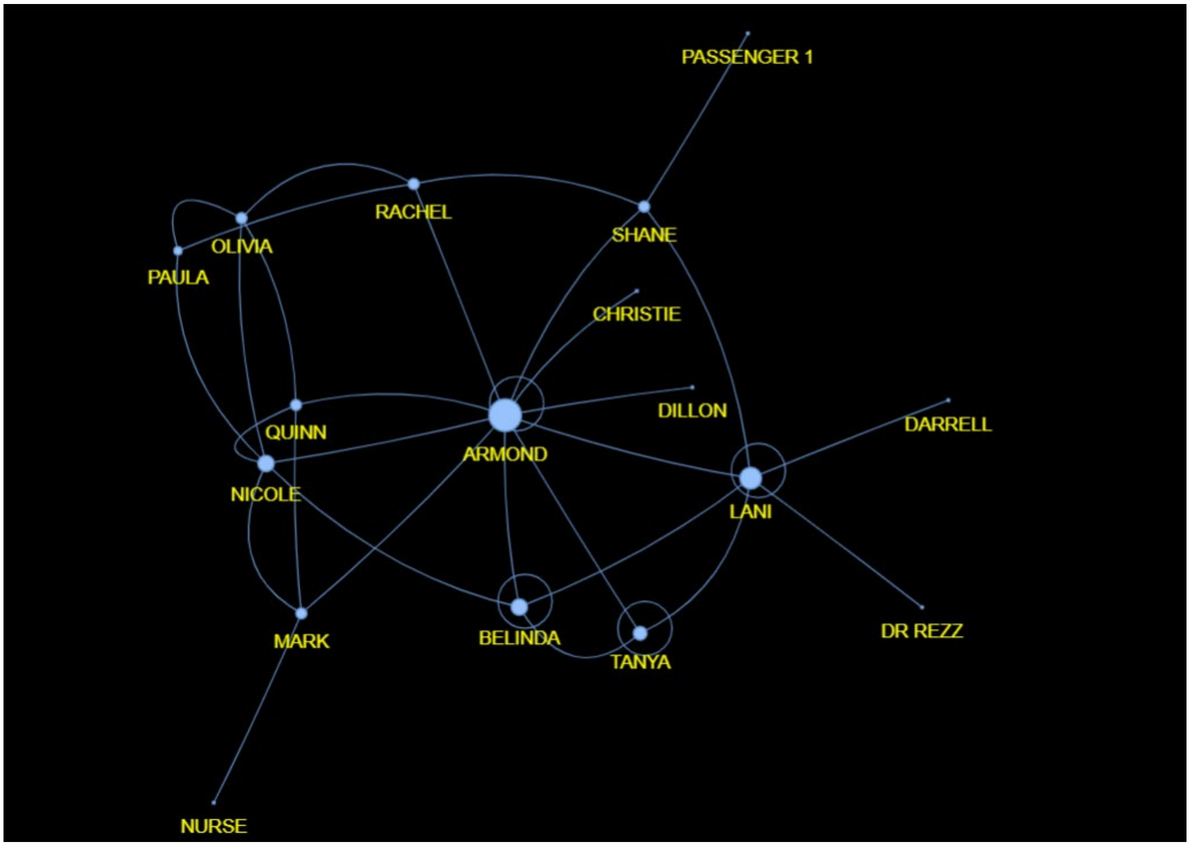
Here is an extract at the edges.



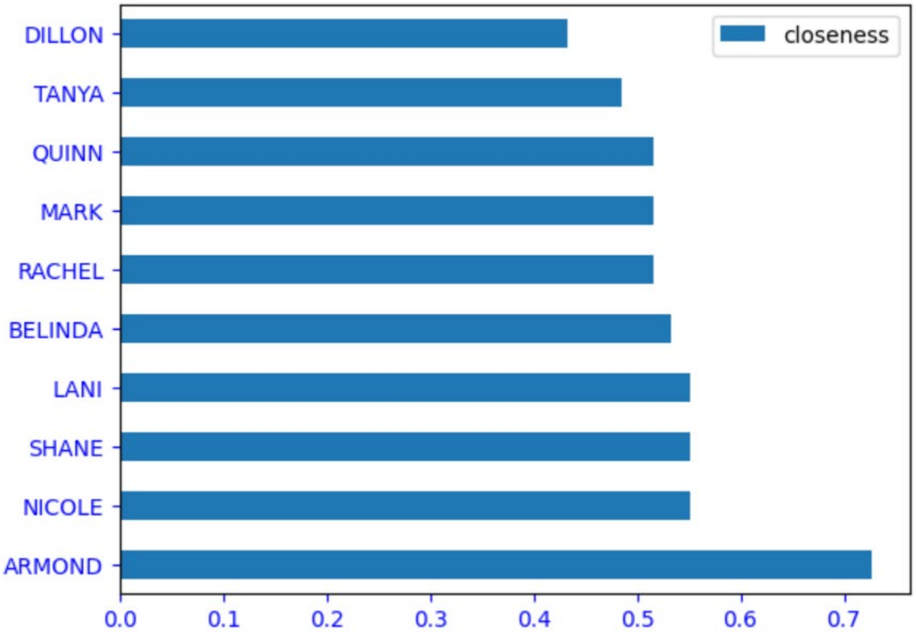
```
# show edges
G.edges
```

```
EdgeView([('OLIVIA', 'PAULA'), ('OLIVIA', 'RACHEL'), ('OLIVIA', 'NICOLE'), ('OLIVIA', 'QUINN'), ('PAULA',
HEL', 'ARMOND'), ('SHANE', 'ARMOND'), ('SHANE', 'PASSENGER 1'), ('SHANE', 'LANI'), ('ARMOND', 'LANI'), ('
K'), ('ARMOND', 'BELINDA'), ('ARMOND', 'CHRISTIE'), ('ARMOND', 'ARMOND'), ('ARMOND', 'QUINN'), ('ARMOND',
'DR REZZ'), ('LANI', 'DARRELL'), ('LANI', 'TANYA'), ('TANYA', 'BELINDA'), ('TANYA', 'TANYA'), ('BELINDA',
RK', 'NICOLE'), ('MARK', 'NURSE'), ('QUINN', 'NICOLE')])
```

The network graph below shows all the edges and nodes contained in episode 1 of *The white Lotus*. The size of the node represents decrease centrality. Thus, Armond has the highest degree in the network, followed by Lani. This means that Armond has interactions with more characters (nodes) than any other character in episode 1.

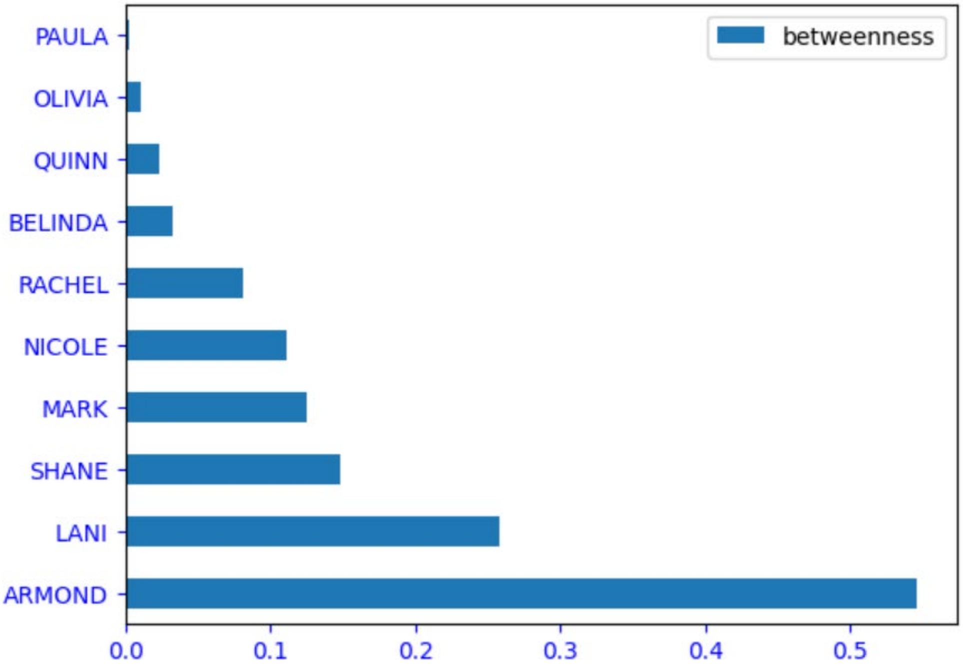


I also calculated other centralities of the network which I have presented as graphs. Below is the closeness centrality graph.

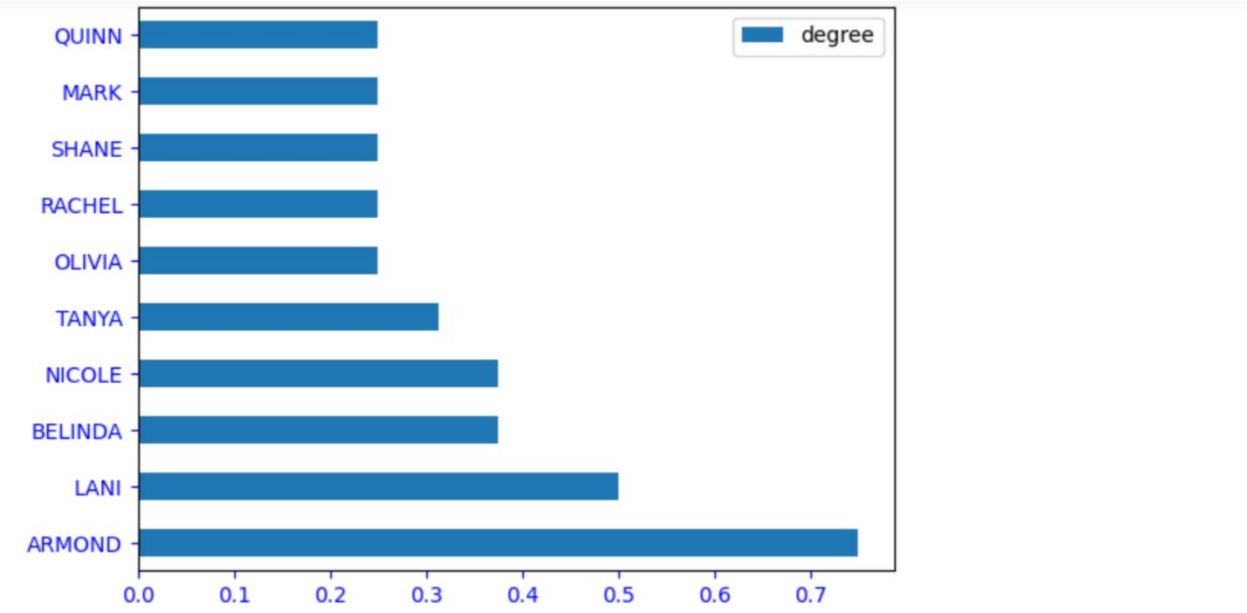


Unsurprisingly, Armond scores high on closeness centrality as well.

For betweenness centrality the difference between the nodes is distinct.



Finally, here is the graph for degree centrality.



These preliminary analyses have shown me that my dataset is sufficient to answer my research question "Who are the most important characters in the TV show *The White Lotus*?"

What's next?

I will perform the analyses presented above on the entire dataset (i.e., season 1 of the TV show), then I will choose the top five characters and further analyze the evolution of their importance throughout the season. I will use centrality measures as the criteria for selecting important characters in the show.

Lastly, I will perform natural language processing on the dialogue of the chosen top five characters to discover topics contained in the conversations they have with other characters.