Lucas Sipos

Problem 1:

```
fruits = {
   'Apple': {"red", "sweet", "round", "juicy", "fruit"},
   'Banana': {"yellow", "sweet", "long", "peel", "fruit"},
   'Orange': {"orange", "round", "juicy", "citrus", "fruit"},
   'Grape': {"purple", "small", "juicy", "peel", "fruit"},
   'Lemon': {"yellow", "sour", "citrus", "peel", "fruit"}
}
```

a.

```
Jaccard Distance Matrix:
[0.0, 0.75, 0.57, 0.75, 0.89]
[0.75, 0.0, 0.89, 0.75, 0.57]
[0.57, 0.89, 0.0, 0.75, 0.75]
[0.75, 0.75, 0.75, 0.0, 0.75]
[0.89, 0.57, 0.75, 0.75, 0.0]
```

b. The triangle inequality holds for all pairs of fruits in the Jaccard Distance matrix, there are no violations.

c. Comparison of Jaccard Distance and Dice Coefficient:

```
Dice Coefficient Matrix:
[1.0, 0.4, 0.6, 0.4, 0.2]
[0.4, 1.0, 0.2, 0.4, 0.6]
[0.6, 0.2, 1.0, 0.4, 0.4]
[0.4, 0.4, 0.4, 1.0, 0.4]
[0.2, 0.6, 0.4, 0.4, 1.0]
```

- Jaccard Distance tends to give larger values when the two sets have fewer shared features relative to their union. It's stricter, as it considers the total number of features in both sets.
- Dice Coefficient generally gives lower values than Jaccard Distance for the same comparisons because it emphasises the size of the intersection relative to the combined size of both sets, making it a bit more relaxed.

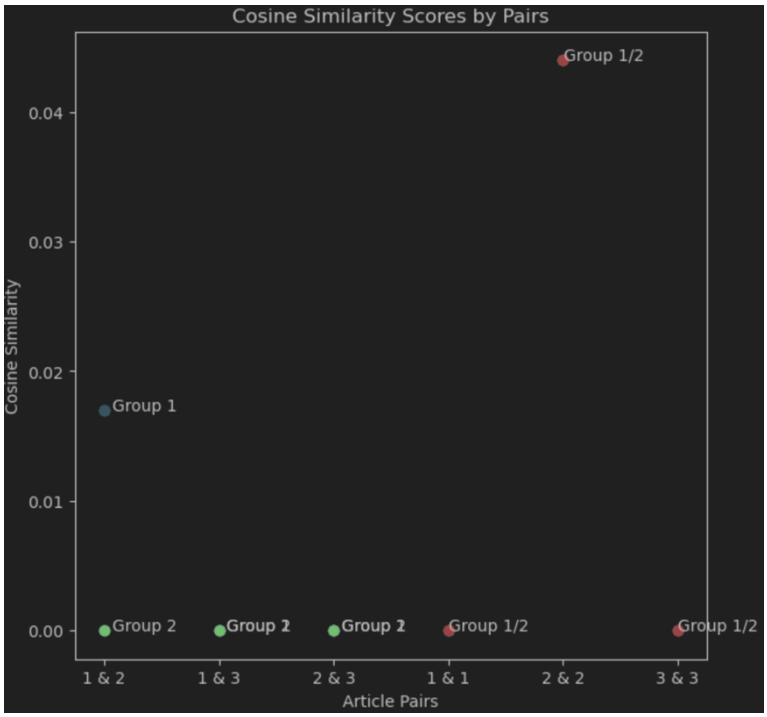
Problem 2:

a. group1_articles = [
   ("UN Climate Change 1", "The COP28 summit aimed to bring together world leaders to accelerate climate action."),

("UN Climate Change 2", "Countries are working to phase out coal and other fossil fuels to meet climate goals."),
("UN Climate Change 3", "The conference in Dubai focused on finding sustainable solutions for global warming.")
]

group2_articles = [
("Paris Agreement climate action 1",
"The Paris Agreement set a goal to limit global temperature rise to 1.5 degrees Celsius."),
("Paris Agreement climate action 2",
"Countries agreed to regularly report their greenhouse gas emissions under the agreement."),
("Paris Agreement climate action 3",
"The pact encourages nations to take ambitious climate action to meet environmental targets.")
]

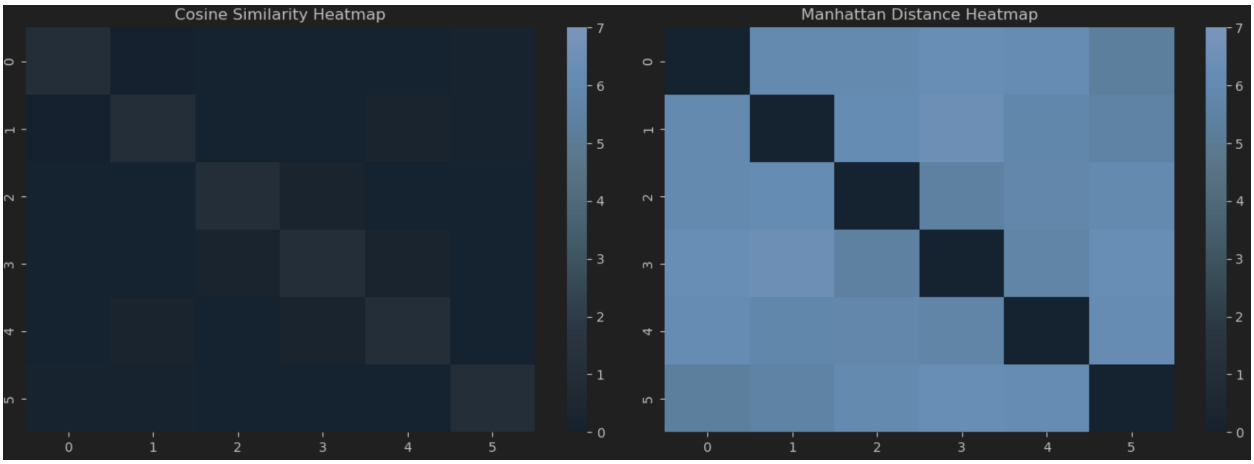
b. We can see that most of the articles do not have clustering because after stripping the stop words, the sentences differ from each other. But let me talk about the Group 1/2 Art. 2 & 2 and Group 1 Art. 1 & 2:



The difference between Group 1 and Group 1/2 is when calculating the tf-idf, the score of their common words differs:
Group 1/2: 'countries': 0.4771, 'countries': 0.4771,
Group 1: 'climate': 0.301, 'climate': 0.301
As you can see the “countries” word has a bigger tf-idf score than “climate”.

c. The scores I have found on the sklearn method are not the same as the ones from the Cosine.py program.

Lucas Sipos



After looking through these heatmaps, I can tell that:

- Both maps have the same diagonal pattern; the difference is that one of them is 1 (Cosine similarity) because A = A, and the other one is 0 (Manhattan distance) because the distance between 2 dots that are on the same coordinates is equal to 0
- Both heatmaps are relatively symmetric
- There are noticeable differences in the specific values and patterns of the heatmaps, especially for items that are not very similar or very different.