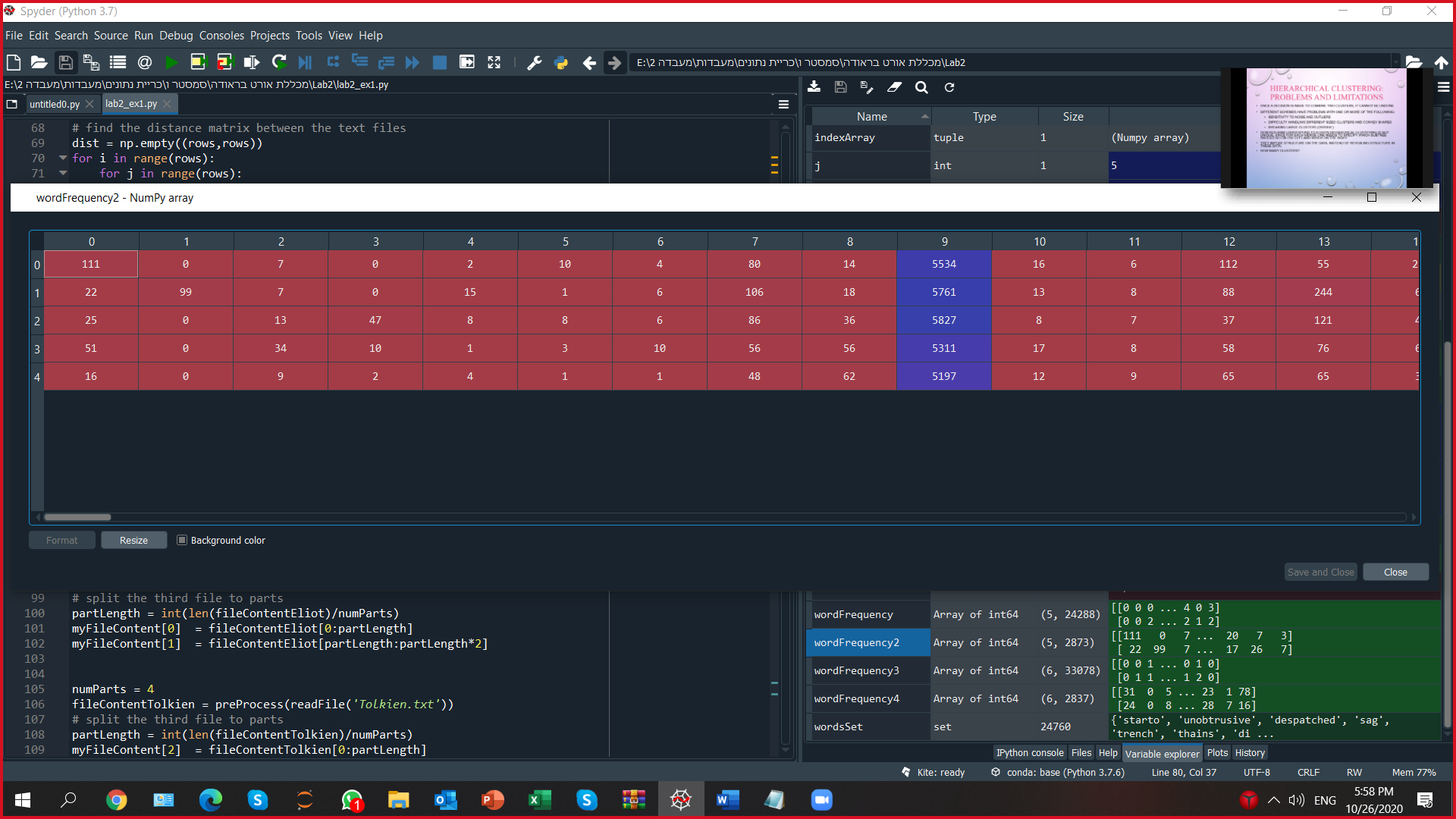
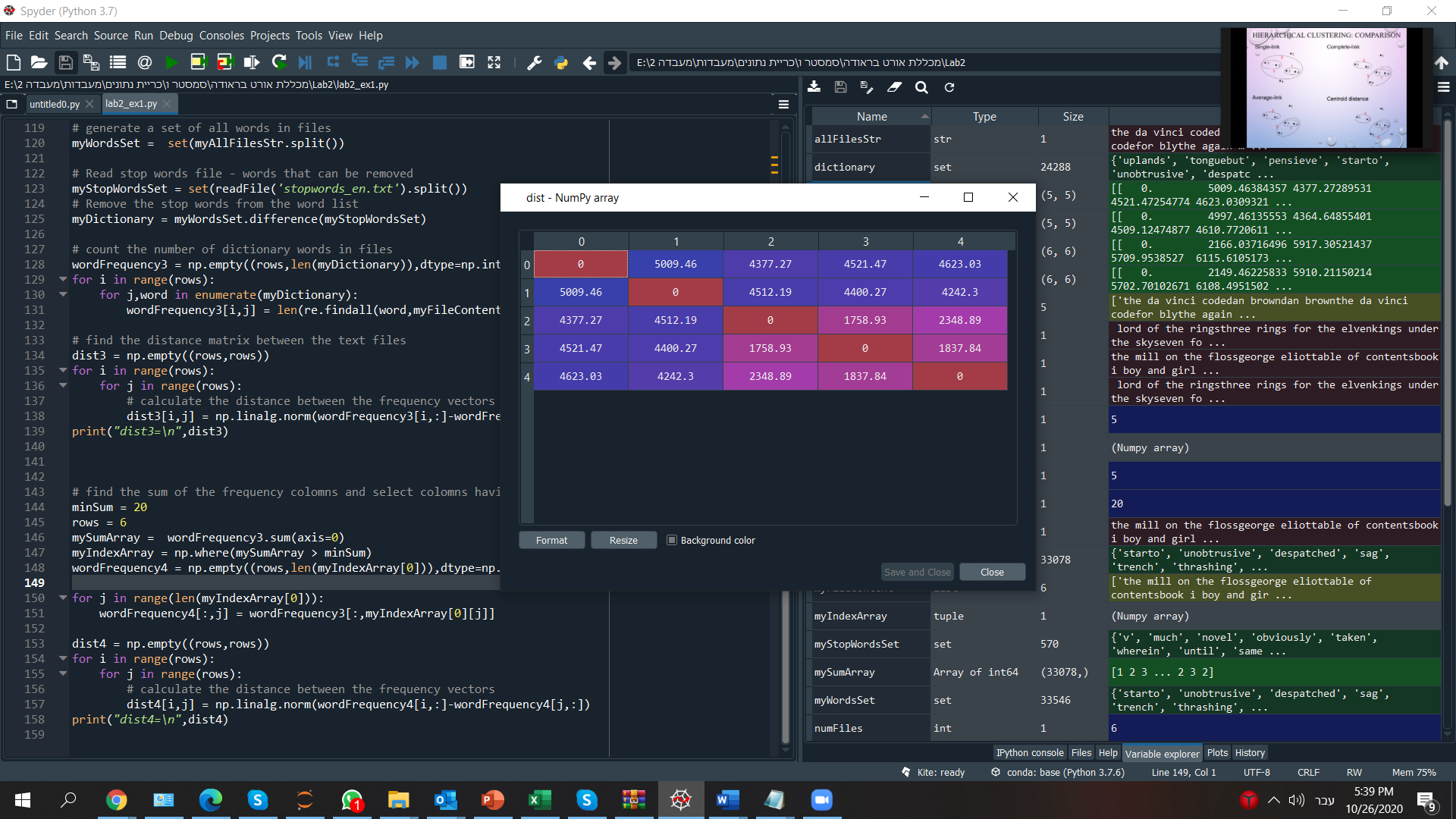
**מעבדה 2 – כריית נתונים**

שם מגיש: אילון בן סימון ת"ז: 312162951

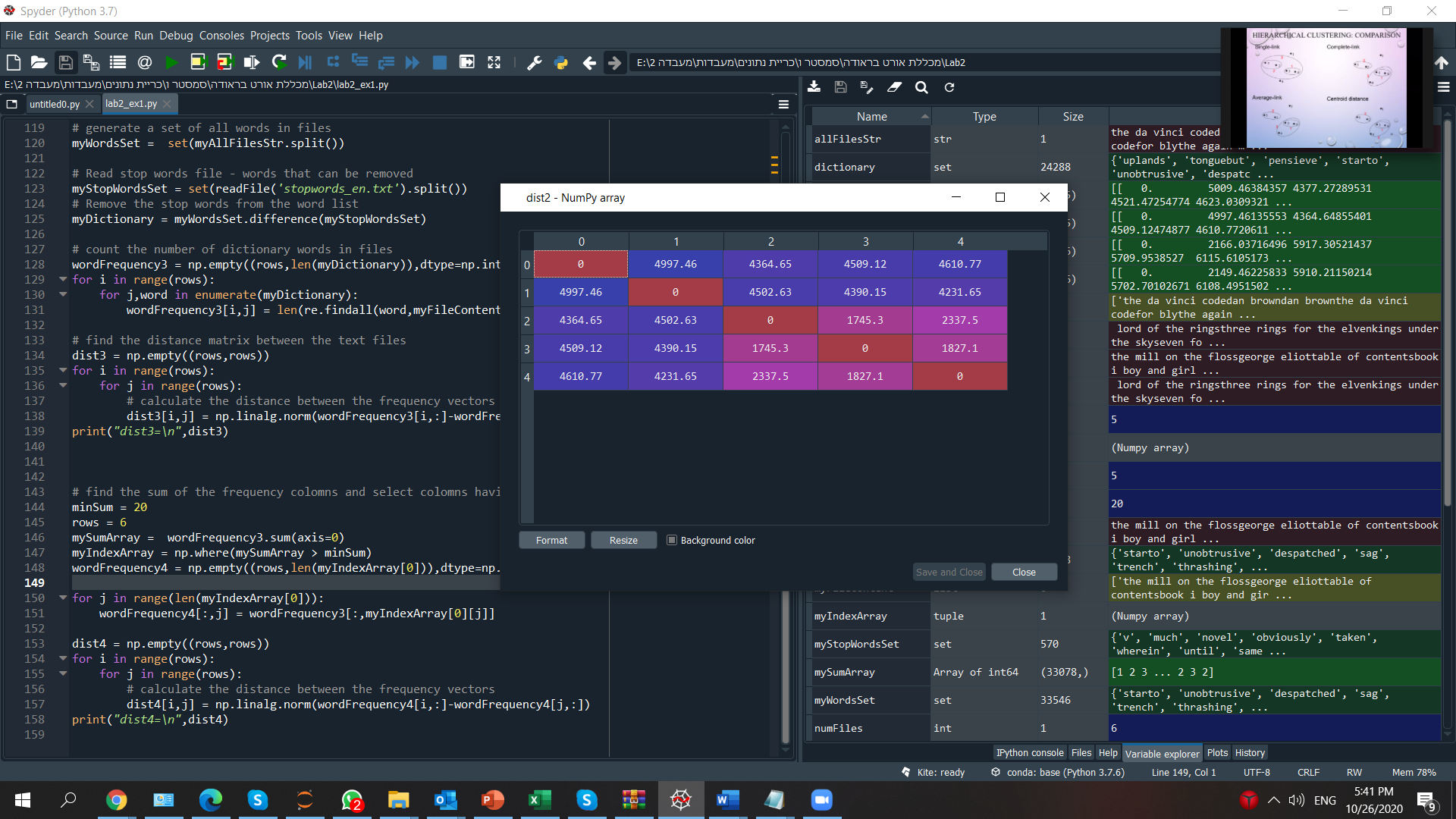
שם מגיש: סער ויקטור ת"ז: 312392822

1. To decrease the frequency matrix dimension build the new frequency matrix **wordFrequency2** in according to the condition (**sumArray**>20). Present results in in the Word file **result.docx**.
2. Analyze the texts similarity using distance matrices **dist** (built from **wordFrequency**) and **dist2**(built from **wordFrequency2**)**.** Write results of analysis in **result.docx** (What is the meaning of distance matrix values? Which distance matrix gives better results?) .

Dist:



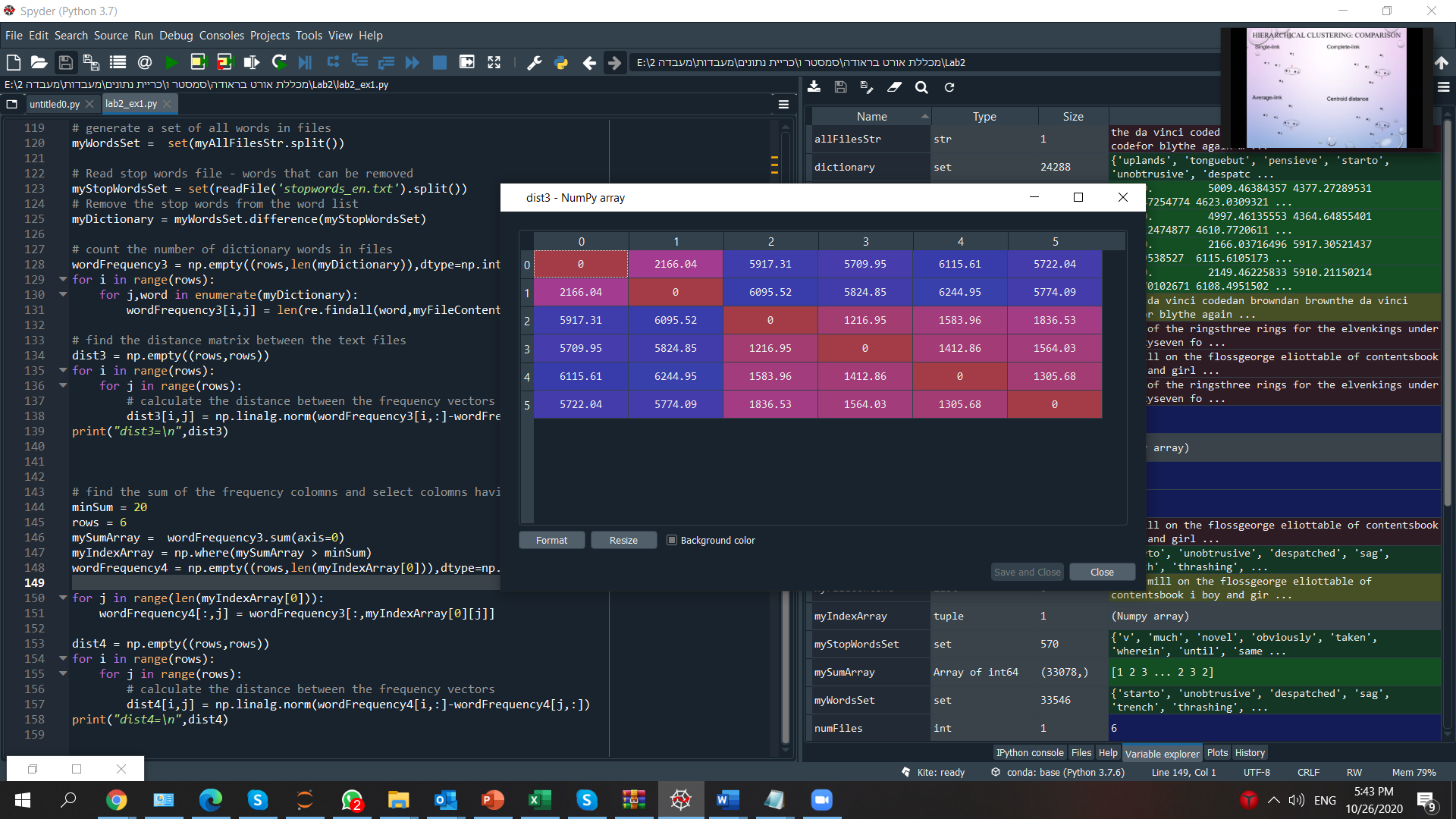
Dist2:



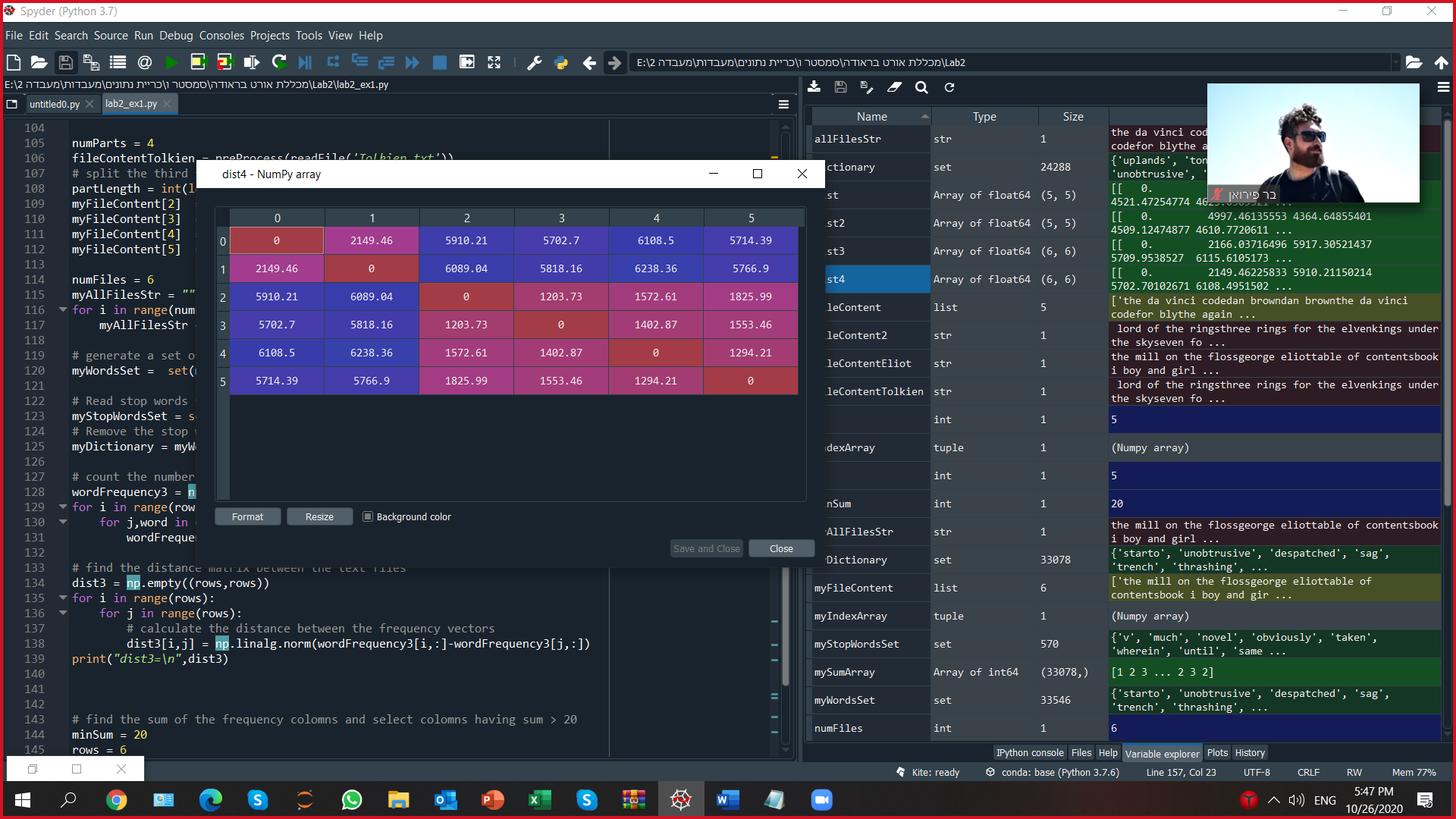
The meaning of distance matrix values is how similar one text to another. The distance matrix that gives better results is the one with the words that appear more than 20 times in all the texts combined. The results are better but not significantly.

1. Analyze the texts similarity of the texts 'Eliot.txt'(divide into 2 parts) and 'Tolkien.txt' (4 parts). Write results of analysis in **result.docx.**

Dist3:



Dist4:



The case here is the same, the new matrix gives better results but not significantly.