LDA Topic Modeling Report

Name: Victor Kwok

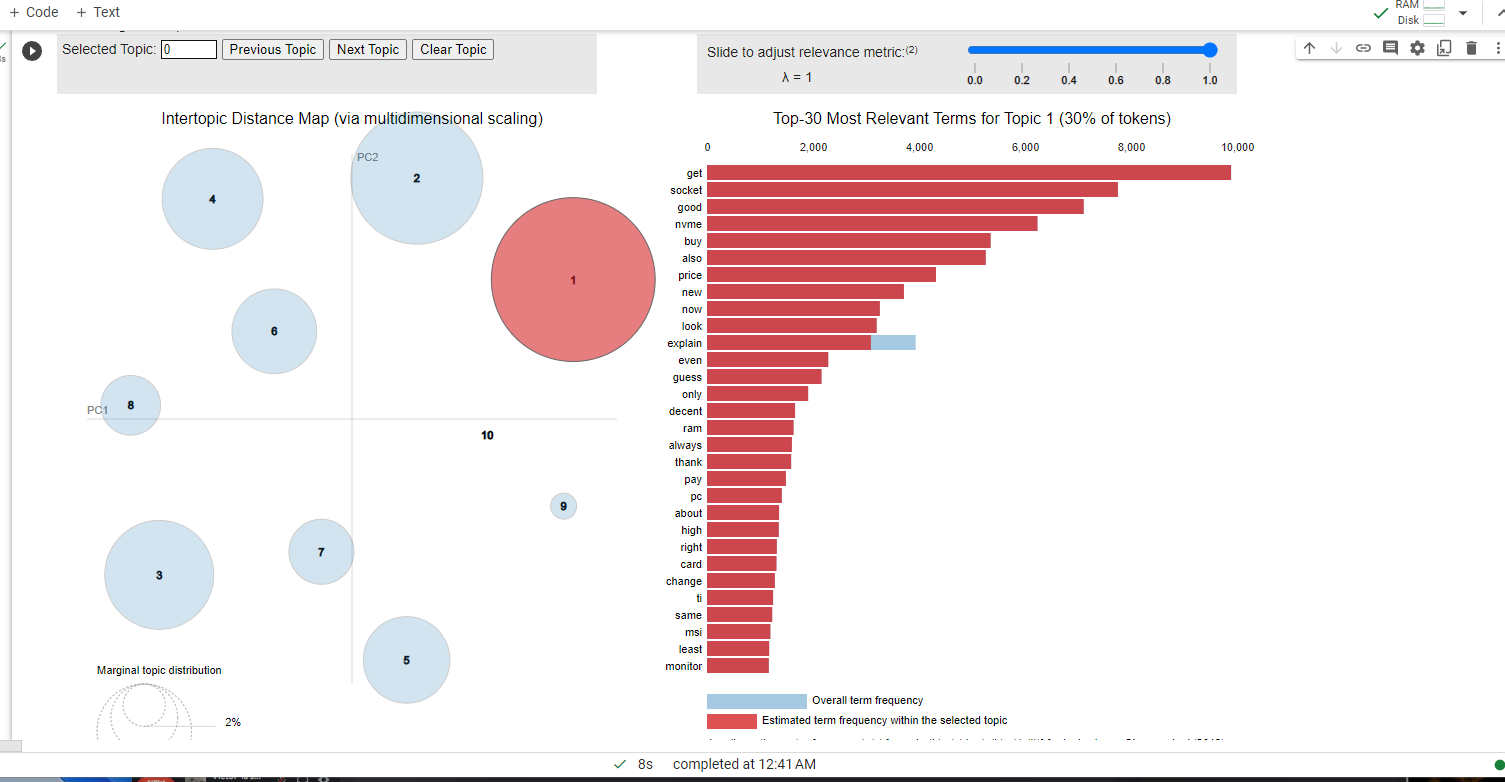
I did Topic modeling using the LDA model to see the topic trend of people’s discussion related to PC components. I grabbed messages from a discord group called Build a pcsale, where people share their thoughts on the recent deals on PC components and what is the current best PC component combination.

Here is a link to my code:

<https://colab.research.google.com/drive/1eaBBQ6lXqW7RFuRXj8uqsn19JZmJq5Ox?usp=sharing>

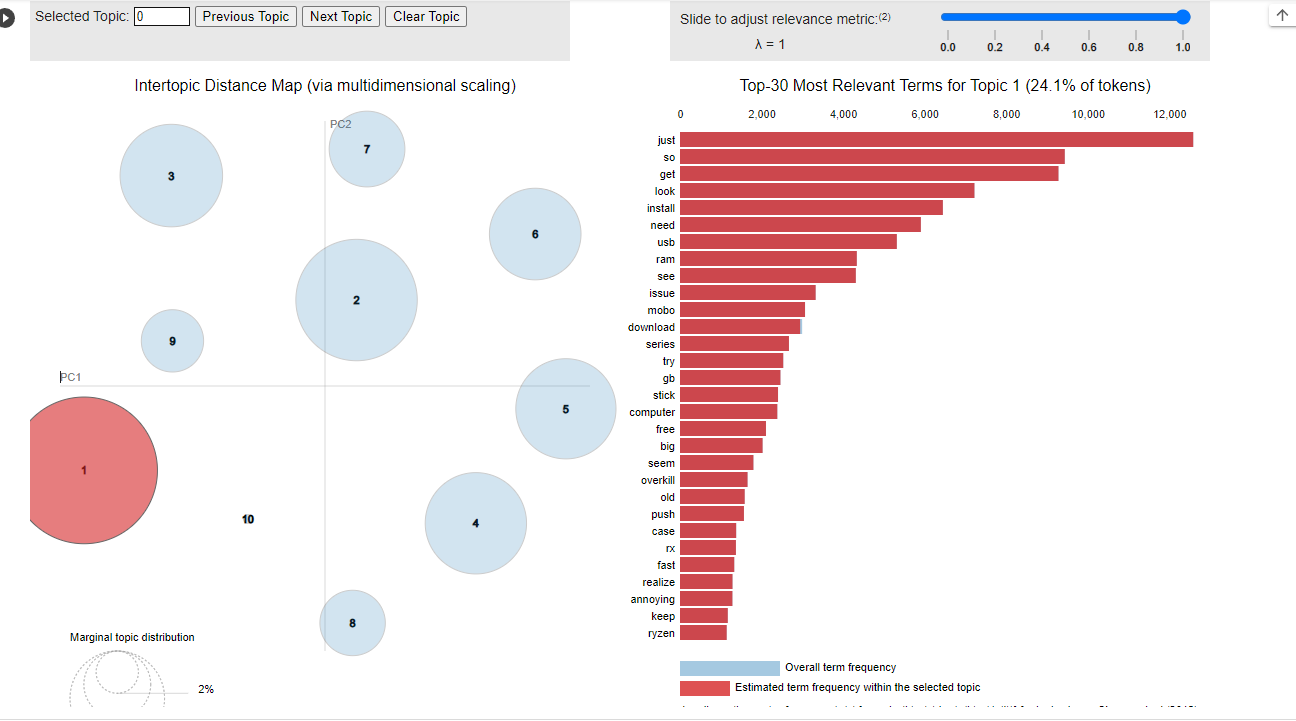
I used an app called Discrub to grab the messages and store the messages into the json. At first I actually intended to grab messages from 2020 to 2023 to see the changes in topic throughout Covid. Unfortunately, grabbing the messages from the discord server took too long as there were around 80k messages sent in 2023 alone. I decided to shorten the range to just 2023 alone and see how the topic of discussion changes throughout the year from Jan 1st of 2023 to Oct 16th 2023. I did a topic modeling on each individual month and here are the results:

January:



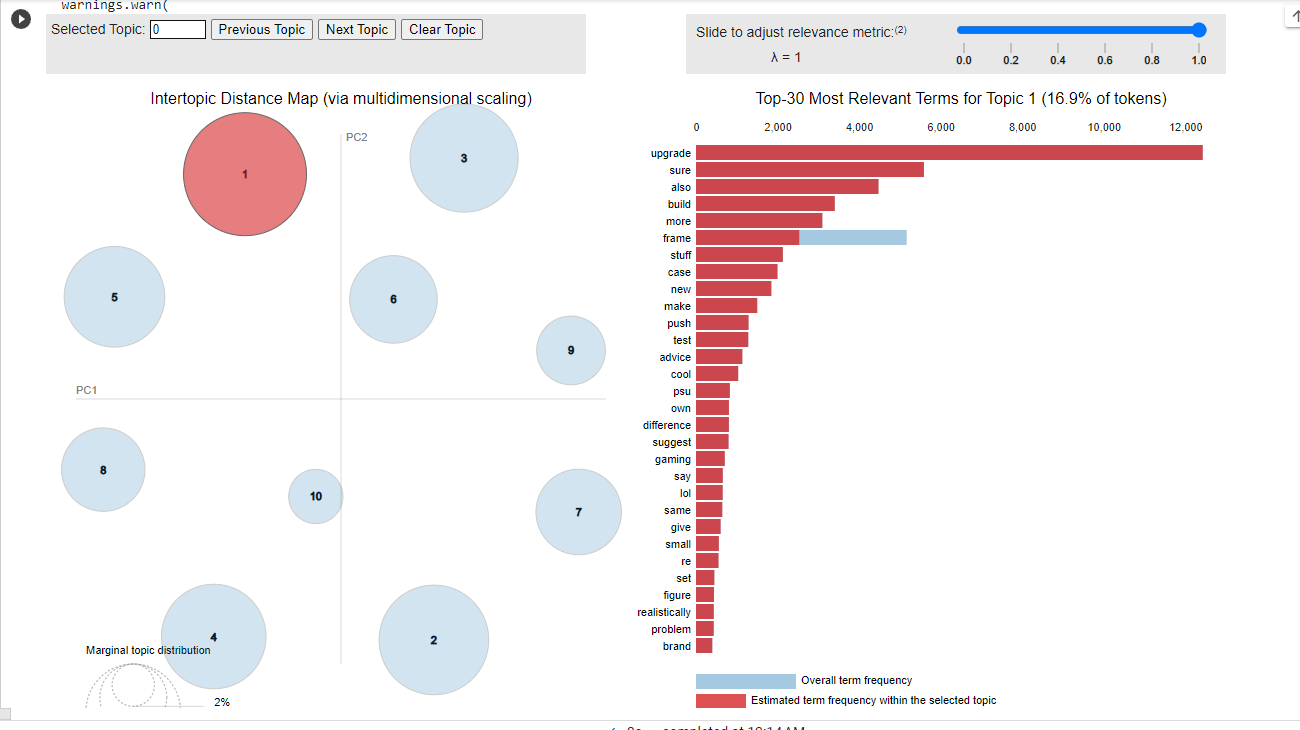
For January the largest topic consist of terms such as get, socket, good, nvme. Nvme stands for

(nonvolatile memory express) and is a type of ssd.

February:

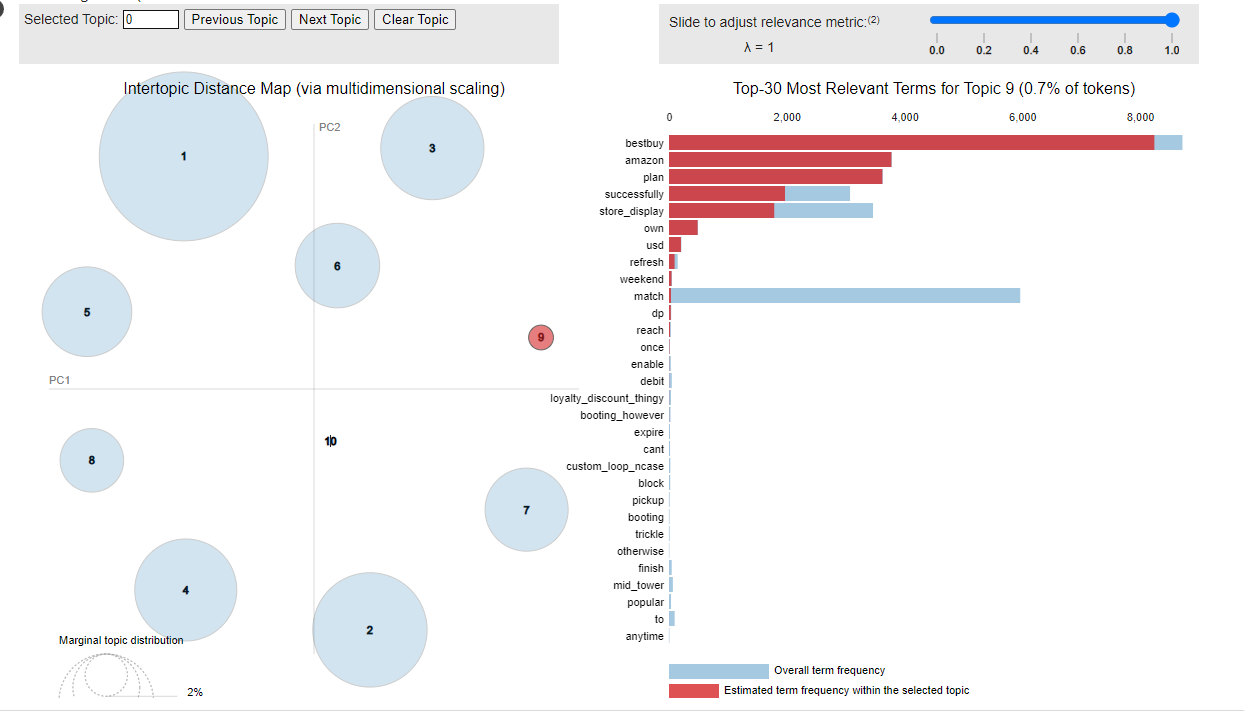
The biggest topic for February consist of terms such as just, so , get , look which are all “useless” term. There are some interesting topics though such as topic number 2 shows terms like deals, gaming and well, which shows that gaming and perhaps deals regarding pc components are being discussed.

March:

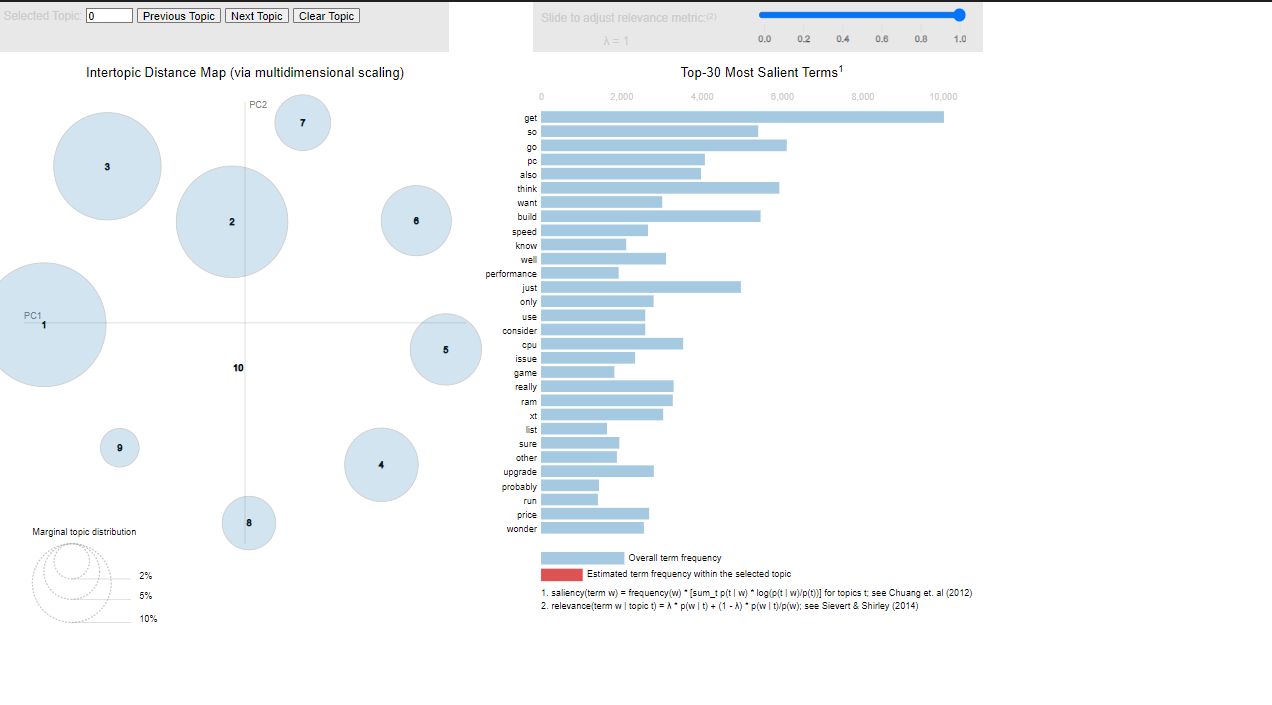


The biggest topic in March seems to be related to upgrading and building. Interestingly, March also has a topic (topic #2), which consists of the word “fan”. This suggests that during this month people were discussing the topic of fans within their pc.

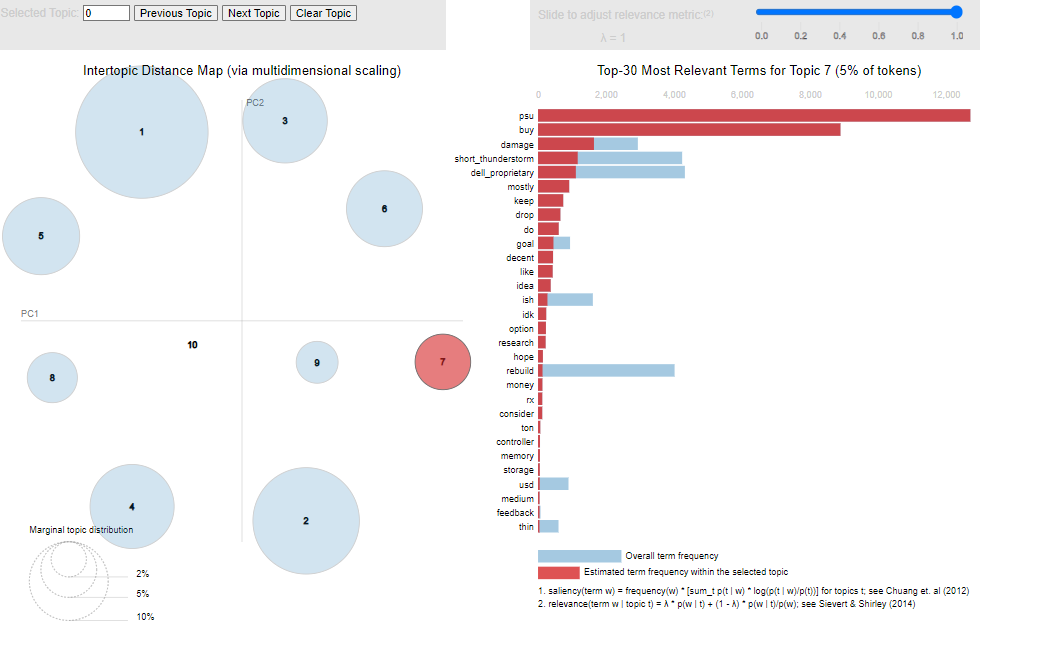
April:



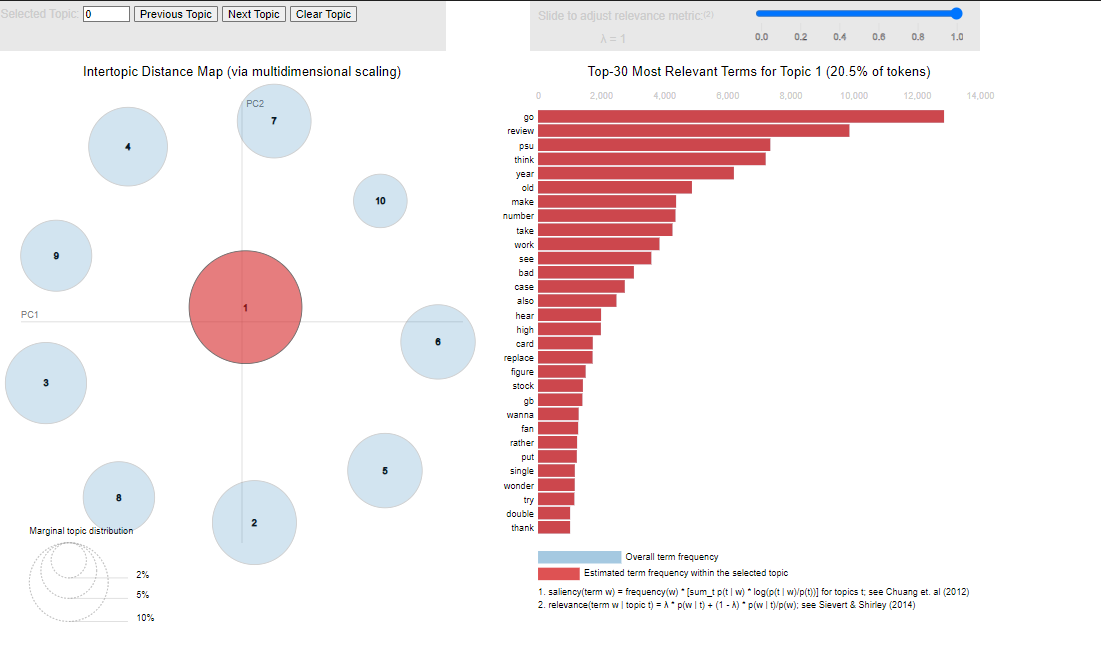
The Biggest topic in April talks about building, pc and budget. Other topics that were discussed this month include, Topic 4: cpu, booting, problem; Topic 2: Bundles; Topic 7: Games. The most interesting topic was topic 9 which was Best- Buy and Amazon.

May:

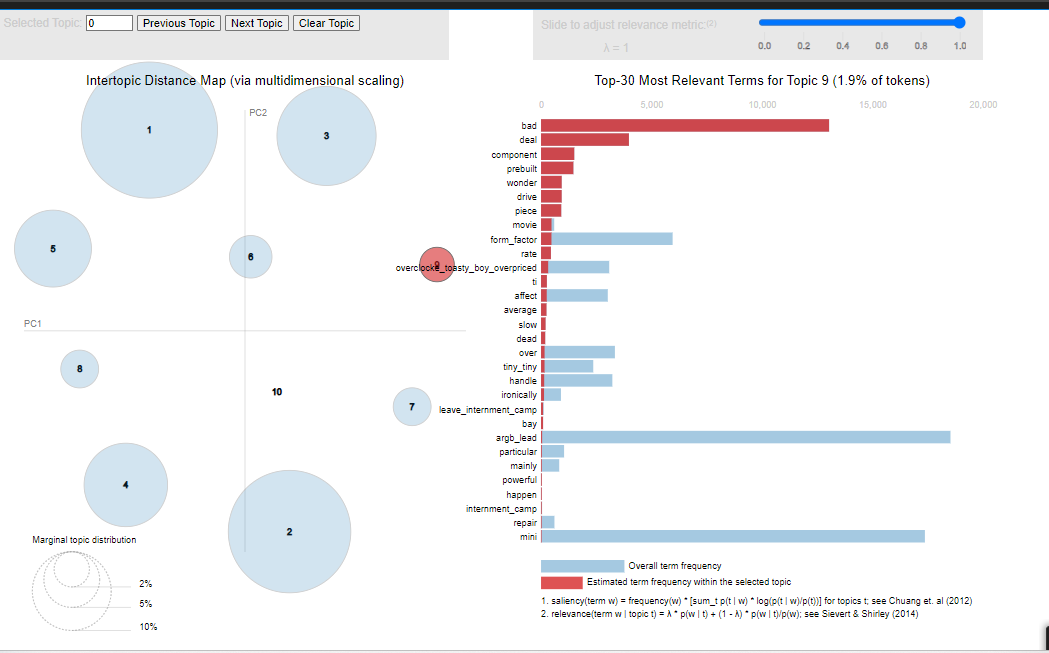
The Biggest topic in May were “useless” term: get, just, need and try. The most interesting topic this month was topic 7, speed and game.

June:

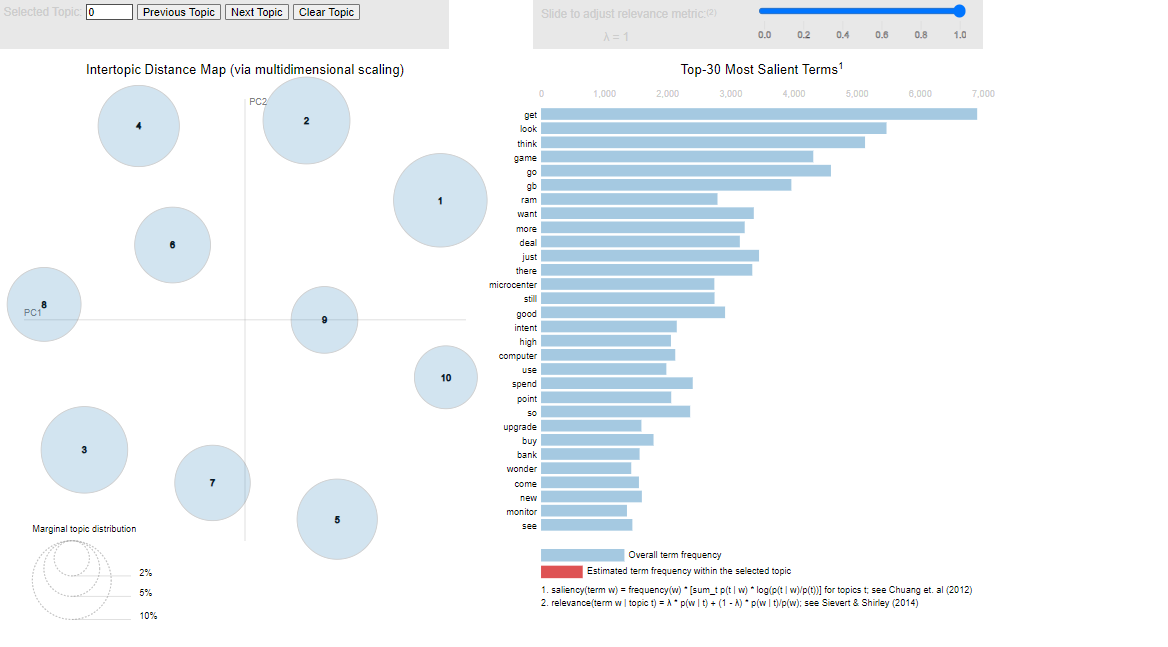
The Biggest topic in June consist of terms: try, know, replace. The most interesting topic this month was topic 7, psu, buy and damage.

July:

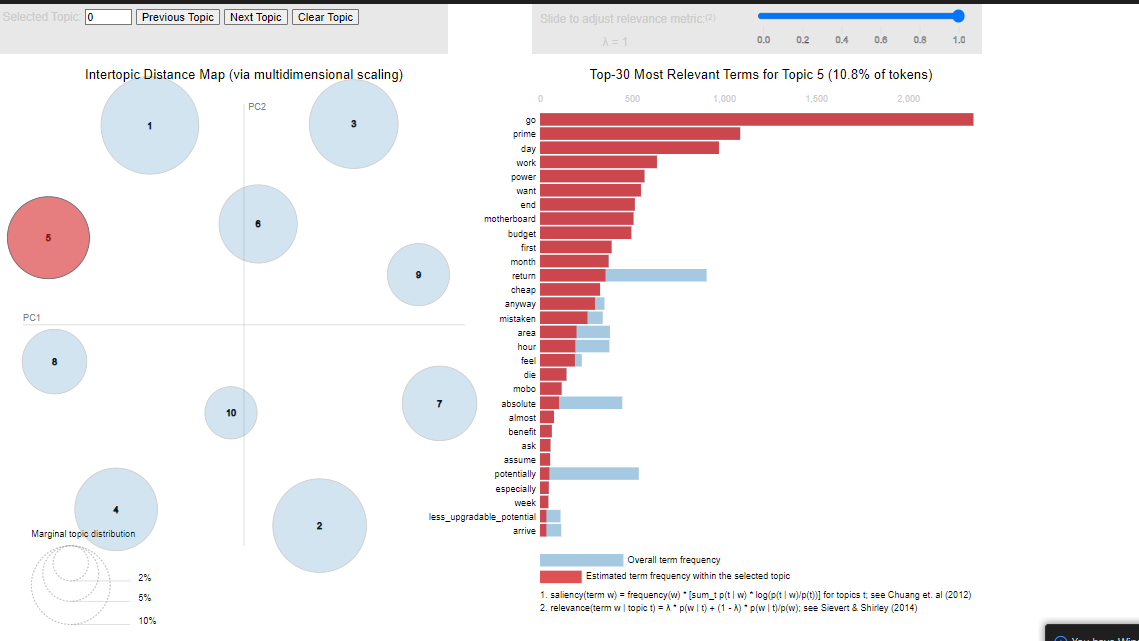
The topics in july make the most sense. The biggest topic consist of terms such as Review, psu. Topic 4 consists of terms such as computer and thermaltake. Topic 9 consist of terms such as budget. Topic 3 consist of terms such as Buy goods. Topic 6 consist of terms such as monitors. Topic 10 consist of the terms such as components.

August:

The Biggest topic in August consist of terms: Total, just, motherboard. The most interesting topic this month was topic 9 which consist of bad deal. It seems that people got bad deals on pc components in August.

September:

September had the most even out topic size. Each topic circle seem to have similar radius on the graph. The biggest topics this month are mainly just useless terms such as: just, there, good, so, issue. However there is first month where microcenter was mentioned in topic 6 which also consist of terms gb and upgrade. There is also topic 10 which talked about ram, high and worth.

October:

This month gave me the most surprising result as it was able to have prime day as its term in one of the topics (topic 5). This is because Amazon hosted its prime day on October 10th and lasted until the 11th.

Limitations:

* In the messages, people refer to the gpu through their Gen number for example, people refer to the GeForce RTX 4090 ti gpu as simply 4090 ti. The model filters the number 4090 or it simply counts the “ti” as a word of its own.
* Because we are using messages from discord, the messages will consist of a lot of “useless” words such as go, so, look, just, try, etc. Was able to get the model to filter out only some of the “useless” words

I have attached a zip folder in my turnin with all the datasets I have used and the Topic modeling for each month if results are not clear in the images.