Week 7

☑ Study the distribution:

The distribution is studied, with the maximum value of 923 (the value that is actually bigger than 0), so I choose the vector length as 923 instead of others.

Choosing the right number of cluster:

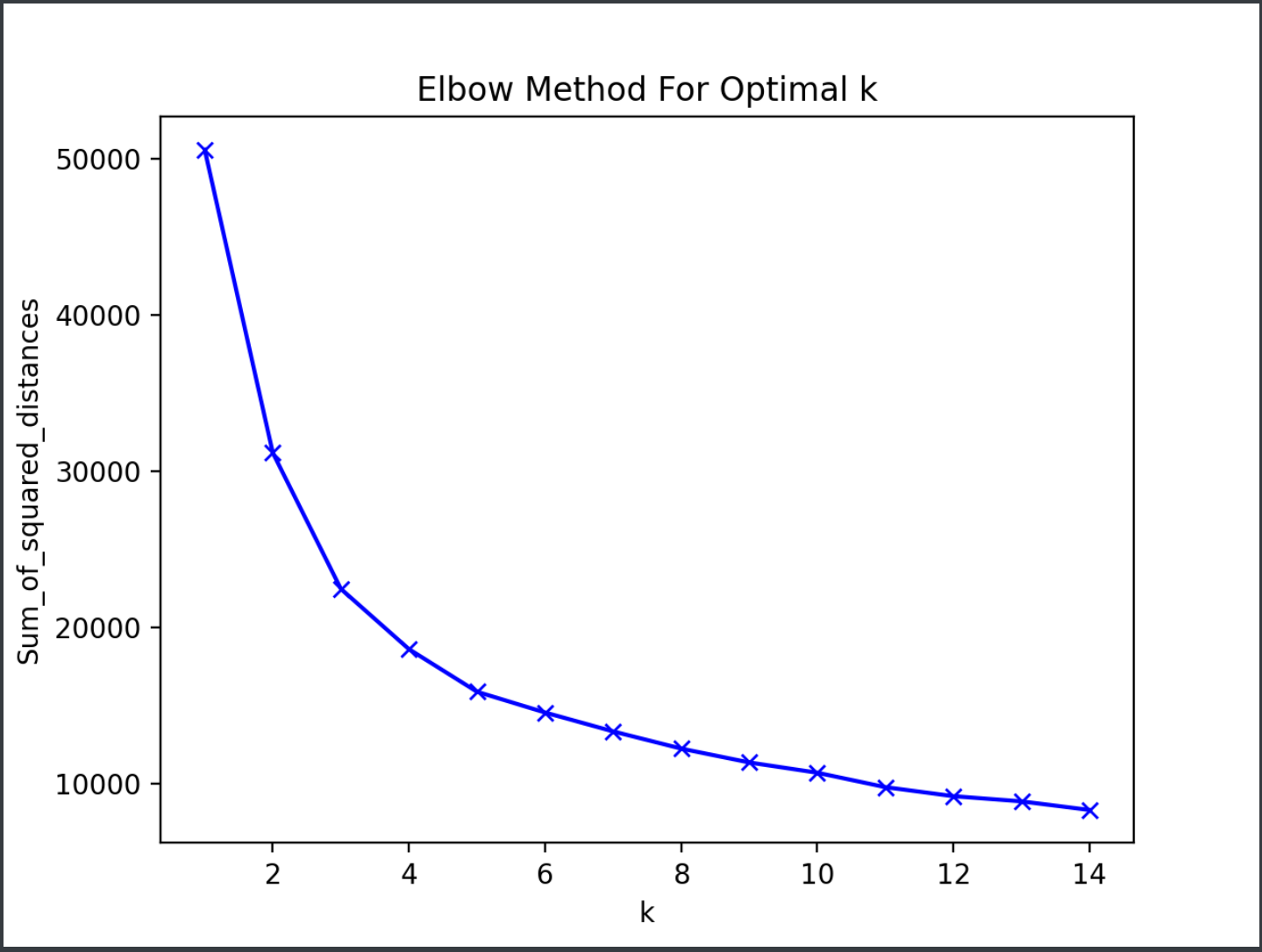
https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/

☑ Discuss events relation with the predicted cluster, plot each cluster average postion vs events...

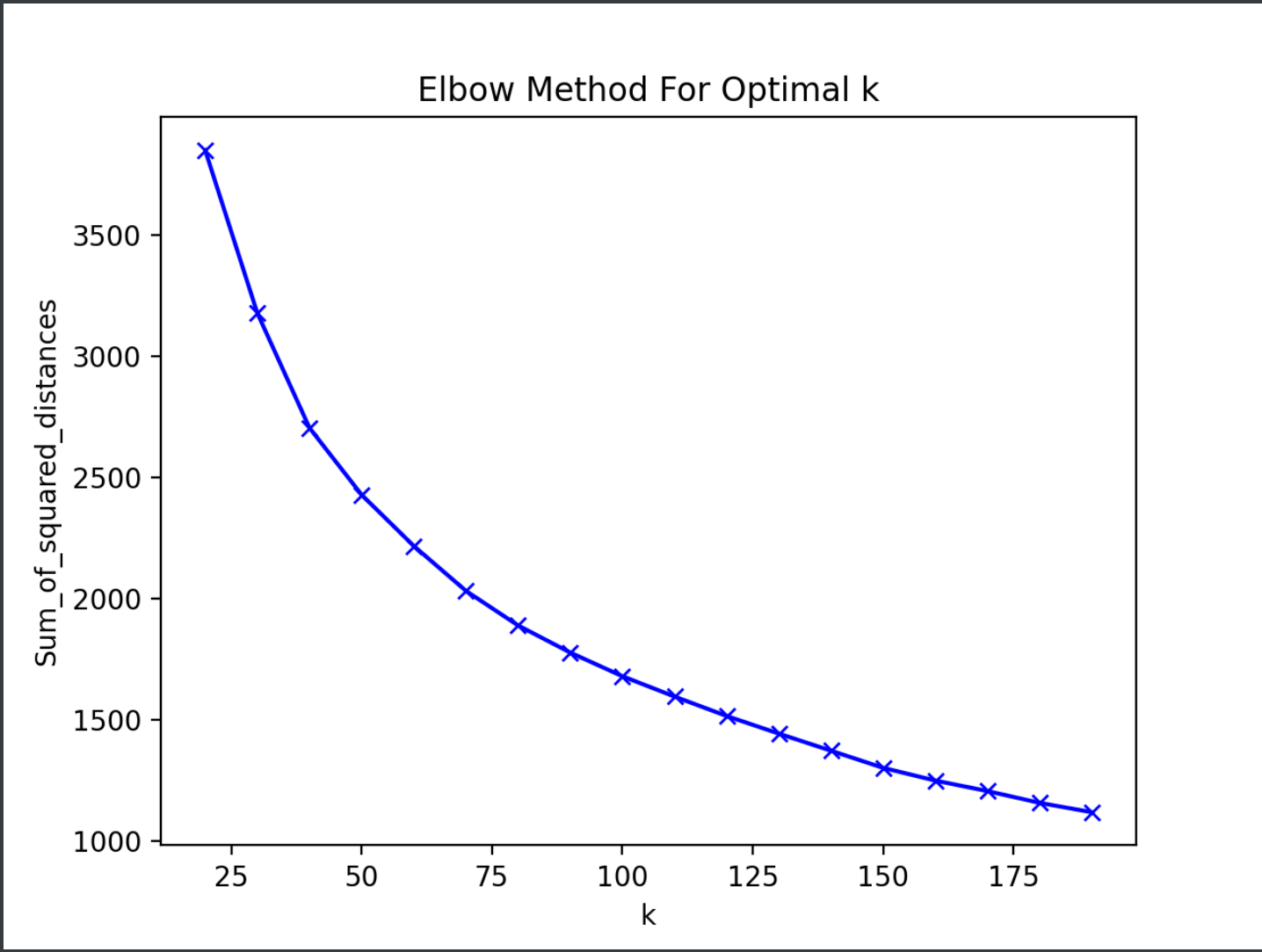
1. I examine the optimal K value of the cluster using the "Gap Statistics":
   Paper of 2001 Stanford: https://web.stanford.edu/~hastie/Papers/gap.pdf
2. The “Elbow Method” for the kmeans cluster https://blog.cambridgespark.com/how-to-determine-the-optimal-number-of-clusters-for-k-means-clustering-14f27070048f
3. The plot of the elbow is:

BTW the optimal K is hard to calculate (as when n_cluster is large the calculation process is really hard..., So one solution is to select the k for that, another is to use one GPU server to calculate the whole optimal K)

For non filter one elbow method:



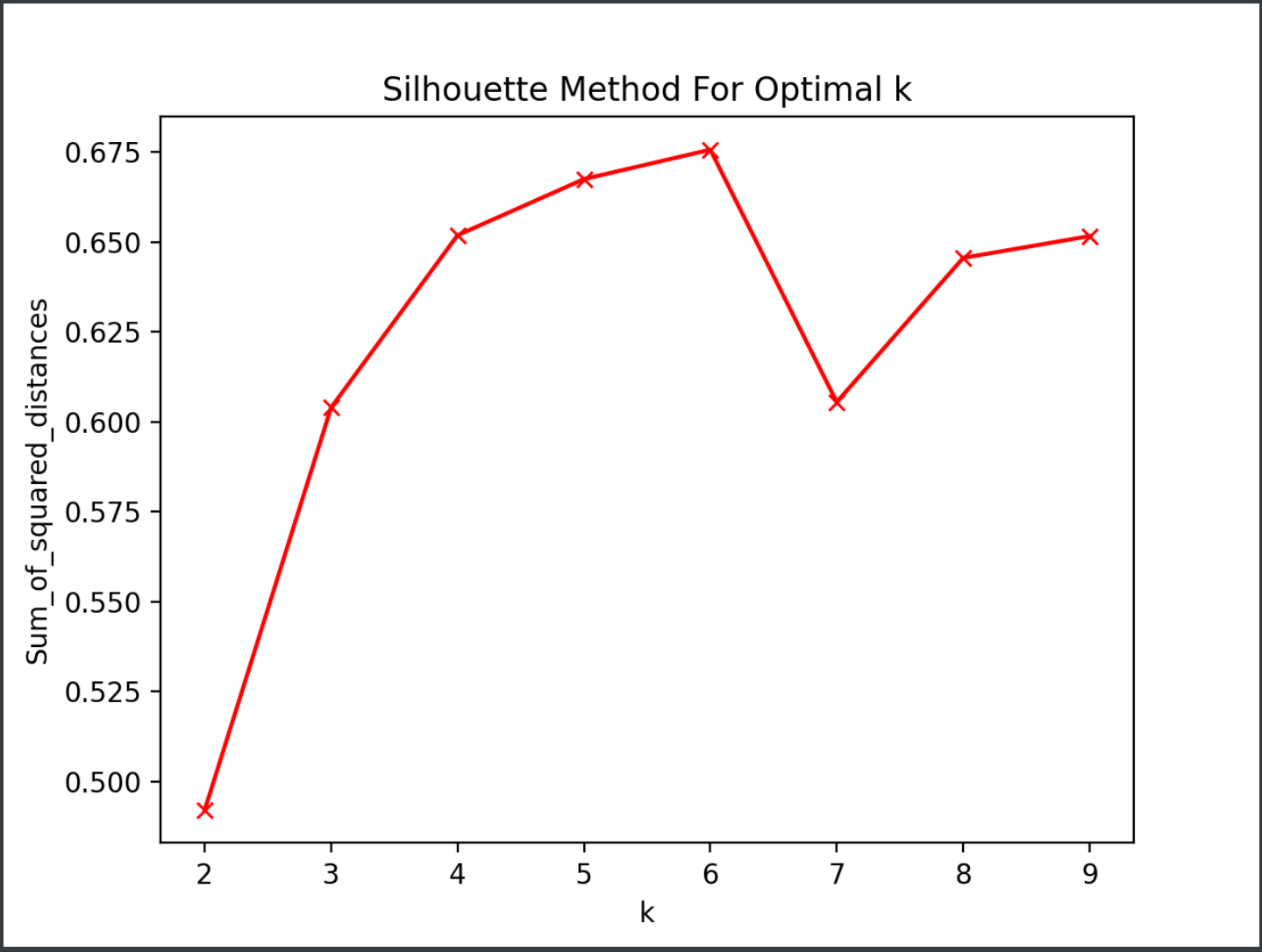
For filter one elbow method:



It is not reaching elbow in 200, so change

Another method:

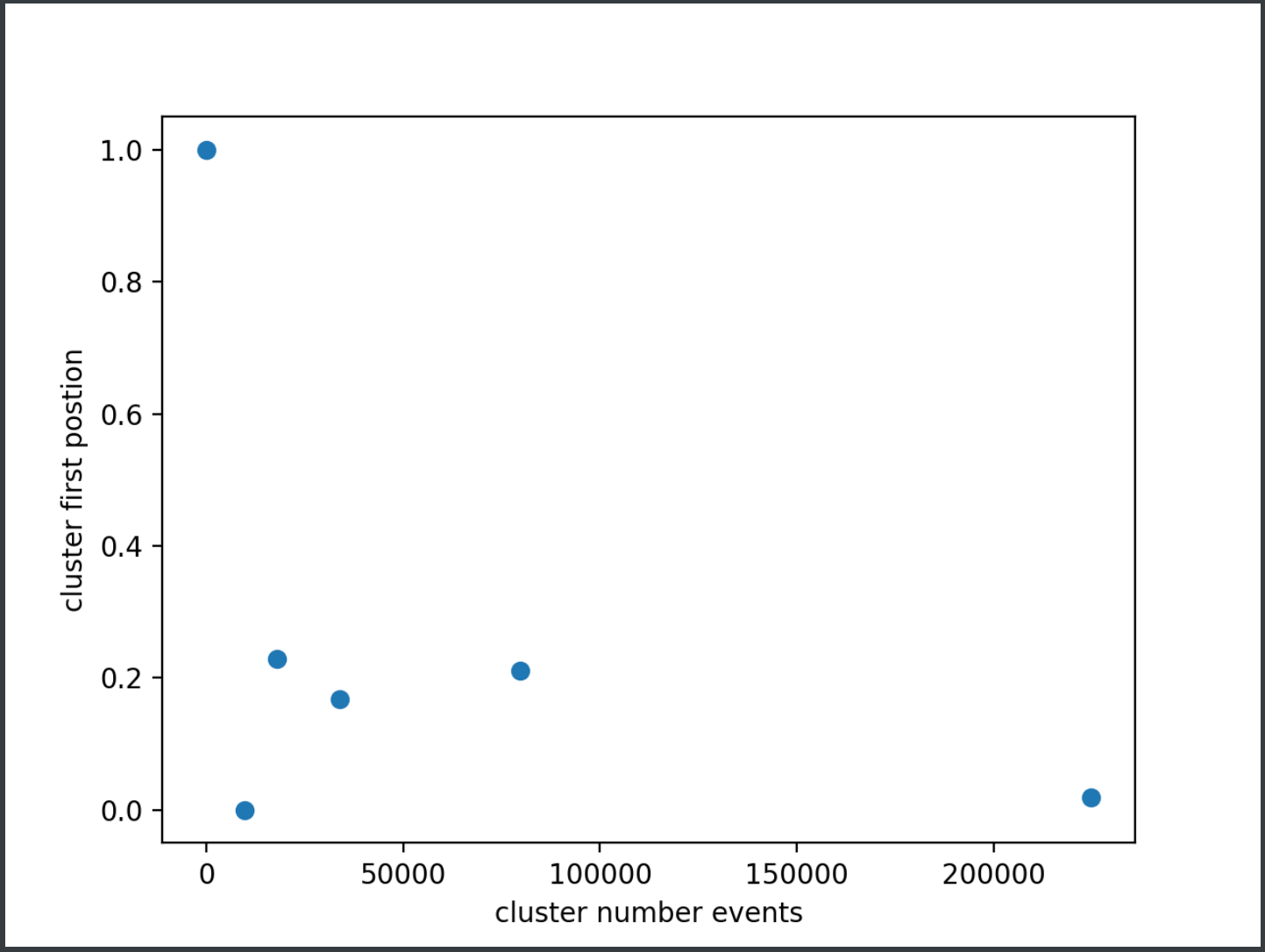
Silhoutte_score:



From the optimal, if use the

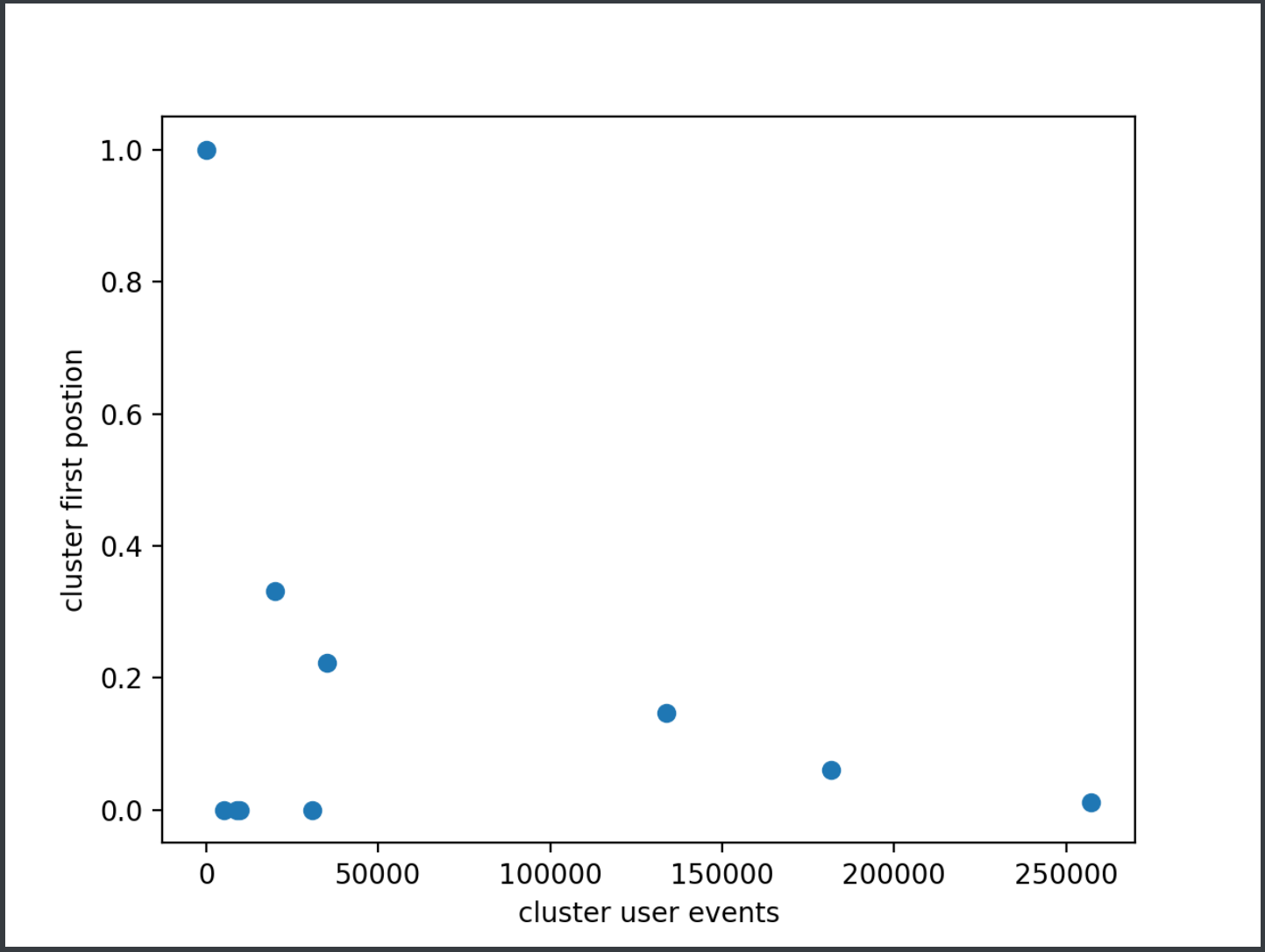
If the optimal number of cluster is 6: the scatter plot is as follows.

Cluster first pos and events



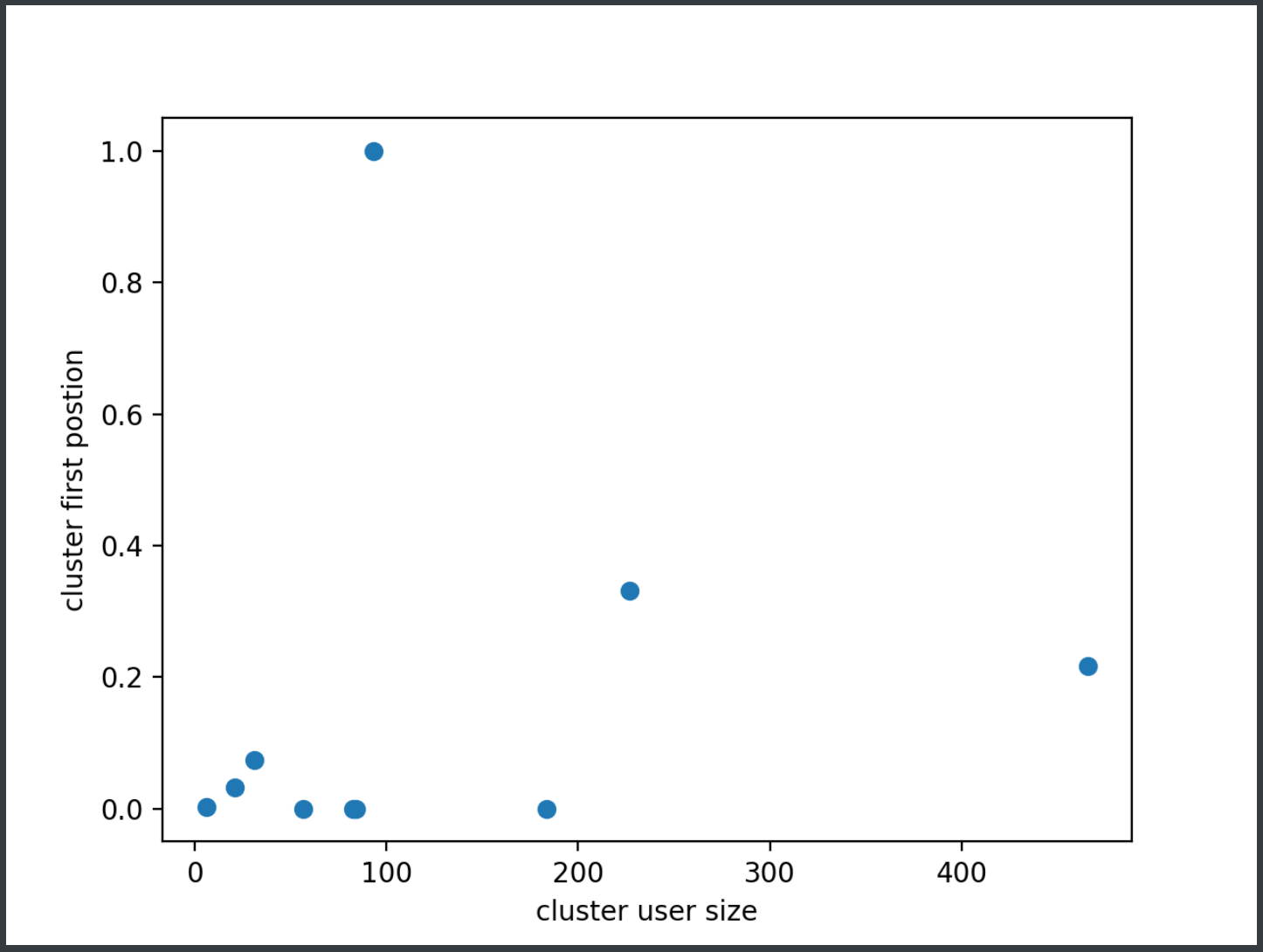
If cluster with ten events:

The result is



Cluster First pos vs users

There are repo with very high value which interrupts...



Then I check the specific work events:

画Heatmap, 取平均，每个vector都画一个heatmap(50xK) Outlier 去掉

Random取5w个repo 带上cluster label 然后再画散点 每组不同颜色