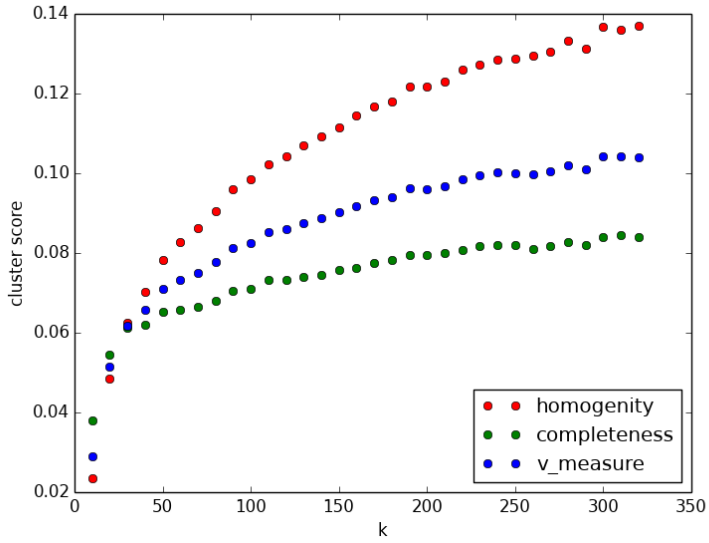
# CS229: Meeting Notes Week 6

*(originally sent via email)*

## K-Means Results

I ran k-means clustering for different k's on the 50 topics LDA data. I measured the homogeneity, completeness and V-measure with the observed category as the cluster assignment. I replaced all the game categories (ARCADE, RACING, etc.) with a single category GAME, so there are totally 25 categories. I have attached a plot of this.




I also printed the frequency of categories in each cluster to see what the clusters look like, I have attached that file here too.

## Manual Annotations

The clusters are very impure with respect to category so I think either categories are a varied mix of application "types", or our topics are not actually application "types". I tried to observe this by manually assigning some label to each topic and then annotating the apps. I have attached the annotated file here.

There are two interesting things about the annotations. The first is that I could not label each topic as an application "type". I had to label it as sort of a "function"; for example, READ_SMS + WRITE_SMS was labelled as SMS. I think our topics are not actually application "types" like social networking or photography, but just compound permissions: permission pairs that co-occur often.

The second thing is that with these annotations, the malware applications exhibit very regular patterns. For example, SMS + LOC (location) + RESTART (to restart applications), maybe because it sends location data via SMS, and if the user tries to stop the application, it restarts itself. I think this is the reason that SVM is able to work fairly well to separate the malware from non-malware in this space.

## Ideas

"Types" are a higher-level representation than the topics we have now, it may be useful to learn this representation from the data. I have two ideas on doing this that we can discuss in our meeting today.

One issue is that I have not evaluated the LDA model itself, or modified the LDA priors. There is a paper [2] by Hannah Wallach about when the priors matter and [3] about evaluation of LDA. The authors have implemented both methods in Mallet [4]. I watched a video by Blei about what the priors represent and I was planning to read these 2 papers too.

## About the cluster measures

Homogeneity is the same as cluster purity, and completeness is higher when a category is present in a single cluster, and lower if a category is present in multiple clusters. V-measure is the harmonic mean of homogeneity and completeness. It is equivalent to NMI (there is a short proof of that in page 165 of [1]).

[1]: http://www.cs.columbia.edu/~hila/hila-thesis-distributed.pdf
[2]: http://people.cs.umass.edu/~wallach/publications/wallach09rethinking.pdf
[3]: https://www.era.lib.ed.ac.uk/bitstream/1842/4587/1/MurrayI_Evaluation%20Methods%20for.pdf
[4]: http://mallet.cs.umass.edu/topics.php