Currently, we are using K means clustering using two dimensions of the Document (i.e. web page)

1. ‘platform’: whether the document was accessed on a desktop, a tablet, or mobile phone
2. ‘location’: in what country the viewer of the page was located

We plan to run the k-means algorithm for progressively increasing values of k (starting k=2), and pick the value of k which leads to the steepest decrease in the WSS (Within Groups Sum of Squared distance).

We plan to explore the use of remote infrastructure (either Cloud, or computing clusters at Gatech) to speed up this task – currently, running a clustering algorithm with two dimensions on the ~25 million records of the file ‘events.csv’ takes >2hr per iteration on a personal computer (12 GB memory).

The following dimensions of the Document (web page) also hold promise for use in clustering. We will trade off the improvement in the prediction of our model with the computing cost of including the following in our clustering process:

* publisher\_id (the publisher or host of the web page – anonymised by Outbrain in the training data set)
* document\_category (the category of the web page, as determined and anonymised by Outbrain in the training data set)

Dimensions of the advertisement that could be used in clustering:

* campaign\_id – every advertisement is part of a unique campaign, and may have content similarities with other advertisements of that campaign
* advertiser\_id – the advertiser may give an indication of the type of the advertisement’s content