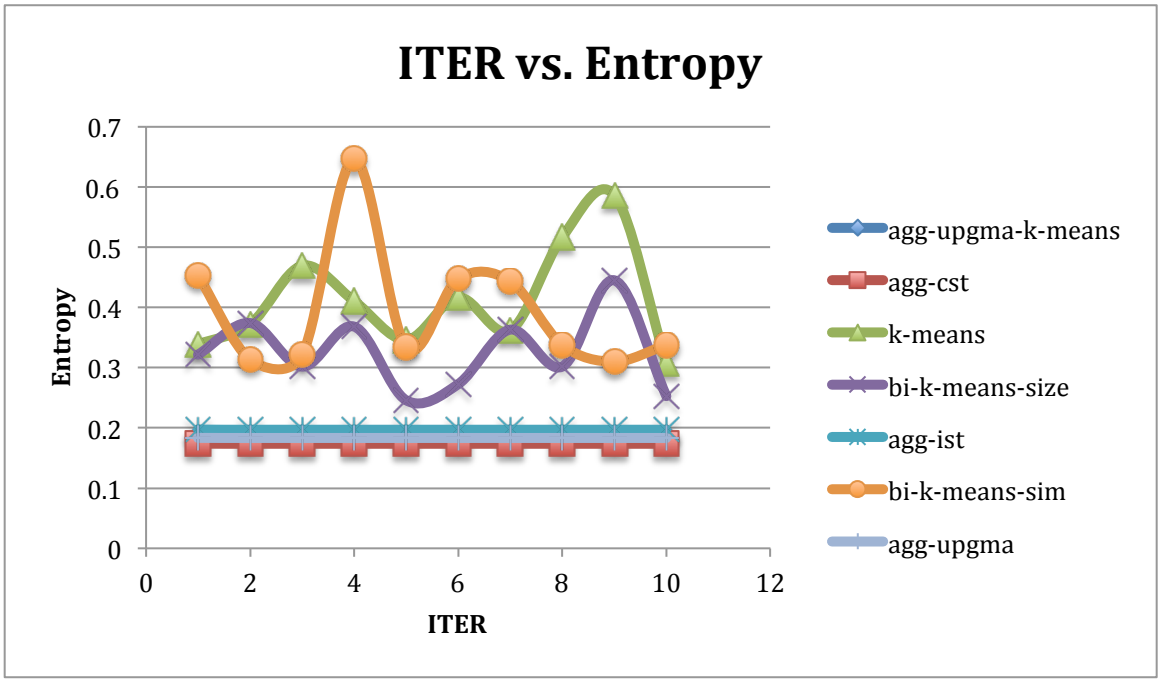
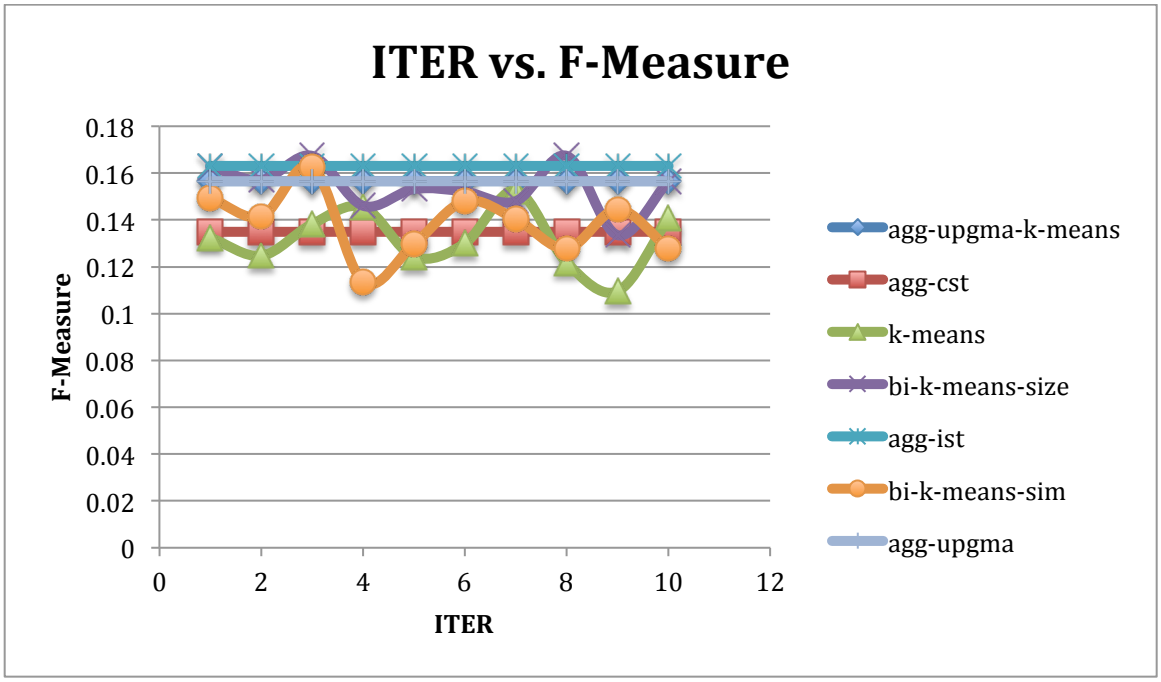
Varying ITER

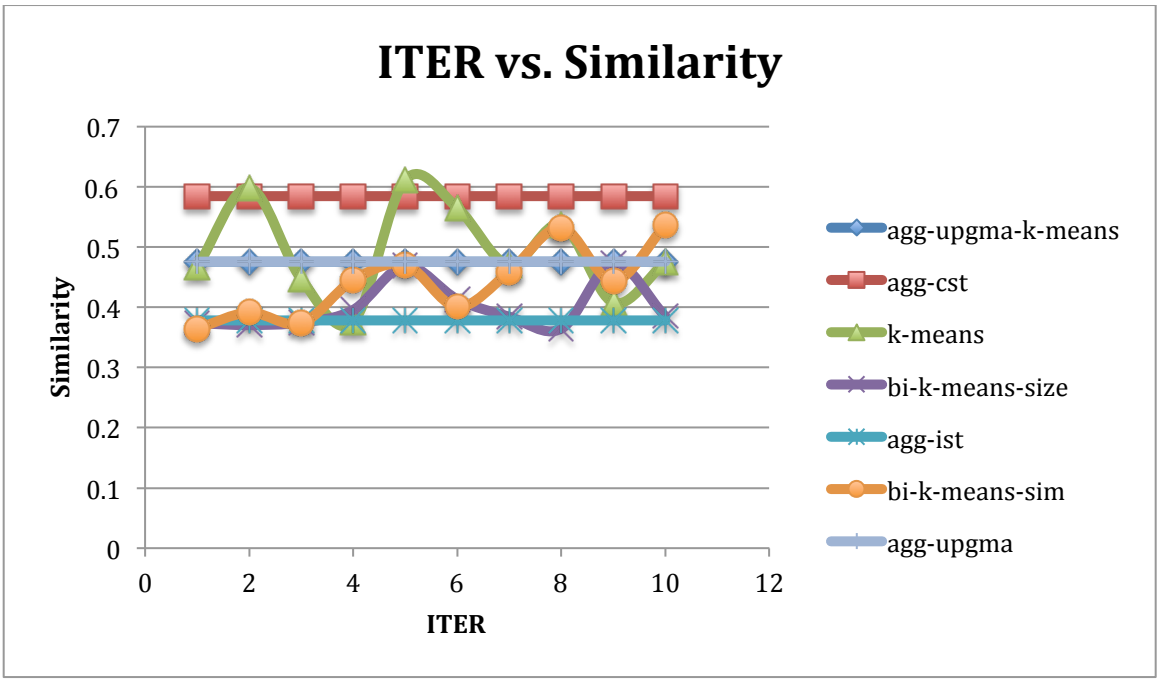
- The three Agglomerative approaches did on the ITER attribute, and therefore did not vary with different values.
- The Agglomerative approaches, in general, will produce higher quality results (Lower entropy, Higher F-Measure, Higher Similarity, Lower Silhouette). They are also computationally more expensive.
- The random hills for the K-Means related approaches are due to the fact that entities are selected at random to being with.
- **All results were based on one trial.**



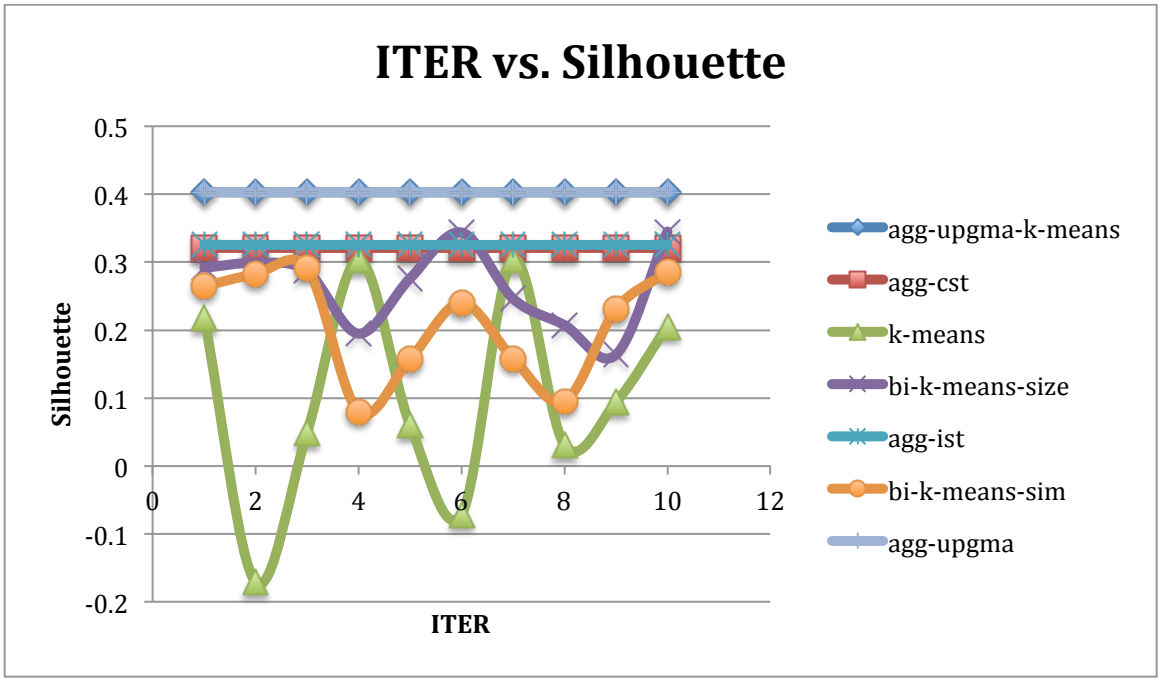
Analyzing only the K-means approaches, both Bi-K-Means approaches returned, on average, better entropy than normal K-means.



The F-Measure for all of the clustering solutions was quite similar. Bi-k-means based on largest cluster scored quite exceptionally when compared to the agglomerative approaches. However, Bi-K-Means based on similarity scored on par with normal K-means, which is undesired. Perhaps this is due to randomness.



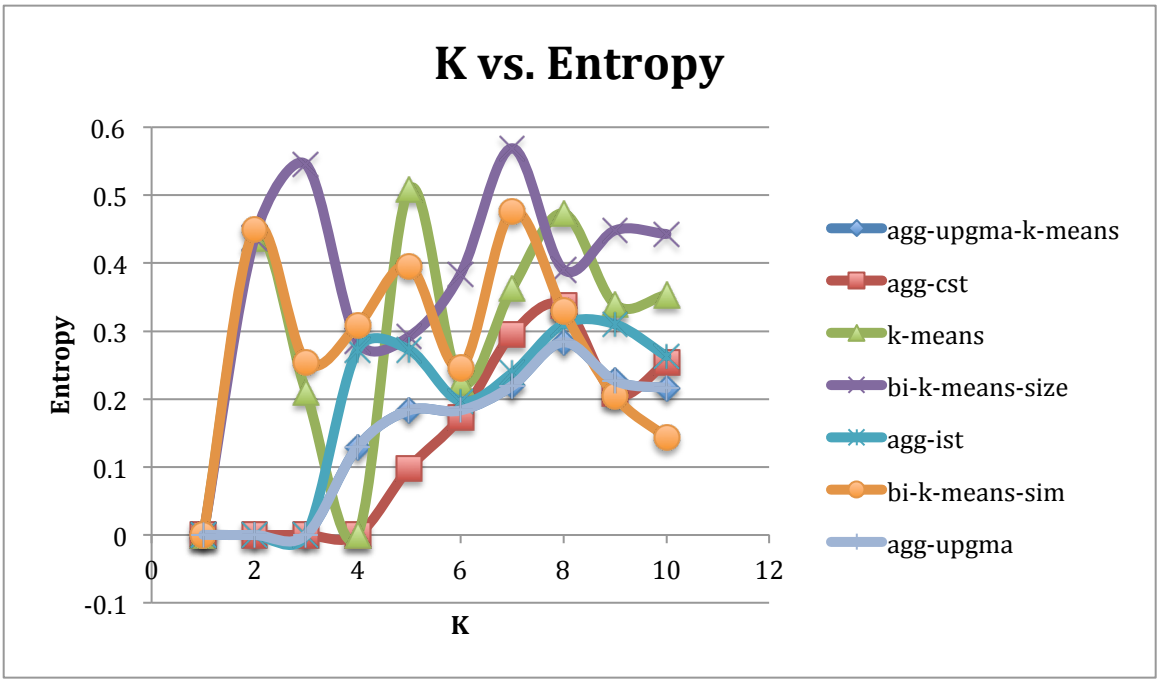
The variation in ITER didn't have a major impact of the similarity measure.



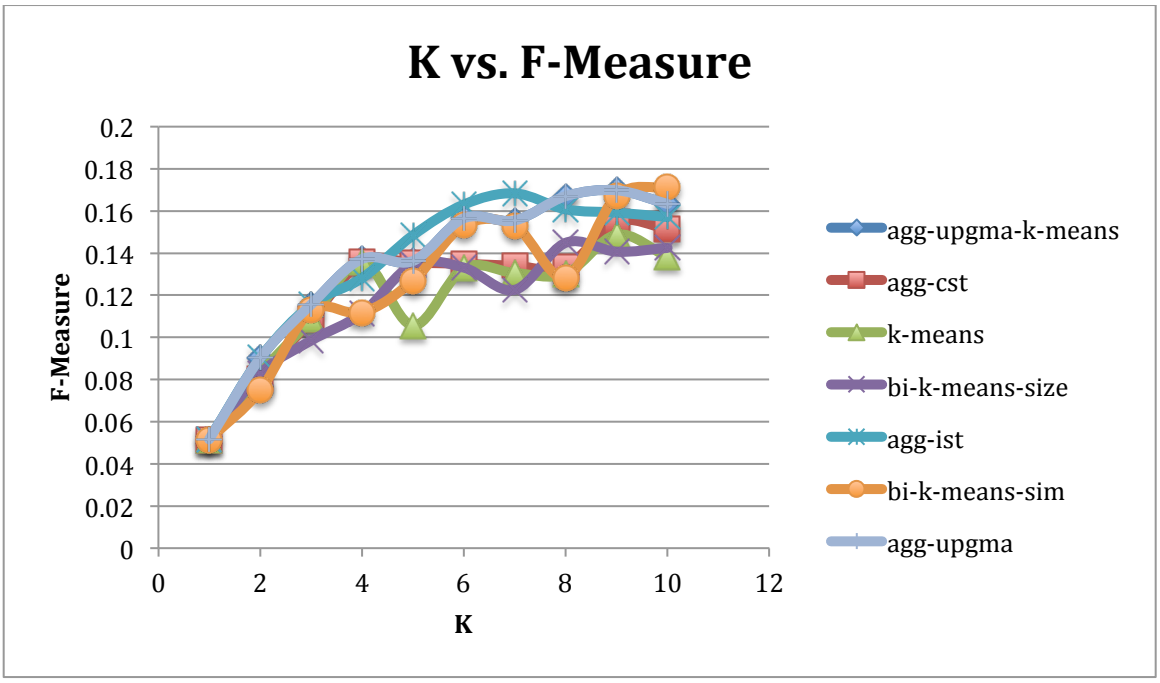
In all cases, the K-Means related algorithm outperformed the agglomerative approaches, with Bi-K-Means performing the best. Normal K-means seemed to be all over the map, even spiking into negative territory (very bad).

Varying Clusters

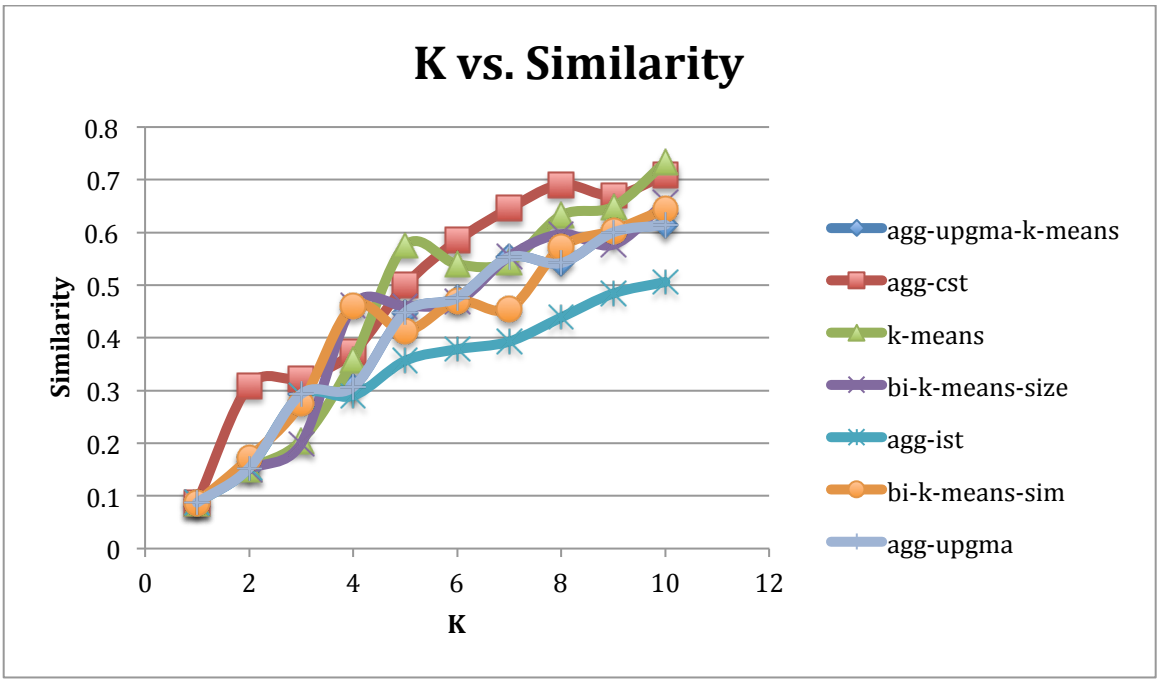
- The Agglomerative approaches, in general, will produce higher quality results (Lower entropy, Higher F-Measure, Higher Similarity, Lower Silhouette). They are also computationally more expensive.
- The random hills for the K-Means related approaches are due to the fact that entities are selected at random to being with.
- **All results were based one trial.**



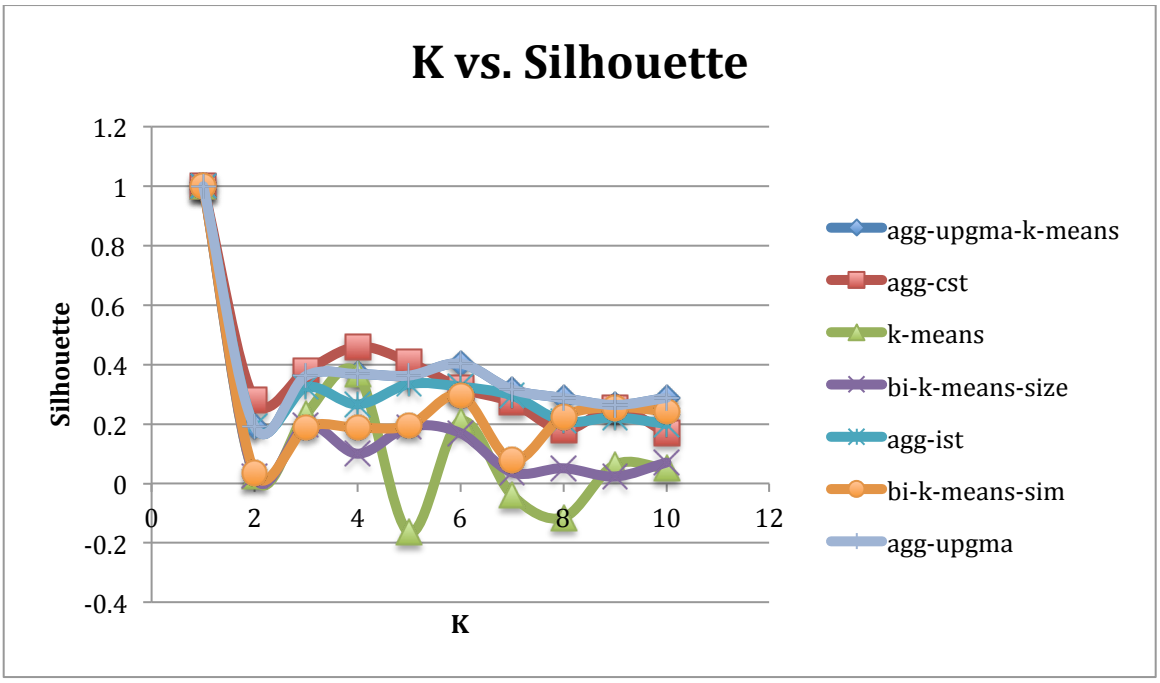
The entropy of all the solutions appeared best around 6, which is good because there were 6 topics to begin with.



The F-Measure steadily increased with the number of clusters. A possible explanation for this is that the size of the clusters reduced, therefore making more "precise" clusters, increasing our F-Measure.



As the number of clusters increased, the overall similarity of the clustering solution increased. This is possible due to the size of the clusters. Like the results from F-Measure, if there are fewer entities in a cluster, they likely have more in common (similar).



With the exception of the sharp drop in the beginning, the size of the cluster didn't have a direct effect on the results for the silhouette coefficient.