

Output Clusters for Each value of K are listed below

For each K, A loop for 300 iterations is run to find the minimum RSS by choosing 300 different random seeds for clustering the documents and the ones with the Least value of RSS are reported below. As K increases, RSS decreases. A graph is plot and attached below for the following values of RSS

k=2

cluster 0

['#tcot.txt', 'facebook.txt', 'Rolling Stone.txt', '#katycats.txt', 'Obama.txt', '#SEO.txt', '#iHeartRadio.txt', 'Russia.txt', 'White House.txt', 'Putin.txt', '#katyperry.txt', '#darkhorse.txt', 'WhatsApp.txt', 'user privacy.txt', '#Instagram.txt', 'foreign policy.txt', 'Ukraine.txt', 'Rand Paul.txt', '#TaylorSwift.txt', 'Jan Koum.txt']

cluster 1

['jeremy lin.txt', 'linsanity.txt', 'James Harden.txt', 'houston nba.txt', 'Zuckerberg.txt', 'toyota center.txt', '#ladygaga.txt', '#rockets.txt', '#socialmedia.txt', 'kevin mchale.txt', '@DwightHoward.txt', '#sxsw.txt']

RSS 15754.3778746

converged in 12 iterations

k=4

cluster 0

['#tcot.txt', 'Obama.txt', 'Russia.txt', 'White House.txt', 'Putin.txt', 'user privacy.txt', 'foreign policy.txt', 'Ukraine.txt', 'Rand Paul.txt', 'Jan Koum.txt']

cluster 1

['Rolling Stone.txt', '#katyperry.txt', '#darkhorse.txt', '#ladygaga.txt', 'WhatsApp.txt', '#TaylorSwift.txt', '#sxsw.txt']

cluster 2

['jeremy lin.txt', 'linsanity.txt', '#katycats.txt', 'James Harden.txt', 'houston nba.txt', 'toyota center.txt', '#rockets.txt', 'kevin mchale.txt', '@DwightHoward.txt']

cluster 3

['facebook.txt', '#SEO.txt', '#iHeartRadio.txt', 'Zuckerberg.txt', '#Instagram.txt', '#socialmedia.txt']

purity 0.84375

RSS 14500.3065141

converged in 12 iterations

k=6

cluster 0

['Rolling Stone.txt', '#katycats.txt', '#iHeartRadio.txt', '#katyperry.txt', '#darkhorse.txt', '#ladygaga.txt', '#TaylorSwift.txt']

cluster 1

['Russia.txt', 'Putin.txt', 'Ukraine.txt']

cluster 2

['jeremy lin.txt', 'linsanity.txt', 'James Harden.txt', 'houston nba.txt', 'toyota center.txt', '#rockets.txt', 'kevin mchale.txt', '@DwightHoward.txt']

cluster 3

['#tcot.txt', 'Obama.txt', 'White House.txt', '#socialmedia.txt']

cluster 4

['#SEO.txt', 'foreign policy.txt', 'Rand Paul.txt']

cluster 5

['facebook.txt', 'Zuckerberg.txt', 'WhatsApp.txt', 'user privacy.txt', '#Instagram.txt', 'Jan Koum.txt', '#sxsw.txt']

RSS 13149.2963311

converged in 12 iterations

k=8

cluster 0

['facebook.txt', 'Rolling Stone.txt', '#katyperry.txt', '#darkhorse.txt', '#ladygaga.txt', 'WhatsApp.txt', '#Instagram.txt', '#TaylorSwift.txt', 'Jan Koum.txt', '#sxsw.txt']

cluster 1

['#katycats.txt', '#SEO.txt', '#iHeartRadio.txt']

cluster 2

['#tcot.txt']

cluster 3

['foreign policy.txt', 'Rand Paul.txt']

cluster 4

['Zuckerberg.txt', 'user privacy.txt', '#socialmedia.txt']

cluster 5

['jeremy lin.txt', 'linsanity.txt', 'James Harden.txt', 'houston nba.txt', 'toyota center.txt', '#rockets.txt', 'kevin mchale.txt', '@DwightHoward.txt']

cluster 6

['Russia.txt', 'Putin.txt', 'Ukraine.txt']

cluster 7

['Obama.txt', 'White House.txt']

RSS 11764.3856755

converged in 12 iterations

Purity for K=4

=> purity 0.84375 I.e 84%

A graph showing the RSS Vs K for $k=0,2,4,6,8$. As k increases, RSS decreases as shown below.

