Results

Ans e)

Clustering for all categories

printing dimensions for tfidf

(18846, 151006)

Performing Analysis for SVD

('For dimension: ', 2)

('homogeneity score -- ', 0.23781762952861971)

('adjusted rand score -- ', 0.07025120372350796)

('For dimension: ', 3)

('homogeneity score -- ', 0.32155064548877371)

('adjusted rand score -- ', 0.13279977500658063)

('For dimension: ', 4)

('homogeneity score -- ', 0.37164809742731553)

('adjusted rand score -- ', 0.17335781645034551)

('For dimension: ', 5)

('homogeneity score -- ', 0.39960579546726593)

('adjusted rand score -- ', 0.19931033480994439)

('For dimension: ', 6)

('homogeneity score -- ', 0.386802109699248)

('adjusted rand score -- ', 0.19147176025155418)

('For dimension: ', 7)

('homogeneity score -- ', 0.38309667459033914)

('adjusted rand score -- ', 0.19262655980606697)

('For dimension: ', 8)

('homogeneity score -- ', 0.40361992825071902)

('adjusted rand score -- ', 0.21994938900975783)

('For dimension: ', 9)

('homogeneity score -- ', 0.42427899754845866)

('adjusted rand score -- ', 0.23144454834270814)

('For dimension: ', 10)

('homogeneity score -- ', 0.43588050472103079)

('adjusted rand score -- ', 0.25704005552884029)

('For dimension: ', 11)

('homogeneity score -- ', 0.42796288102006041)

('adjusted rand score -- ', 0.24119587319349861)

('For dimension: ', 12)

('homogeneity score -- ', 0.42067221256246257)

('adjusted rand score -- ', 0.24703152565003939)

('For dimension: ', 13)

('homogeneity score -- ', 0.43181783574899302)

('adjusted rand score -- ', 0.2502054491289511)

('For dimension: ', 14)

('homogeneity score -- ', 0.43106437595861086)

('adjusted rand score -- ', 0.25874744556528811)

('For dimension: ', 15)

('homogeneity score -- ', 0.43123611933479539)

('adjusted rand score -- ', 0.26281487868404757)

('For dimension: ', 16)

('homogeneity score -- ', 0.43489029595041251)

('adjusted rand score -- ', 0.25813195988121895)

('For dimension: ', 17)

('homogeneity score -- ', 0.43531440588226772)

('adjusted rand score -- ', 0.26716425667635907)

('For dimension: ', 18)

('homogeneity score -- ', 0.42189836055558677)

('adjusted rand score -- ', 0.24311782444644087)

('For dimension: ', 19)

('homogeneity score -- ', 0.42267870730435386)

('adjusted rand score -- ', 0.24725798521949538)

('For dimension: ', 20)

('homogeneity score -- ', 0.42681528009069652)

('adjusted rand score -- ', 0.26011200369426252)

best results at n\_components for SVD = 17

('Confusion Matrix : ,

array([[ 7, 0, 24, 220, 2, 2, 43, 274, 38, 0, 0, 41, 95,

1, 4, 7, 38, 0, 0, 3],

[ 0, 328, 13, 1, 0, 0, 147, 5, 0, 0, 2, 80, 39,

2, 12, 0, 112, 10, 201, 21],

[ 0, 87, 8, 0, 0, 0, 65, 3, 0, 0, 2, 55, 11,

2, 14, 3, 29, 34, 661, 11],

[ 0, 212, 6, 0, 0, 0, 80, 1, 0, 6, 6, 102, 5,

5, 32, 0, 24, 402, 92, 9],

[ 0, 325, 8, 0, 0, 0, 211, 1, 0, 4, 9, 60, 6,

1, 40, 0, 26, 239, 24, 9],

[ 0, 158, 10, 0, 0, 0, 97, 1, 0, 2, 0, 102, 8,

19, 6, 0, 103, 1, 454, 27],

[ 0, 179, 9, 1, 0, 0, 326, 0, 0, 36, 79, 109, 12,

2, 52, 0, 5, 131, 26, 8],

[ 0, 12, 4, 0, 0, 0, 129, 3, 0, 0, 592, 129, 33,

5, 45, 5, 14, 2, 4, 13],

[ 0, 4, 13, 3, 0, 0, 107, 0, 0, 1, 399, 290, 43,

0, 21, 0, 92, 3, 0, 20],

[ 0, 4, 45, 0, 2, 0, 183, 4, 2, 589, 2, 104, 41,

0, 9, 1, 3, 0, 0, 5],

[ 0, 5, 9, 1, 0, 1, 84, 0, 5, 845, 1, 22, 6,

0, 14, 0, 3, 0, 0, 3],

[ 1, 38, 15, 0, 0, 0, 35, 0, 0, 0, 0, 71, 63,

663, 9, 30, 36, 0, 9, 21],

[ 1, 312, 14, 1, 0, 0, 199, 10, 0, 5, 99, 137, 46,

14, 11, 2, 54, 25, 11, 43],

[ 2, 55, 202, 7, 0, 0, 127, 14, 1, 1, 7, 142, 375,

0, 8, 2, 20, 0, 2, 25],

[ 0, 16, 14, 0, 0, 0, 74, 10, 1, 3, 2, 55, 96,

0, 3, 5, 23, 2, 1, 682],

[ 43, 8, 10, 757, 2, 0, 44, 5, 1, 0, 0, 36, 67,

0, 5, 3, 12, 1, 1, 2],

[ 0, 3, 16, 2, 3, 0, 39, 5, 3, 0, 3, 66, 41,

5, 14, 690, 13, 0, 0, 7],

[ 0, 2, 33, 5, 471, 264, 42, 5, 0, 0, 0, 21, 77,

0, 13, 4, 3, 0, 0, 0],

[174, 2, 15, 7, 2, 0, 44, 3, 2, 1, 1, 66, 247,

2, 55, 129, 4, 0, 0, 21],

[ 14, 1, 19, 194, 2, 1, 47, 72, 76, 1, 1, 37, 71,

0, 23, 59, 5, 0, 0, 5]]))

('homogeneity score : ', 0.42616603433003425)

('completeness score : ', 0.45377092616322373)

('adjusted rand score : ', 0.25729621390083574)

('adjusted mutual info score : ', 0.42430845657393135)

Performing Analysis for NMF

('For dimension: ', 2)

('homogeneity score -- ', 0.18974848198890773)

('adjusted rand score -- ', 0.071307735985516763)

('For dimension: ', 3)

('homogeneity score -- ', 0.2631747936188194)

('adjusted rand score -- ', 0.12300178855528618)

('For dimension: ', 4)

('homogeneity score -- ', 0.28402342221186744)

('adjusted rand score -- ', 0.11829754344722444)

('For dimension: ', 5)

('homogeneity score -- ', 0.3394409256134554)

('adjusted rand score -- ', 0.15425694612374616)

('For dimension: ', 6)

('homogeneity score -- ', 0.36680198654206858)

('adjusted rand score -- ', 0.19042266888216089)

('For dimension: ', 7)

('homogeneity score -- ', 0.33857872429576202)

('adjusted rand score -- ', 0.15844153398367938)

('For dimension: ', 8)

('homogeneity score -- ', 0.35686374046463004)

('adjusted rand score -- ', 0.17720410634681477)

('For dimension: ', 9)

('homogeneity score -- ', 0.37197557456559133)

('adjusted rand score -- ', 0.20266265433278219)

('For dimension: ', 10)

('homogeneity score -- ', 0.35692101423925138)

('adjusted rand score -- ', 0.18384692448077364)

('For dimension: ', 11)

('homogeneity score -- ', 0.37997664201254805)

('adjusted rand score -- ', 0.20505829712817586)

('For dimension: ', 12)

('homogeneity score -- ', 0.36309106015312576)

('adjusted rand score -- ', 0.19150998684362294)

('For dimension: ', 13)

('homogeneity score -- ', 0.41496263806276484)

('adjusted rand score -- ', 0.2312298698624764)

('For dimension: ', 14)

('homogeneity score -- ', 0.36809062611799886)

('adjusted rand score -- ', 0.20060679612008772)

('For dimension: ', 15)

('homogeneity score -- ', 0.3703091520508115)

('adjusted rand score -- ', 0.19193693277383758)

('For dimension: ', 16)

('homogeneity score -- ', 0.40214420340151791)

('adjusted rand score -- ', 0.2277486306652762)

('For dimension: ', 17)

('homogeneity score -- ', 0.39520965344945436)

('adjusted rand score -- ', 0.21174689024371177)

('For dimension: ', 18)

('homogeneity score -- ', 0.4157394810789265)

('adjusted rand score -- ', 0.25585512675552946)

('For dimension: ', 19)

('homogeneity score -- ', 0.41087061959160509)

('adjusted rand score -- ', 0.25068791130689783)

('For dimension: ', 20)

('homogeneity score -- ', 0.40810516714869371)

('adjusted rand score -- ', 0.25462757156975196)

best results at n\_components = 18

('Confusion Matrix : \n', array([[ 3, 0, 0, 4, 1, 3, 346, 0, 52, 0, 35, 259, 15,

12, 25, 40, 0, 1, 2, 1],

[ 44, 2, 211, 56, 1, 1, 5, 9, 134, 263, 76, 26, 29,

0, 1, 95, 9, 1, 2, 8],

[ 21, 0, 212, 13, 1, 0, 2, 5, 52, 515, 44, 6, 10,

1, 3, 39, 31, 0, 10, 20],

[ 48, 13, 111, 24, 3, 0, 5, 13, 107, 101, 95, 6, 27,

0, 2, 30, 383, 0, 0, 14],

[113, 9, 94, 28, 0, 0, 4, 21, 270, 45, 63, 13, 21,

0, 0, 31, 242, 0, 0, 9],

[ 3, 2, 185, 52, 14, 0, 4, 5, 83, 438, 88, 6, 13,

0, 0, 92, 1, 0, 0, 2],

[156, 50, 48, 7, 0, 0, 4, 78, 339, 23, 81, 4, 10,

5, 0, 6, 133, 0, 8, 23],

[188, 0, 1, 22, 0, 0, 0, 583, 77, 2, 64, 7, 3,

2, 18, 15, 2, 0, 5, 1],

[175, 2, 0, 26, 0, 0, 8, 452, 62, 0, 146, 6, 9,

3, 16, 84, 3, 0, 1, 3],

[ 16, 657, 2, 9, 0, 4, 2, 3, 194, 0, 54, 15, 28,

0, 0, 2, 0, 0, 5, 3],

[ 3, 865, 0, 4, 0, 0, 3, 1, 84, 0, 17, 4, 15,

0, 0, 2, 0, 1, 0, 0],

[ 15, 1, 12, 24, 490, 0, 1, 3, 72, 11, 95, 4, 21,

5, 101, 54, 1, 0, 26, 55],

[175, 17, 47, 76, 5, 0, 4, 188, 178, 22, 108, 13, 26,

1, 5, 63, 41, 0, 3, 12],

[ 66, 2, 11, 55, 0, 1, 15, 44, 134, 3, 118, 23, 435,

12, 35, 19, 0, 0, 5, 12],

[ 18, 5, 3, 710, 0, 1, 1, 2, 56, 3, 27, 14, 13,

0, 9, 18, 2, 0, 4, 101],

[ 3, 0, 3, 8, 0, 2, 840, 2, 37, 1, 20, 6, 3,

52, 10, 10, 0, 0, 0, 0],

[ 3, 1, 0, 6, 2, 2, 4, 2, 65, 0, 39, 12, 14,

0, 649, 13, 0, 0, 93, 5],

[ 3, 1, 1, 0, 0, 504, 14, 0, 68, 0, 14, 10, 12,

3, 47, 2, 0, 256, 4, 1],

[ 16, 3, 0, 36, 1, 7, 13, 8, 98, 0, 43, 7, 10,

203, 319, 2, 0, 0, 5, 4],

[ 3, 2, 1, 2, 0, 5, 326, 1, 71, 0, 33, 77, 13,

16, 70, 3, 0, 0, 1, 4]]))

('homogeneity score : ', 0.41616000249809543)

('completeness score : ', 0.44077174991104395)

('adjusted rand score : ', 0.25646247548046824)

('adusted mutual info score : ', 0.41427134440003177)

Varying K from 10 to 100

Completeness score in navy

Homogeneity score in orange

As observed, optimum value for both lie between 20 to 30



Varying K from 20 to 30 (Plotting graph between 20 to 30 only)

Optimum k = 22 (Both values are high)

This is the value of k for pure clustering



For best value of K = 22, scores are :

('homogeneity score -- ', 0.43776410507486702)

('completeness score -- ', 0.44639959770517912)

('adjusted rand score -- ', 0.2616845204833142)

('adusted mutual info score -- ', 0.4357516669194223)

Ans f )

Truncating features with SVD with number of components as 17 :

K-Means clustering with number of clusters : 6

Trace : 10233 i.e. 10233 docs correctly clustered

Scores are :

Confusion Matrix -- :

[[1479 18 501 287 2 137]

[ 2 3301 85 1009 13 481]

[ 26 2 2027 361 1 208]

[ 12 417 1007 1714 15 787]

[ 2 33 106 824 1498 1516]

[ 1 216 9 491 44 214]]

('homogeneity score -- ', 0.38174473459741737)

('completeness score -- ', 0.37789707283151758)

('adjusted rand score -- ', 0.2931731305878762)

('adjusted mutual info score -- ', 0.37765619407508566)