Data Mining

Abgabe 5

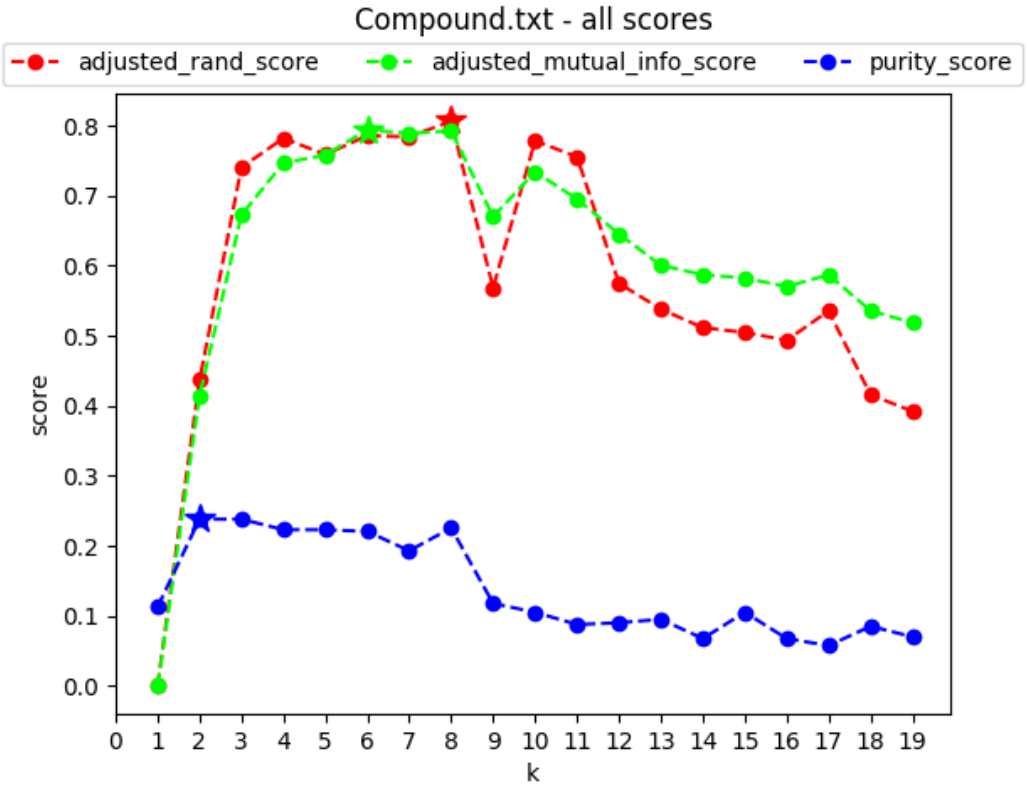
Besten k's

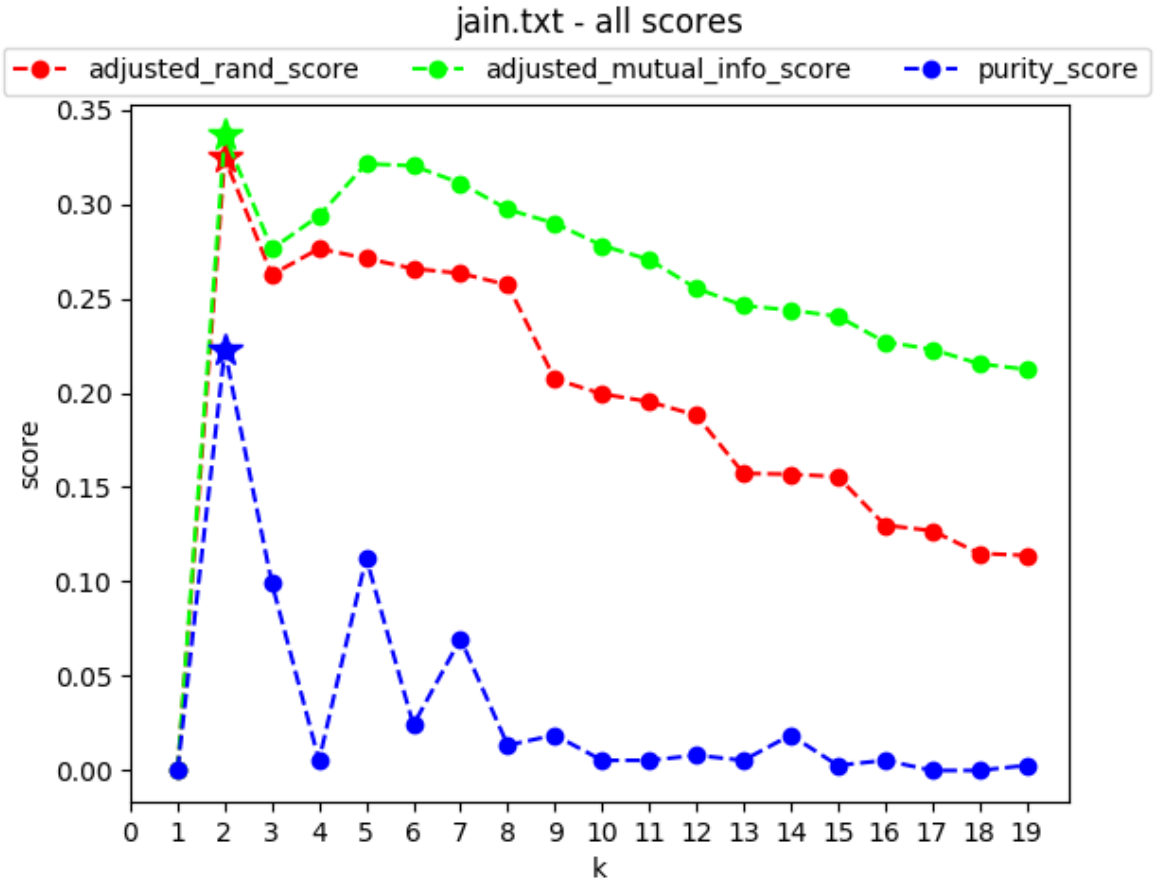
Für das "jain" Datenset ist es einfach, da alle drei Bewertungsmethoden das selbe Ergebnis geliefert haben.
k = 2

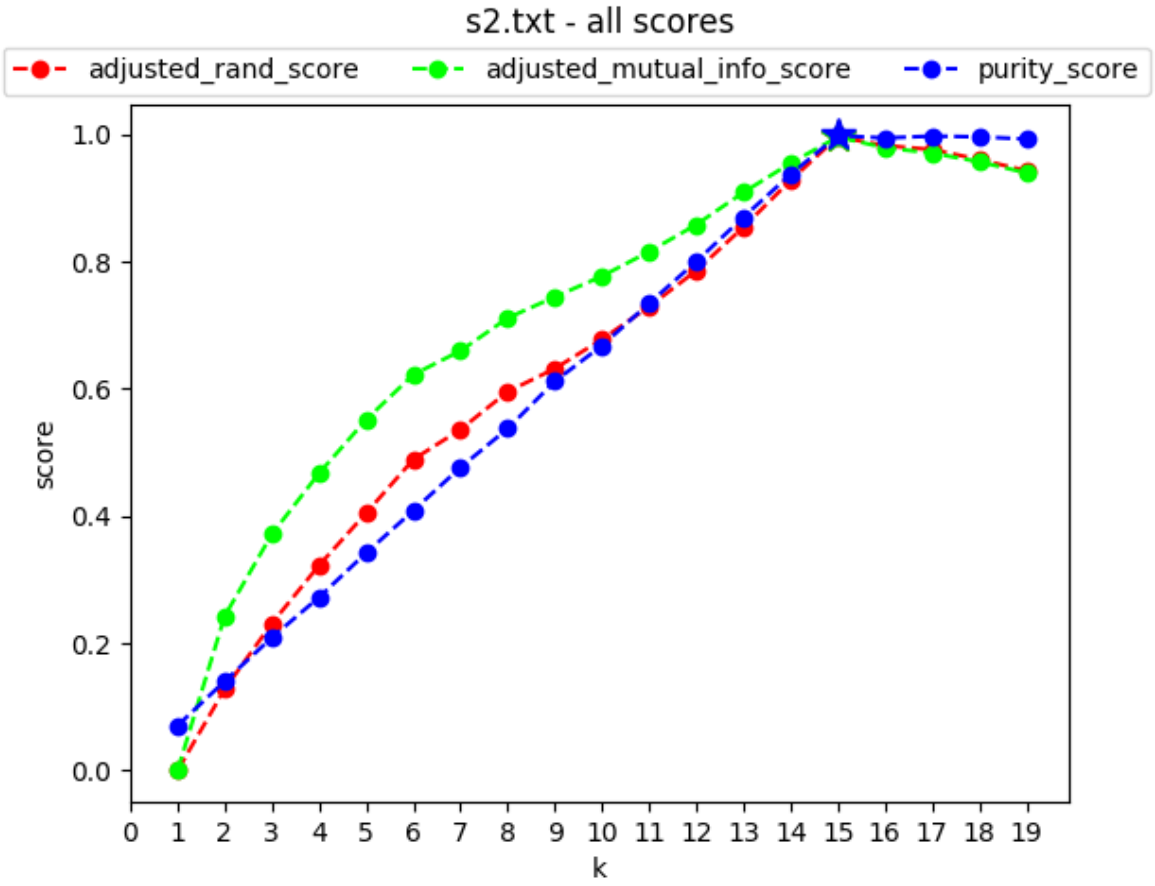
Für das "Compound" Datenset sind sich die Methoden nicht einig.
Rand score sieht k = 8, mutual info score sieht k = 6 und purity score sieht k = 2 als bestes an.
Somit lässt sich es hier nicht exakt sagen, doch nach optischer Bewertung von k = 2, 6, 8 haben wir uns entschieden.
k = 6

Für das "s2" Datenset herrschte Einigkeit, was die bewertung anging, alle zeigten den höchsten score bei k = 15.
Begründet sehe ich das darin, das wir als Ausgangslabels 15 cluster hatten, somit gibt es dort eine fast 100 %ige
Übereinstimmung. Diese Ausgangslabel wurden nach dem optischen Aspekt gewählt.
k = 15

Zufällige Startzentren für dazu, dass es unter umständen ein lokales Maximum gibt, welche jedoch keineswegs das beste clustering darstellt. Somit ist die Bewertung über die Methoden nicht unbedingt endgültig, für ein k sondern es kann mit anderen zufäligen Startwerten ein noch höherer Score erreicht werden.In unserer Analyse haben wir dies, durch mehrmaliges durchlaufen des fürjedes k ausgeschlossen.




jain.txt - all scores
adjusted_rand_score
adjusted_mutual_info_score
purity_score
score
k


s2.txt - all scores
adjusted_rand_score
adjusted_mutual_info_score
purity_score
score
k