

# Klasterovanje ćelija pankreasa embriona miševa

## Seminarski rad iz Istraživanja Podataka 2

Petar Zečević 169/2016

Aleksandra Ružić 47/2016

### Sažetak

Ovaj rad predstavlja istraživanje podataka o ćelijama pankreasa uzetim iz embriona miševa. Podaci su uzeti iz 3 različite studije. Korišćena su dva algoritma klasterovanja, kmeans i spektralno, da bi se utvrdio broj vrsta diferenciranih ćelija u raznim stupnjevima razvoja embriona. Utvrđeno je da se podaci najbolje dele u 2 ili 3 klastera kao i da je kmeans bolji algoritm za ove podatke.

### Sadržaj

Sažetak.....	1
Uvod.....	2
Podaci.....	2
Preprocesiranje.....	3
Metode.....	4
Rezultati.....	4
E12.5.....	5
E13.5.....	7
E14.5.....	8
E15.5.....	11
E17.5.....	12
3140920.....	15
Grupa 2699154, 2699155 i 3140916.....	16
Grupa 2699157, 3140917 i 3140918.....	17
Grupa 3195456 i 3488509.....	20
Grupa 3852752, 3852753, 3852754 i 3852755.....	22
Grupa 3140915, 3140916, 3140917 i 3140918.....	23
Zaključak.....	25
Dodatak - kodovi.....	27

## Uvod

U ovom radu proučavaju se podaci o proteinima generisanim u ćelijama miševa iz različitih tkiva i u različitim stepenima razvoja. Podaci su podeljeni u 16 tabela. Jedna tabela sadrži podatke o ćelijama tkiva pankreasa jednog miša koji je u nekom stadijumu embrionalnog razvoja. Podaci su preprocesirani tako da kolone tabele predstavljaju genetske sekvene kojima se proizvode proteini, a redovi su ćelije čija se proizvodnja proteina koristi, označene jedinstvenim indeksima.

U zavisnosti od stepena razvoja embriona iz kog se ćelije izvlače i vrste ćelije koje se posmatraju, zavisi raspodela sekveni. Informacija o vrsti ćelije nije poznata. Broj vrsta ćelije zavisi od stepena razvoja embriona. Broj vrsta je nepoznat, ali je stepen razvoja za većinu ćelija poznat. Cilj rada je da se na osnovu raspodele sekveni odredi broj vrsta ćelija za dati stepen razvoja.

## Podaci

Podaci su prikupljeni iz 3 različite studije:

1. „Lineage dynamics of pancreatic development at single cell resolution”[1]

Podaci iz ove studije su iz tabela GSM2699154-GSM2699157 i GSM3140915-GSM3140920. Ćelije uzorka su uzete iz pankreasa. Uzorak se uzimao 12 (GSM3140915), 14 (GSM2699154, GSM2699155, GSM3140916) i 17 (GSM2699157, GSM3140917, GSM3140918) dana nakon začeća.

2. „Defining multistep cell fate decision pathways during pancreatic development at single-cell resolution”[2]

Podaci iz ove studije su iz tabela GSM3195456 i GSM3488509. Ćelije uzorka su uzete iz pankreasa. Uzorak se uzimao 14 dana nakon začeća.

3. „Comprehensive single-cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis”[3]

Podaci iz ove studije su iz tabela GSM3852752-GSM3852755. Ćelije uzorka su uzete iz pankreasa. Uzorak se uzimao 12 (GSM3852752), 13 (GSM3852753), 14 (GSM3852754) i 15 (GSM3852755) dana nakon začeća.

Ove studije prate i proučavaju diferencijaciju ćelija pankreasa. Naime, u pankreasu postoji 3 različita tipa ćelija: egzokrine, endokrine i ćelije koje oblažu Virsungov izvodni kanal. Egzokrine ćelije grade tkivo koje luči sokove za varenje, dok endokrine ćelije luče hormone. Endokrine ćelije se dele na pet tipova: alfa, beta, delta, epsilon i PP ćelije. Sve ćelije pankreasa u embriogenezi nastaju od istog tipa ćelija. Taj tip ćelija se formira u 8 danu i raste do 10-11 dana. Između 11 i 13 dana se diferenciraju 3 tipa ćelija. Od jednog tipa nastaju egzokrine, od drugog endokrine i ćelije kanala i od trećeg samo ćelije kanala. Nakon 13 dana, te ćelije se diferenciraju u dalje podtipove.

Prvi zadatak ovog rada je da isprobava klasterovanje ćelija iz jednog fajla tako da klasteri otprilike predstavljaju različite tipove ćelija pankreasa. Dakle, suština ovog zahteva je da ispita da li sličnost po datim atributima implicira da su ćelije istog tipa ili da će se diferencirati u isti tip u kasnijem

stadijumu razvoja.

Drugi zadatak ovog rada je da poređenjem klasterovanja podataka iz različitih studija koji sadrže ćelije iz istog stadijuma razvoja ispita sličnost samih podataka. Naime, fajlovi koji sadrže ćelije uzete iz istog dana nakon začeća bi trebalo da imaju istu raspodelu diferenciranih ćelija.

S obzirom da nije unapred poznato koja ćelija je kog tipa i da autori nemaju adekvatno znanje biologije, rezultate ovog rada bi trebalo predati stručnom licu koje bi moglo zaključiti na osnovu datih podataka odgovore na data pitanja. Ono čime će se rad baviti je samo klasterovanje i proučavanje osobina klastera koje bi moglo da uputi neko stručno lice na dalje zaključke.

## Preprocesiranje

Sirovi podaci su bili u tabelama čiji su indeksi označavali gene, a kolone ćelije. Identifikatori gena su bili šifre koje su počinjale se ENSMUSG. Šifre su bile uparene sa nazivom u tabeli dozvoljenih gena (common\_mouse\_list.csv).

Prvi korak preprocesiranja je bilo izbacivanje gena koji ne postoje u tabeli dozvoljenih gena. Drugi korak je bilo transponovanje tabela tako da ćelije budu redovi, a geni kolone. Imena ćelija su zamenjena kompozicijom broja koji označava fajl i rednog broja reda.

Sledeći korak je bio izbacivanje gena čija je zastupljenost u ćelijama svih tabela manja od zadatog procenata. Taj procenat je po dogovoru 1%. Nakon toga su se izbacivale ćelije čija je proizvodnja proteina manja od zadatih granica. Te granice su po dogovoru bar 500 različitih gena i bar 1000 ukupno ispoljenih gena. Poslednji korak je bilo izbacivanje gena kojise ne koriste u svim tabelama. Umesto šifri, za identifikatore gena su postavljeni imena gena iz tabele dozvoljenih gena.

Fajl	Broj redova	Nula kol.	Ne-nula kol.
<b>2699154</b>	3495	48	1
<b>2699155</b>	4306	54	0
<b>2699156</b>	4410	53	0
<b>2699157</b>	2224	89	0
<b>3140915</b>	12682	8	11
<b>3140916</b>	9009	12	7
<b>3140917</b>	5571	11	1
<b>3140918</b>	5643	7	1
<b>3140919</b>	7888	9	38
<b>3140920</b>	7897	9	38
<b>3195456</b>	3427	19	30
<b>3488509</b>	6877	15	14
<b>3852752</b>	11182	4	45
<b>3852753</b>	5166	27	47
<b>3852754</b>	10187	21	47
<b>3852755</b>	11431	27	42

Preprocesirane tabele imaju 12208 kolona. Broj redova, nula kolona i kolona bez ijedne nule za svaku tabelu je dat iznad. Treba primetiti da tabele 3140919 i 3140920 imaju isti broj specifičnih kolona. Utvrđeno je da se ceo fajl 3140919 se sadrži u fajlu 3140920. Ovo je greška u podacima. Analiziraće se samo fajl 3140920.

## Metode

Korišćeni su kmeans i spektralno klasterovanje iz scikit-learn biblioteke u jeziku Python. Broj klastera za koji su se puštali algoritmi su 2-7. Zahtevani broj klastera se zaključio na osnovu značenja podataka tj. broja mogućih diferenciranih tipova ćelija u dotoj fazi embriogeneze. Klasterovane su sve tabele pojedinačno, spojene tabele koje sadrže podatke iz istog dana i spojene tabele koje sadrže podatke iz različitih dana. Pri klasterovanju pojedinačnih tabela, izbacivale su se nula kolone te tabele. Grupe tabele koje su spajane su: 2699154, 2699155 i 3140916 (14. dan), 2699157, 3140917, 3140918 (17. dan), 3195456 i 3488509 (14. dan), 3140915, 3140916, 3140917 i 3140918 (12., 14. i 17. dan), 3852752, 3852753, 3852754 i 3852755 (12., 13., 14. i 15. dan).

Ocene klasterovanja koje su korišćene su: silhouette, davies-bouldin i calinski-harabasz score. Za vizuelizaciju je korišćeno TSNE klasterovanje nad koordinatama dobijenim PCA redukcijom koje obuhvataju bar 95% varijanse.

## Rezultati

S obzirom da su podaci dati u obliku retkih matrica, najočiglednije odlike klastera na koje se možemo fokusirati su skroz nula kolone i kolone bez ijedne nule. Naime, ako postoji atribut koji među ćelijama u jednom klasteru nikad ne uzima vrednost nule ili uvek uzima vrednost nule, može se reći da je taj atribut jedna od bitnih odlika klastera.

Porediće se fajlovi koji predstavljaju ćelije uzete u istom danu nakon začeća. Uporediće se veličine klastera, broj zajedničkih nula kolona i broj zajedničkih ne-nula kolona. Treba uzeti u obzir da se algoritmi nisu upotrebili nad istim skupom kolona za sve fajlove jer su pre algoritma izbačene nula kolone za fajl nad kojim se upotrebljava algoritam klasterovanja.

Treba napomenuti da će se posmatrati promena pripadnosti ćelija klasterima pri povećanju broja klastera i pri tome će se često pominjati da se neki klaster formiran u slučaju sa manjim brojem klastera razdvojio u slučaju sa većim brojem klastera. To i slični izrazi se ne upotrebljavaju doslovno, već samo u smislu opažanja promene pripadnosti grupe ćelija.

## E12.5

Prvo će se posmatrati tabele koje sadrže podatke prikupljene 12. dana nakon začeća. To su fajlovi 3140915 i 3852752. Te dve tabele imaju redom 8 i 4 nula i 11 i 45 ne-nula kolona, a imaju zajedničke 3 nula kolone (M58665#En1, M08631#Tgfb1i1, M48528#Nkx1-2) i 9 ne-nula kolona (M12848#Rps5, M74129#Rpl13a, M50708#Ftl1, M20460#Rps27a, M49517#Rps23, M93674#Rpl41, M45128#Rpl18a, M30744#Rps3, M37563#Rps16).

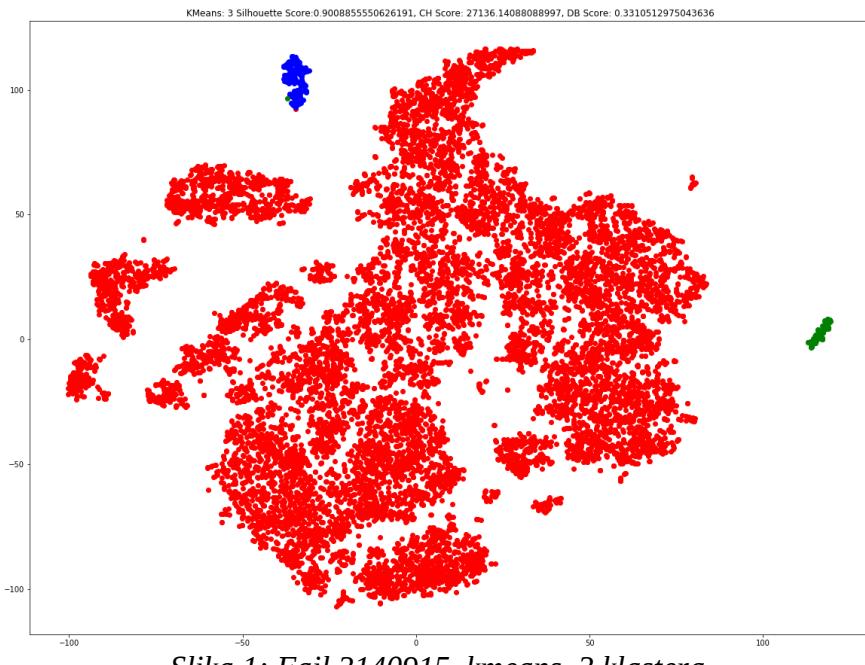
### Kmeans:

U slučaju 2 klastera, nema očiglednih sličnosti između fajlova. Odnosi veličina klastera su potpuno drugačiji (veličine klastera u fajlu 3140915 su 12502 i 180, a u fajlu 3852752 su 4702 i 6480), brojevi nula i ne-nula kolona su daleko od sličnih. Ono što je možda zanimljivo u fajlu 3852752 je da veći klaster ima samo ne-nula kolona koje su ne-nula u celom fajlu i da klaster u fajlu 3852752 veličine 4702 ima sve ne-nula kolone koje ima i klaster iz fajla 3140915 veličine 12502.

U slučaju sa 3 klastera u oba fajla treći klaster se formira od čelija koje se izdvajaju iz dva klastera koja postoje u slučaju sa 2 klastera. Zanimljivo je primetiti da treći klaster koji sadrži 90 čelija u fajlu 3140915 nastaje tako što se 89 čelija izdvoji iz većeg klastera (12502 čelije) i tačno 1 čelija iz manjeg klastera (180 čelija). Izdvajanjem samo te jedne čelije, broj ne-nula kolona se povećava za jedan, što znači da je jedino ta čelija imala nulu u tom atributu- M37742#Eef1a1. Izdvajanjem 89 čelija iz većeg klastera, broj ne-nula kolona se sa 15 povećava na 45.

U slučaju sa 4 klastera, u oba fajla, četvrti klaster nastaje podelom jednog od klastera iz prethodnog slučaja. U fajlu 3140915 klaster veličine 12413 se deli na klastera veličina 10487 i 1926, a u fajlu 3852752 klaster veličine 4817 se deli na klastera veličina 4815 i 2. To implicira da preseci specifičnih kolona ostaju vrlo slični, eventualno se proširuju (jer broj specifičnih kolona izbacivanjem čelija iz klastera može samo da se poveća). U fajlu 3140915 ova podela klastera, uzrokuje značajan pad u ocenama (senka je za 3 klastera bila 0.9, a za 4 je 0.53). Klasteri iz fajla 3852752 veličina 4815 i 2 sadrže redom 32 i 44 ne-nula kolona i po 3 nula kolone celog fajla 3852752 (klaster od kog su nastali je imao 32 zajedničke ne-nula kolone). Klasteri iz fajla 3140915 veličina 10487 i 1926 sadrže redom 9 i 11 ne-nula kolona i 5 i 8 nula kolona celog fajla 3140915 (klaster od kog su nastali je imao 9 zajedničkih ne-nula kolona i 5 zajedničkih nula kolona).

Među fajlovima nema sličnosti ni po ocenama klasterovanja. Prvi fajl ima izuzetno dobre ocene za 2 i 3 klastera (senka za 2 klastera je 0.92, a za 3 je 0.9), dok su mu ostale ocene dosta lošije (već za 4 klastera je senka 0.53, dok je za 5 klastera 0.37). Ocene drugog fajla se ne razlikuju međusobno toliko. Takođe, drugi fajl nema nijednu toliko dobru ocenu. Najbolje ocene ima za 2 klastera (senka je 0.38).

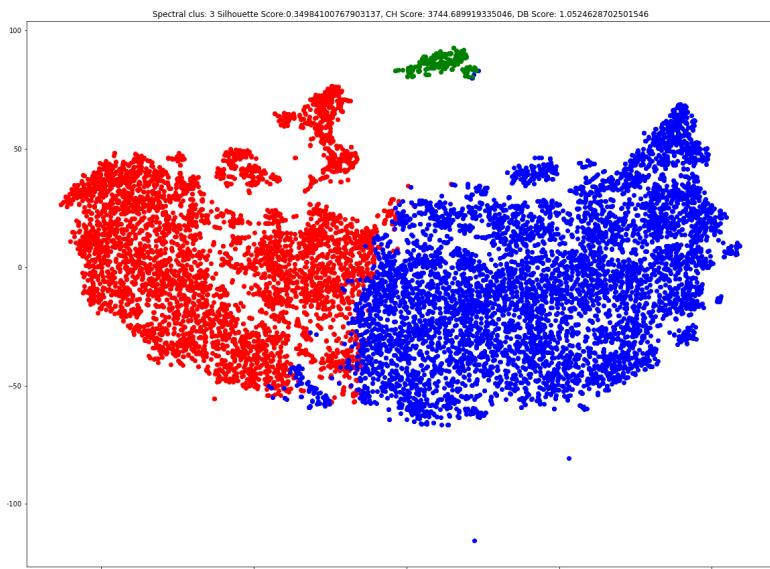


*Slika 1: Fajl 3140915, kmeans, 3 klastera*

## **Spektralno klasterovanje:**

Fajlovi se dosta razlikuju u odnosima veličina klastera. U fajlu 3140915, za 2 klastera, formiraju se klasteri veličina 12500 i 182 sa 0 i 339 nula kolona i 15 i 28 ne-nula kolona. U fajlu 3852752, formiraju se klasteri veličina 6327 i 4855 sa 13 i 7 nula kolona i 153 i 45 ne-nula kolona. U slučaju 3 klastera, u fajlu 3140915 dva klastera nastaju od ćelija koje se nalaze u većem klasteru u slučaju 2 klastera. U fajlu 3852752, u slučaju 3 klastera, 20 ćelija manjeg klastera prelazi u veći klaster iz slučaja sa 2 klastera, a po 107 i 108 ćelija iz oba zajedno formiraju novi klaster. U fajlu 3140915, formiraju se klasteri veličine 12410, 182 i 90 sa 0, 339 i 575 nula kolona i 47, 28 i 33 ne-nula kolone. U fajlu 3852752, formiraju se klasteri veličine 4728, 6239 i 215 sa 7, 15 i 597 nula kolona i 45, 171 i 116 ne-nula kolona.

Ocene za spektralno klasterovanje su približno iste kao i ocene za kmeans za oba fajla i sve brojeve klastera, sem u fajlu 3140915 za 4 klastera gde su ocene za spektralno drastično lošije (za kmeans je senka 0.53, dok je u za spektralno 0.29). Najbolje ocene u fajlu 3140915 se ostvaruju za 2 klastera (senka je 0.92), dok se u fajlu 3852752 najbolje ocene ostvaruju za 3 klastera (senka je 0.35).



Slika 2: Fajl 3852752, spektralno, 3 klastera

U fajlu 3140915 klasteri koji se formiraju za broj klastera do 4, su gotovo isti kao i klasteri koje formira kmeans. Za 4 klastera i više, ta sličnost nestaje. Za 4 klastera, kmeans formira klastera veličine 10487, 1926, 179 i 90, dok spektralno klasterovanje formira klastera veličine 11726, 684, 182 i 90. Poslednja dva klastera u oba algoritma su gotovo ista. Treba primetiti da se u poslednjim klasterima nalaze istih 89 celija, a da ta jedna celija pravi veliku razliku u broju nula kolona (klaster formiran kmeans-om ima 611 nula kolona, a klaster formiran spektralnim klasterovanjem ima 567 nula kolona). Celija koja se nalazi u klasteru koji je formirao kmeans je 3140915\_5629, a celija koja se nalazi u klasteru koje je formiralo spektralno klasterovanje je 3140915\_10886.

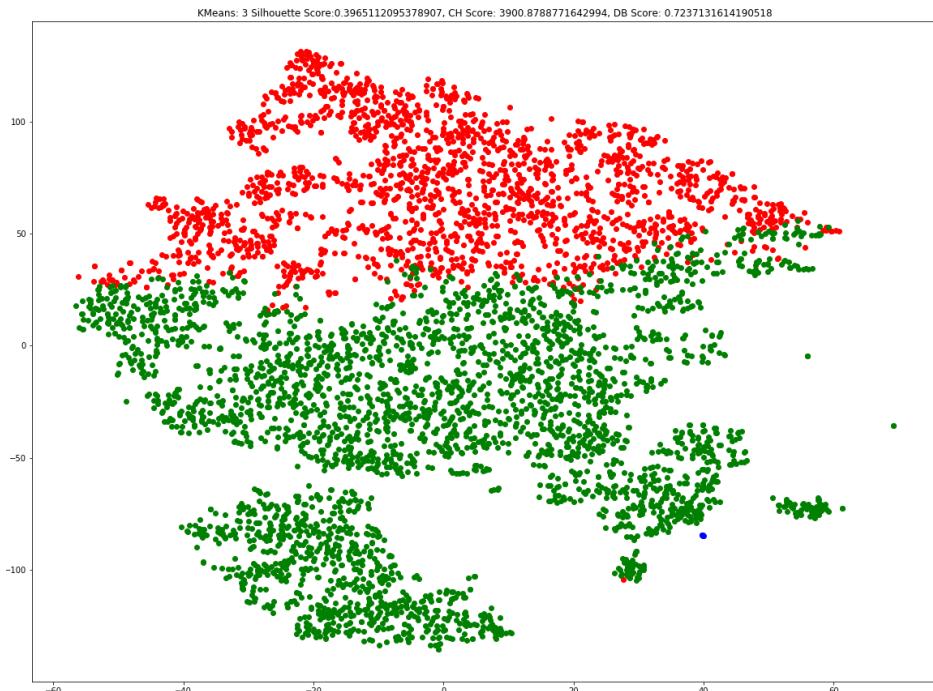
## E13.5

Celije iz ove grupe su uzete 13. dana nakon začeća. Ovoj grupi pripada samo fajl GSM3852753. On ima 5166 redova i 27 nula i 47 ne-nula kolona.

### Kmeans:

Za dva klastera, formiraju se klasteri veličine 3274 i 1892. Njihove nula i ne-nula kolone su 8 i 47, 64 i 317. Za tri klastera, formiraju se klasteri veličine 1891, 5 i 3270. Njihove nula i ne-nula kolone su 64 i 318, 9677 i 159, 9 i 64. Drugi klaster je nastao od 5 celija koje pripadaju većem klasteru iz prethodnog slučaja, dok su ostale celije tog većeg klastera završile u trećem klasteru zajedno sa jednom celijom iz manjeg klastera. Treba primetiti da treći klaster ima više specifičnih kolona od prvog klastera u prethodnom slučaju, iako nije nastao samo njegovim smanjivanjem.

Brojevi specifičnih kolona po klasteru se povećavaju sa brojem klastera, što znači da su klasteri u neku ruku više razdvojeni. Treba uzeti u obzir da je najbolja ocena za 3 klastera (senka je 0.4).



Slika 3: Fajl 3852753, kmeans, 3 klastera

### Spektralno klasterovanje:

U slučaju sa 2 klastera, veličine klastera su 2574 i 2592 sa 12 i 52 nula kolone i 47 i 262 ne-nula kolone. U slučaju 3 klastera, veličine klastera su 1764, 2607 i 795 sa 71, 22 i 22 nula kolone i 326, 108 i 50 ne-nula kolona. Srednji klaster sadrži čelije iz oba klastera iz prethodnog slučaja, prvi sadrži samo čelije iz drugog, a drugi samo čelije iz prvog klastera.

Ocene ovog klasterovanja su dosta slične ocenama kmeans-a. Značajnije se razlikuju za 3 i 4 klastera (senka kmeans-a za 3 i 4 klastera su 0.4 i 0.28 dok su za spektralno 0.23 i 0.19). Najbolje ocene spektralnog klasterovanja su za 2 klastera (senka je 0.34)

## E14.5

Ovoj grupi pripadaju fajlovi 2699154, 2699155, 3140916, 3488509, 3852754 i 3195456. Imaju redom 3495, 4306, 9009, 6877, 10187 i 3427 redova, 48, 54, 12, 15, 21 i 19 nula kolona i 1, 0, 7, 14, 47 i 30 ne-nula kolona. Imaju zajedničkih 6 ne-nula kolona.

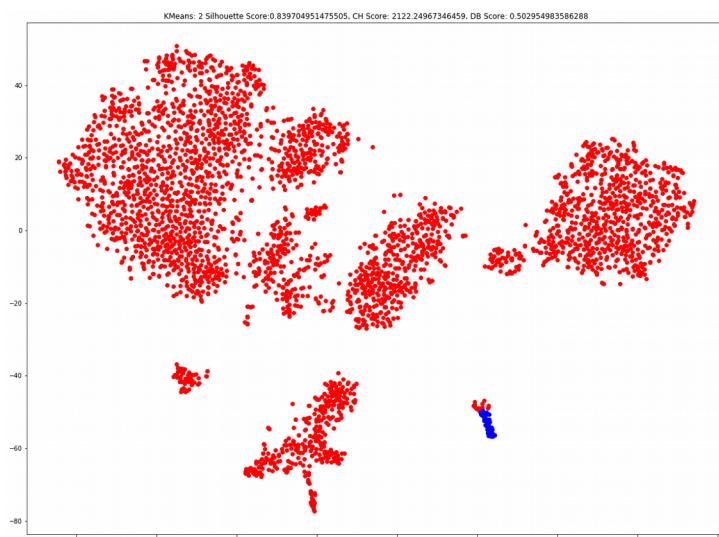
### Kmeans:

U slučaju sa dva klastera može se primetiti da fajlovi 2699154, 2699155, 3140916 i 3195456 imaju slične veličine klastera u odnosu na veličine fajla (za svaki ovaj fajl važida je preko 96% čelija u jednom klasteru) i donekle slične brojeve specifičnih kolona. Takođe, imaju jako dobre ocene za ovo klasterovanje (mada mora se primetiti da 2699155 ima izuzetno dobre ocene- senke je 0.95, dok je za 2699154, 3140916 i 3195456 senka 0.84, 0.79 i 0.89). U sva četiri fajla, veći klaster nema nula kolone.

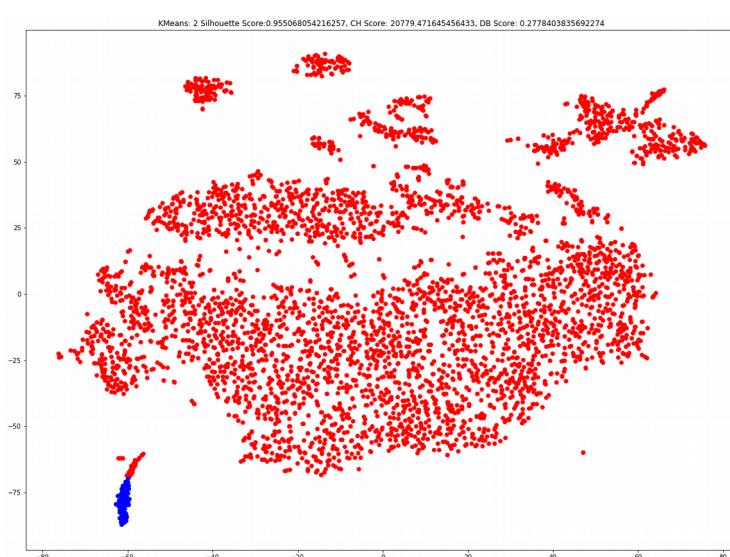
Sa povećanjem broja klastera na 3, ocene za fajl 2699154 naglo opadaju. To se ne dešava za ostale fajlove. Ocena za 2699155 ostaje približno ista (senka je 0.95), dok za 3140916 i 3488509 ocene dostižu maksimum (senke su 0.8 i 0.65).

Sa povećanjem broja klastera na 4, ocene za fajl 3140916 naglo opadaju (za 4 klastera senka je 0.39). Dva od ta četiri klastera su isti kao u slučaju sa tri klastera (to su klasteri veličine 277 i 81). Treći i četvrti klaster sadrže čelije iz većeg klastera u slučaju tri klastera. Manji od ta dva, ima dosta više specifičnih kolona od „roditeljskog“ klastera, dok veći ima skoro isto. Ovoliki pad ocena sa povećanjem broja klastera se ne dešava za ostale fajlove. Ocene za 2699155, 3488509 i 3852754 se blago smanjuju, dok za se 3195456 ocene blago povećavaju.

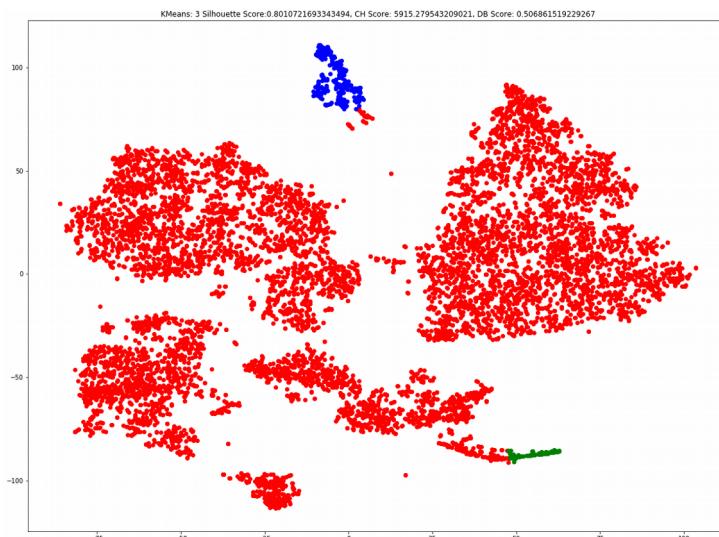
Fajl 2699155 ima izuzetno dobre ocene za broj klastera do 5. Za 6 klastera, ocene naglo opadaju. Za 5 klastera, većinu čelija sadrži jedan veliki klaster (4224), dok su ostala četiri jako mali (48, 17, 14 i 3). Za 6 klastera, izdvaja se i jedan klaster srednje veličine (838).



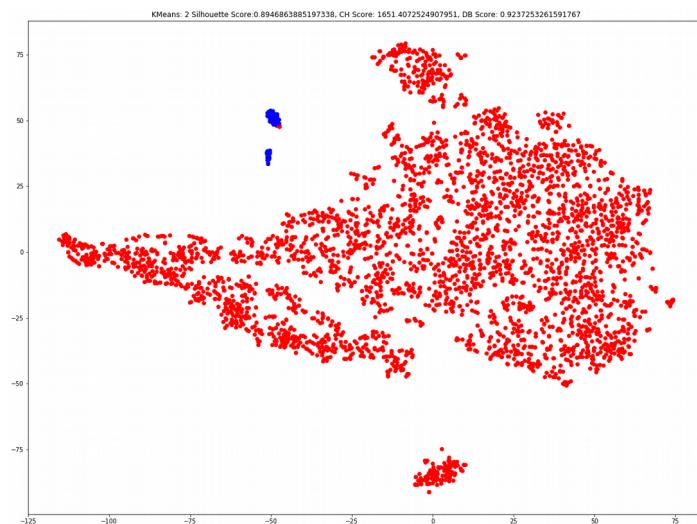
Slika 4: Fajl 2699154, kmeans, 2 klastera



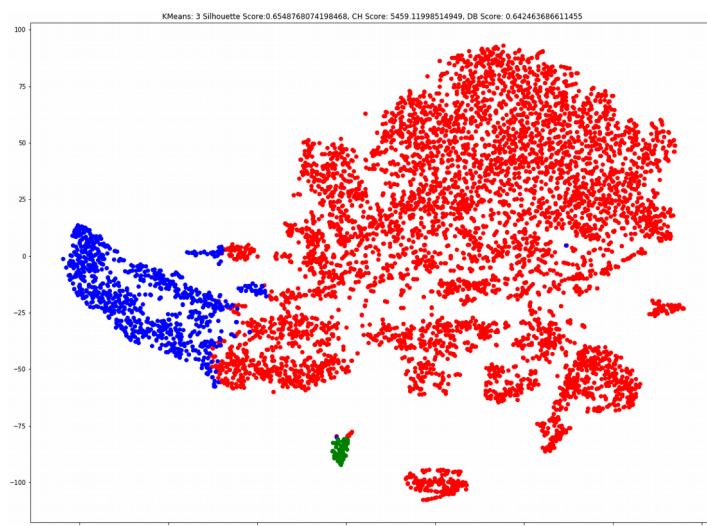
Slika 5: Fajl 2699155, kmeans, 2 klastera



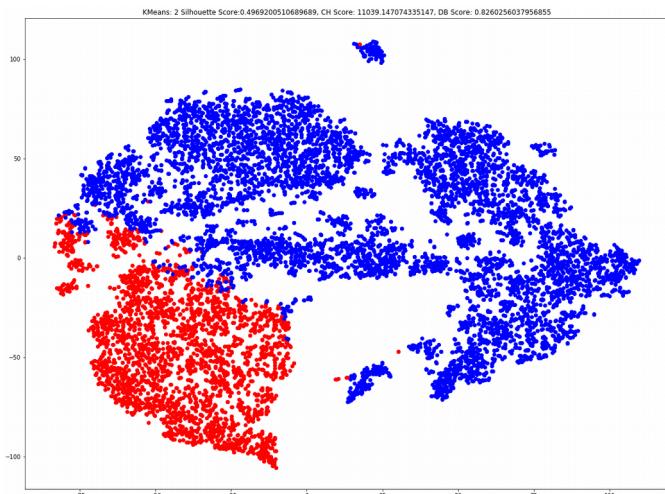
Slika 6: Fajl 3140916, kmeans, 3 klastera



Slika 7: Fajl 3195456, kmeans, 2 klastera



Slika 8: Fajl 3488509, kmeans, 3 klastera



Slika 9: Fajl 3852754, kmeans, 2 klastera

## Spektralno klasterovanje:

U slučaju sa dva klastera može se primetiti da fajlovi 2699155, 3140916 i 3195456 imaju slične veličine klastera u odnosu na veličine fajla i donekle slične brojve specifičnih kolona. Takođe, imaju jako dobre ocene za ovo klasterovanje (mada mora se primetiti da 2699155 ima izuzetno dobre ocene- senka je 0.95, dok je za 3140916 i 3195456 senka 0.79 i 0.87). U sva tri fajla, veći klaster nema nula kolone.

Ocene fajla 2699155 naglo opadaju sa povećanjem broja klastera na 3. Treći klaster sadrži 258 celija iz većeg klastera u slučaju 2 klastera. Taj klaster ima 510 nula i 7 ne-nula kolona.

Ocene fajlova 3140916, 3195456 i 3488509 naglo opadaju sa povećanjem broja klastera na 4. Treći klaster sadrži po 85, 76 i 372 celija iz većeg klastera u slučaju 2 klastera za svaki fajl. U fajlu 3488509, jedna celija iz manjeg klastera prelazi u veći, i time se broj ne-nula kolona manjeg klastera povećao za 1.

Ocene ovog klasterovanja su uglavnom lošije od ocena kmeans-a. Fajl 2699155 ima jako dobru ocenu za 2 klastera (senka je 0.95). Fajlovi 3140916 i 3195456 imaju jako dobre ocene za 2 i 3 klastera (senke su 0.79 i 0.8, 0.87 i 0.69). Fajl 3488509 ima solidne ocene za 2 i 3 klastera (senke su 0.61 i 0.63). Sve ostale ocene su loše. Fajl 3852754 ima relativno loše ocene kod oba algoritma što verovatno znači da nije pogodan za ovakve vrste klasterovanja.

## E15.5

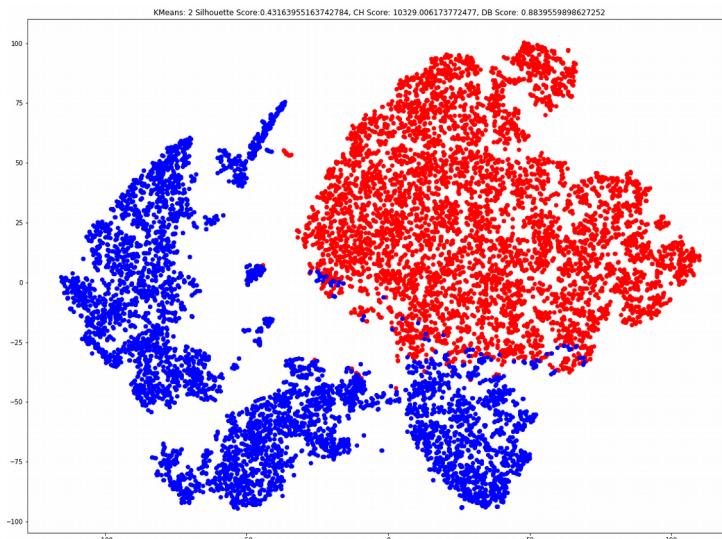
Celije ove grupe su uzete 15 dana nakon začeća. Ovoj grupi pripada samo fajl 3852755. On ima 11431 redova, 27 nula i 42 ne-nula kolona.

### Kmeans:

Za dva klastera, formiraju se klasteri veličine 6091 i 5340. Njihove nula i ne-nula kolone su 30 i 181, 5 i 42. Za tri klastera, formiraju se klasteri veličine 5017, 1833 i 4581. Njihove nula i ne-nula kolone su 6 i 42, 73 i 363, 57 i 168. Treći klaster se sastoji od određenog broja celija iz prvog i

drugog klastera iz prethodnog slučaja, dok se prvi i drugi sastoje samo od ćelija odgovarajućih klastera.

Što se više povećava broj klastera, broj specifičnih kolona po klasteru raste, što ukazuje na bolju razdvojenost klastera. To je kontradiktorno lošijim ocenama za veće brojeve klastera. Najbolja ocena je bila za 2 klastera (senka je 0.43).



Slika 10: Fajl 3852755, kmeans, 2 klastera

### Spektralno klasterovanje:

Za slučaj od 2 klastera, veličine klastera su 4035 i 7396. Klasteri imaju po 7 i 31 nula kolona i po 46 i 108 ne-nula kolona. Za slučaj od 3 klastera, treći klaster se formira od 90 ćelija iz manjeg klastera u prethodnom slučaju. Takođe, 6 ćelija iz većeg klastera prelaze u manji klaster. Novi klaster ima 1092 nula kolone i 244 ne-nula kolona.

Ocene su slabije nego za kmeans. Značajno su manje ocene za 4, 5 i 6 klastera. Najbolje ocene su za 3 klastera (senka je 0.35), a najgore su za 5 (senka je 0.17).

## E17.5

Ovoj grupi pripadaju fajlovi 2699157, 3140917 i 3140918. Zajedničke nula kolone su im M08631#Tgfb1i1, M45330#4933402E13Rik i M58665#En1. Nemaju zajedničke ne-nula kolone jer fajl 2699157 nema ne-nula kolone. Fajlovi 3140917 i 3140918 imaju po jednu ne-nula kolonu i ona je zajednička: M52305#Hbb-bs.

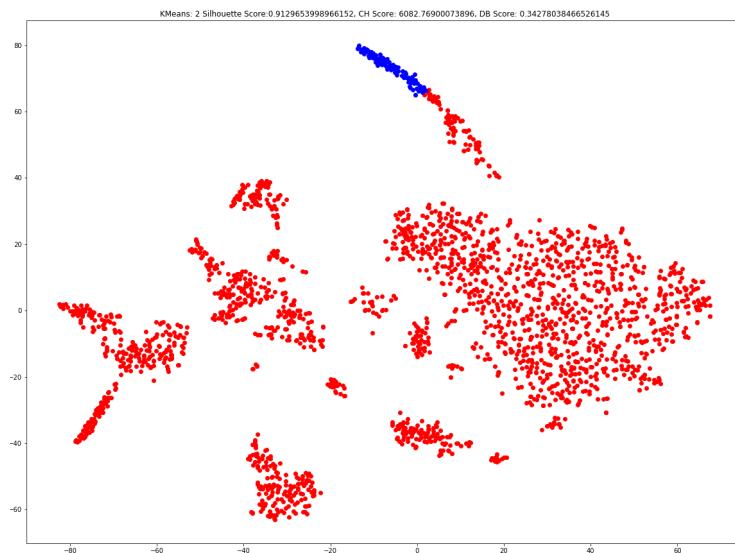
### Kmeans:

U slučaju sa tri klastera, u fajlu 2699157 klaster veličine 2146 se razdvojio na klastera veličine 2119 i 27, dok se u fajlu 3140918 treći klaster formirao izdvajanjem 412 ćelija iz većeg i jedne ćelije iz manjeg klastera. Izdvajanjem te jedne ćelije iz klastera veličine 216, broj nula kolona se povećao za jedan. Dva od tri klastera u fajlu 3140917 su nastali mešavinom elemenata oba klastera iz slučaja sa dva klastera, dok je treći nastao izbacivanjem 8 ćelija iz klastera veličine 230. Tim izbacivanjem se broj nula kolona povećao za 13, a broj ne-nula kolona za 4.

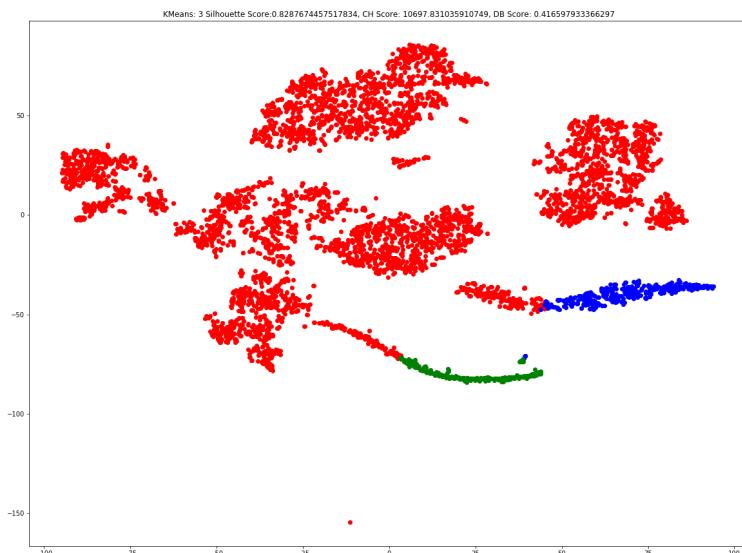
U slučaju sa 4 klastera, u fajlu 2699157 četvrti klaster sadrži 41 celiju iz klastera veličine 78 i 12 celija iz klastera veličine 2119. Izdvajanjem ovih 41 celija, broj nula kolona se povećava sa 1747 na 3226, a broj ne-nula kolona se povećava sa 12 na 26. U fajlu 3140917, četvrti klaster se formirao od 48 celija iz klastera veličine 5010 i 106 celija iz klastera veličine 222. U fajlu 3140918, četvrti klaster je nastao od 80, 48 i 2 celije iz klastera veličine 135, 5016 i 412. Fajlovi 3140917 i 3140918 imaju dosta slične klastere po veličini i brojevima specifičnih kolona. Takođe, njihovi klasteri međusobno imaju dosta zajedničkih specifičnih kolona.

Ostali slučajevi su previše komplikovani za analizu. Može se primetiti da su fajlovi 3140917 i 3140918 u svim slučajevima dosta slični međusobno po veličinama klastera i broju specifičnih kolona po klasteru. Klasteri međusobno dele dosta specifičnih kolona. Fajl 2699157 ne deli ove sličnosti sa druga dva fajla.

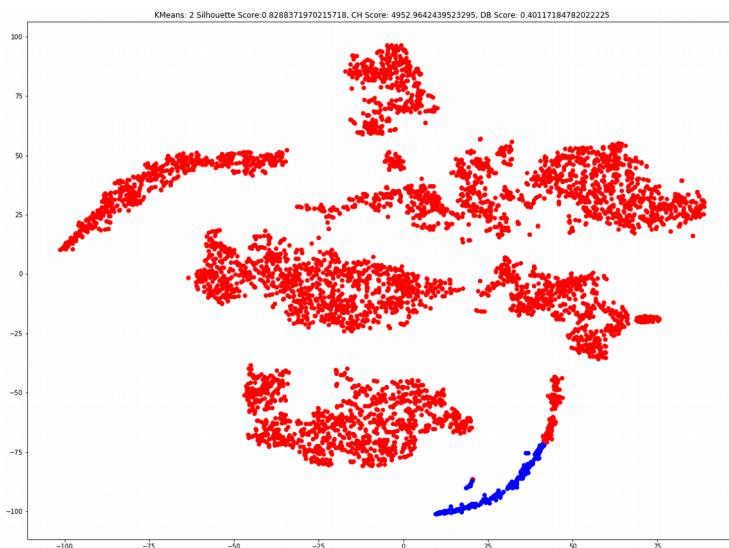
Ocene su odlične u svim fajlovima za sve brojeve klastera sem u fajlu 3140917 za 7 klastera koja je dosta lošija u odnosu na ostale (senka je 0.4). Primećuje se da ocena blago opada sa povećanjem broja klastera. Fajl 2699157 ima najbolje ocene za 2 klastera (senka je 0.91). Fajl 3140917 ima najbolje ocene za 3 klastera (senka je 0.83). Fajl 3140918 ima najbolje ocene za 2 klastera (senka je 0.83).



Slika 11: Fajl 2699157, kmeans, 2 klastera



Slika 12: Fajl 3140917, kmeans, 3 klastera



Slika 13: Fajl 3140918, kmeans, 2 klastera

### Spektralno klasterovanje:

Za slučaj sa 2 klastera dobijamo redom klastere veličina 2132 i 92, 641 i 4930, 4967 i 676. Lako je primetiti da su poslednja dva fajla napravila klastere slične veličine. Klasteri imaju po 0 i 1477, 60 i 0, i 0 i 39 nula kolona i 0 i 10, 12 i 4, i 5 i 8 ne-nula kolona. Za slučaj sa 3 klastera, u fajlu 2699157, treći klaster sadrži 56 celija većeg klastera iz slučaja 2 klastera, dok u druga dva fajla treći klaster sadrži po 297 i 267 celija manjeg klastera iz slučaja 2 klastera. U poslednja dva fajla, u veći klaster je otišlo 17 i 3 celija manjeg, a u manji je otišlo 57 i 52 celije većeg klastera.

Ocene za spektralno klasterovanje su dosta lošije od ocena za kmeans za broj klastera iznad 3. Za sva tri fajla ocene naglo opadaju za 4 klastera (pogotovo za fajl 2699157), dok su ocene za 2 i 3 klastera solidne (mada ipak malo lošije od ocena za kmeans).

# 3140920

Ovaj fajl ima 7897 čelija tj. redova. Ima 9 nula kolona i 38 ne-nula kolona. Nije poznato u kom danu embriogeneze su uzimani uzorci pa će se zato ovaj fajl samostalno analizirati.

## Kmeans

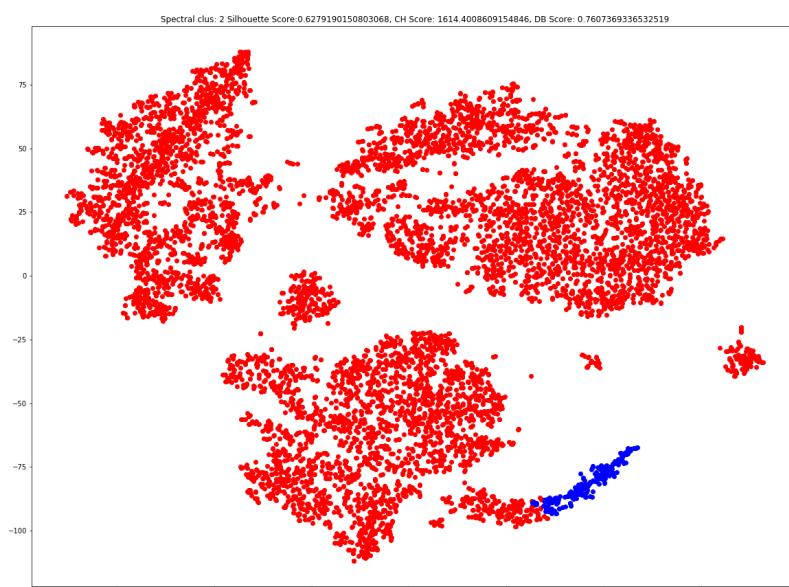
Kad je zadati broj klastera 2, formiraju se klasteri veličina 1190 i 6707. Prvi ima 13 nula kolona i 262 ne-nula kolone, a drugi ima 1 nula i 38 ne-nula kolona (koje su ne-nula za ceo fajl). Za zadati broj klastera jednak tri, formiraju se klasteri veličina 1186, 6562, 149 sa redom brojevima nula i ne-nula kolona 13 i 262, 1 i 38, 474 i 227. Prvi klaster se sastoji od elemenata iz prvog klastera prethodnog slučaja, tačnije ne sadrži samo 4 njegove čelije. Tri od te četiri su se pripojile drugom klasteru iz prethodnog slučaja, a jedna je se spojila sa 148 čelija koje su se izdvojile iz većeg klastera i formirale novi klaster. Zanimljivo je da se brojevi specifičnih kolona nisu promenile za dva veća klastera.

Klasteri koji se formiraju za 4 klastera su veličina 1328, 6040, 149 i 380. Brojevi nula i ne-nula kolona su im redom 13 i 194, 2 i 38, 474 i 227, 30 i 503. Prvi klaster je nastao od 806 čelija prvog i 522 čelije iz drugog klastera prethodnog slučaja. Svih 194 ne-nula i 7 nula kolona deli sa prvim klasterom iz prethodnog slučaja. Drugi klaster se sastoji od čelija iz drugog, treći od čelija iz trećeg, a četvrti od čelija prvog klastera iz prethodnog slučaja.

Brojevi specifičnih kolona se povećavaju sa povećanjem broja klastera. To bi moglo da ukazuje na bolju razdvojenost klastera, iako je najbolja ocena za tri klastera (senka je 0.54). Ocene za 2, 3 i 4 klastera su solidne dok su za 5, 6 i 7 lošije (senka je malo iznad 0.2).

## Spektralno klasterovanje:

Za slučaj sa 2 klastera dobijamo redom klastera veličina 202 i 7686. Klasteri imaju po 347 i 0 nula kolona i 194 i 38 ne-nula kolona. Za slučaj sa 3 klastera, treći klaster sadrži 1437 čelija većeg klastera iz slučaja 2 klastera, dok je u većim klasterima otišlo 49 čelija manjeg klastera.



Slika 14: Fajl 3140920, spektralno, 2 klastera

Najbolje ocene su za 2 klastera (senka je 0.63). Ocene se blago razlikuju od ocena za kmeans, sem za 4 klastera gde je ocena za spektralno klasterovanje dosta lošija (senka za kmeans je 0.42, a za spektralno je 0.2).

## Grupa 2699154, 2699155 i 3140916

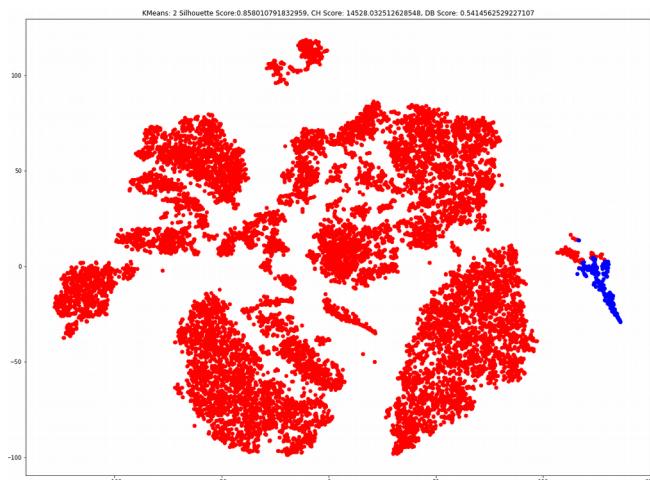
Ova grupa se sastoji od spojenih tabela koje sadrže podatke iz 14. dana nakon začeća. Ima 16810 redova, 8 nula kolona i 0 ne-nula kolona.

### Kmeans:

U slučaju 2 klastera, klasteri su veličine 16490 i 320 i imaju po 3478 i 17, 4229 i 77, 8783 i 226 redova iz svakog fajla, 0 i 117 nula kolona i 0 i 8 ne-nula kolona. Veći klaster sadrži približno srazmerno redova iz svakog fajla, dok manji klaster ima dosta više redova iz fajla 3140916. Taj klaster je jako mali i specifičan. S obzirom da je dobra ocena ovog klasterovanja, ta grupa celija mora da je sa razlogom izdvojena i preporučuje nekom stručnom licu su da detaljnije pogleda tih 320 celija.

U slučaju 3 klastera, klasteri su veličine 12923, 324 i 3563 i imaju po 3388, 17 i 90, 4174, 77 i 55, 5361, 230 i 3418 redova iz svakog fajla, 0, 114 i 21 nula kolona i 0, 8 i 129 ne-nula kolona. Zanimljivo je primetiti da srednji klaster ima ukupno više specifičnih kolona od najmanjeg klastera. Najmanji klaster ima dosta više nula kolona ali prilično manje ne-nula kolona od srednjeg klastera. Treći klaster je nastao izdvajanjem 3563 celija iz većeg klastera iz slučaja sa 2 klastera. Takođe, 4 celije većeg klastera su se pripojile manjem klasteru. Te 4 celije su iz fajla 3140916.

Grupa ima najbolje ocene za 2 klastera (senka je 0.86). Sve ostale ocene su dosta lošije (senka za sve ostale brojeve klastera je 0.42-0.47).



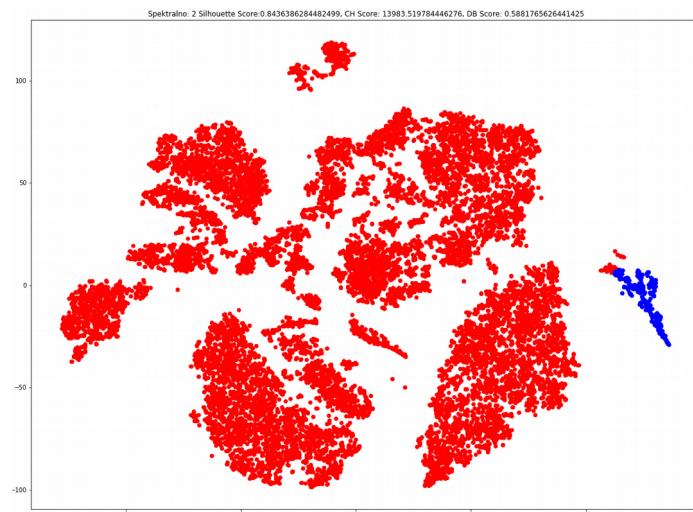
Slika 15: Grupa fajlova 2699154, 2699155 i 3140916, kmeans, 2 klustera

### Spektralno klasterovanje:

U slučaju 2 klastera, klasteri su veličine 16417 i 393 i imaju po 3465 i 30, 4225 i 81, 8727 i 282 redova iz svakog fajla, 0 i 93 nula kolona i 0 i 8 ne-nula kolona. Kao i u slučaju kmeans-a sa 2 klastera, preporučuje nekom stručnom licu su da detaljnije pogleda celija iz manjeg klastera s obzirom da se izdvojio i da klasterovanje daje prilično dobру ocenu.

U slučaju 3 klastera, klasteri su veličine 8827, 7590 i 393 i imaju po 56, 3409 i 30, 51, 4174 i 81, 8720, 7 i 282 redova iz svakog fajla, 3, 13 i 93 nula kolona i 2, 0 i 8 ne-nula kolona. Treći klaster je nastao od 7590 celija iz većeg klastera iz prošlog slučaja. Ovo klasterovanje se razlikuje od kmeans-a za isti broj klastera po slabijim ocenama (senka je 0.18), po manje specifičnih kolona po klasteru i po zastupljenosti celija iz jednog fajla u jednom klasteru. Klasteri su u ovom klasterovanju dosta više podeljeni po fajlovima (najveći klaster se skoro ceo sastoji od celija iz fajla 3140916, najmanji klaster je otprilike tri četvrtine od celija iz fajla 3140916, dok se srednji klaster skoro skroz sastoji od celija iz ostala dva fajla).

Grupa ima najbolje ocene za 2 klastera (senka je 0.84). Sve ostale ocene su dosta lošije (senka za sve ostale brojeve klastera je 0.18-0.21). Ocene su (osim za 2 klastera) dosta lošije u odnosu na ocene za kmeans.



*Slika 16: Grupa fajlova 2699154, 2699155 i 3140916, spektralno, 2 klastera*

## Grupa 2699157, 3140917 i 3140918

Ova grupa se sastoji od spojenih tabela koje sadrže podatke iz 17. dana nakon začeća. Ima 13438 redova, 3 nula kolona i 0 ne-nula kolona.

### Kmeans:

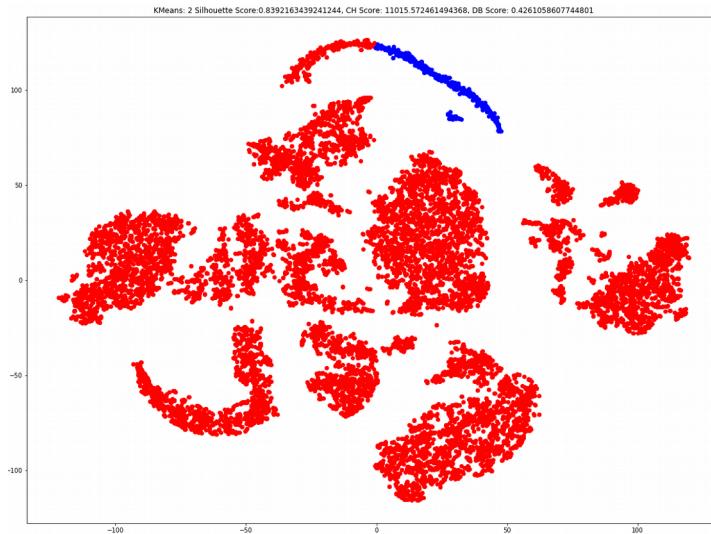
U slučaju sa 2 klastera, klasteri su veličine 12988 i 450 i imaju po 2223 i 1, 5338 i 233, 5427 i 216 redova iz svakog fajla, 0 i 145 nula kolona i 0 i 163 ne-nula kolona. U veći klaster su otišle sve celije fajla 2699157 (samo je jedna zalutala u manji klaster). Druga dva fajla su se slično podelila između dva klastera. S obzirom na kvalitet ocena i to koliko je manji drugi klaster, bilo bi korisno u daljoj analizi da se detaljnije prouče celije tog klastera.

U slučaju sa 3 klastera, klasteri su veličine 12186, 438 i 814 i imaju po 2174, 0 i 50, 4988, 224 i 359, 5024, 214 i 405 redova iz svakog fajla, 0, 150 i 47 nula kolona i 0, 235 i 5 ne-nula kolona. Treći klaster se sastoji od 807 celija većeg i 7 celija manjeg klastera iz slučaja sa 2 klastera. Takođe, 5 celija iz manjeg je prešlo u veći klaster.

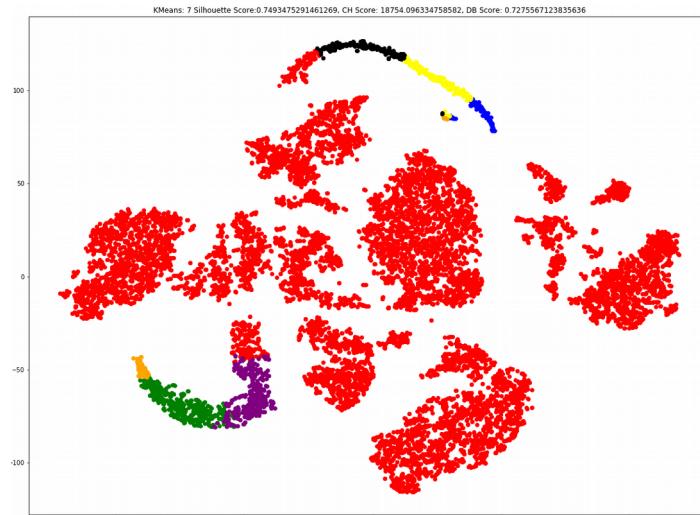
U slučaju sa 4 klastera, klasteri su veličine 12088, 245, 812 i 293 i imaju po 2170, 0, 50 i 4,

4942, 124, 359 i 146, 4976, 121, 403 i 143 redova iz svakog fajla, 0, 251, 47 i 213 nula kolona i 0, 295, 5 i 108 ne-nula kolona. Četvrti klaster se sastoji od 98, 193 i 2 čelija iz najvećeg, najmanjeg i srednjeg klastera iz slučaja sa 3 klastera.

Grupa ima najbolje ocene za 2 klastera (senka je 0.84). Sve ostale ocene su slične i isto prilično dobre (senka za sve ostale brojeve klastera je 0.75-0.83).



Slika 17: Grupa fajlova 2699157, 3140917 i 3140918, kmeans, 2 klastera



Slika 18: Grupa fajlova 2699157, 3140917 i 3140918, kmeans, 7 klastera

## Spektralno klasterovanje:

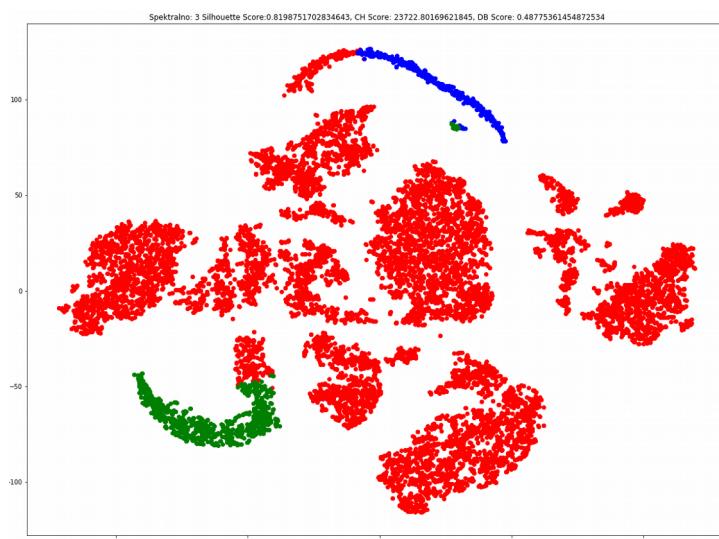
U slučaju 2 klastera, klasteri su veličine 12094 i 1344 i imaju po 2172 i 52, 4930 i 641, 4992 i 651 redova iz svakog fajla, 0 i 33 nula kolona i 0 i 3 ne-nula kolona. Isti broj klastera napravljen kmeans-om ima više ne-nula kolona po klasteru.

U slučaju 3 klastera, klasteri su veličine 11992, 496 i 950 i imaju po 2144, 4 i 76, 4912, 251 i 408, 4936, 241 i 466 redova iz svakog fajla, 0, 137 i 38 nula kolona i 0, 122 i 4 ne-nula kolona. Treći klaster se sastoji od 751 i 199 čelija iz oba klastera, a 97 čelija manjeg klastera je prešlo u veći

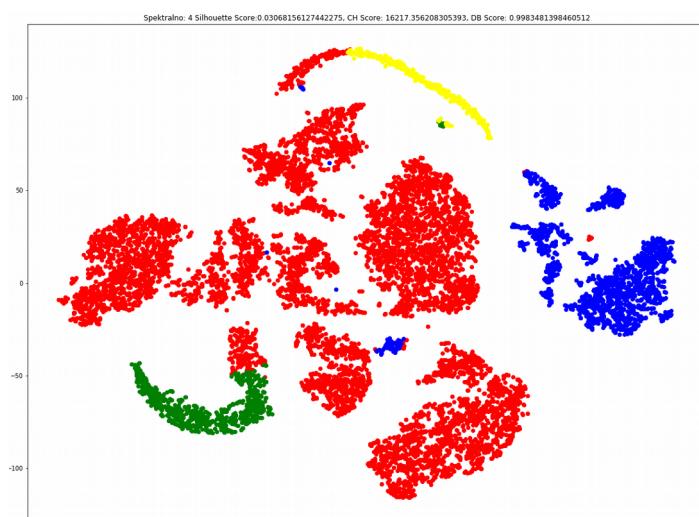
klaster iz prethodnog slučaja. Kao i u prethodnom slučaju, većina čelija iz fajla 2699157 se nalazi u najvećem klasteru.

U slučaju sa 4 klastera, klasteri su veličine 9955, 2038, 948 i 497 i imaju po 129, 2015, 76 i 4, 4902, 11, 406 i 252, 4924, 12, 466 i 241 redova iz svakog fajla, 0, 73, 38 i 137 nula kolona i 0, 0, 4 i 122 ne-nula kolona. Četvrti klaster se sastoji od 2038 čelija iz najvećeg klastera iz slučaja sa 3 klastera. Od tih 2038 čelija, 2015 čelija je iz fajla 2699157 što je malo iznad 90% čelija iz fajla i malo iznad 98% klastera. Takođe, jedna čelija najvećeg klastera je prešla u najmanji, a dve čelije srednjeg klastera su prešle u najveći klaster. Nakon povećanja broja klastera sa 3 na 4, ocene naglo opadaju (senka za 4 klastera je 0.03). To znači da ovakva podela klastera po fajlovima nije odgovarajuća.

Grupa ima najbolje ocene za 3 klastera (senka je 0.82). Sve ostale ocene, osim za 2 klastera (senka je 0.8), su dosta lošije (senka za sve ostale brojeve klastera je 0.03-0.12).



Slika 19: Grupa fajlova 2699157, 3140917 i 3140918,  
spektralno, 3 klastera



Slika 20: Grupa fajlova 2699157, 3140917 i 3140918,  
spektralno, 4 klastera

## Grupa 3195456 i 3488509

Ova grupa se sastoji od spojenih tabela koje sadrže podatke iz 14. dana nakon začeća. Ima 10304 redova, 13 nula kolona i 9 ne-nula kolona.

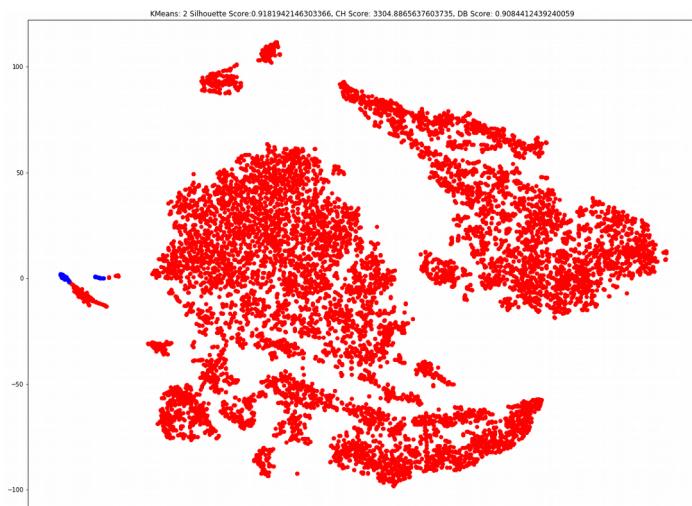
### Kmeans:

U slučaju 2 klastera, klasteri su veličine 10259 i 45 i imaju po 0 i 1710 nula kolona i 13 i 49 ne-nula kolona. Manji klaster se u potpunosti sastoji od čelija iz fajla 3195456. Preporučuje nekom stručnom licu su da detaljnije pogleda čelija iz manjeg klastera s obzirom da klasterovanje ima odlične ocene.

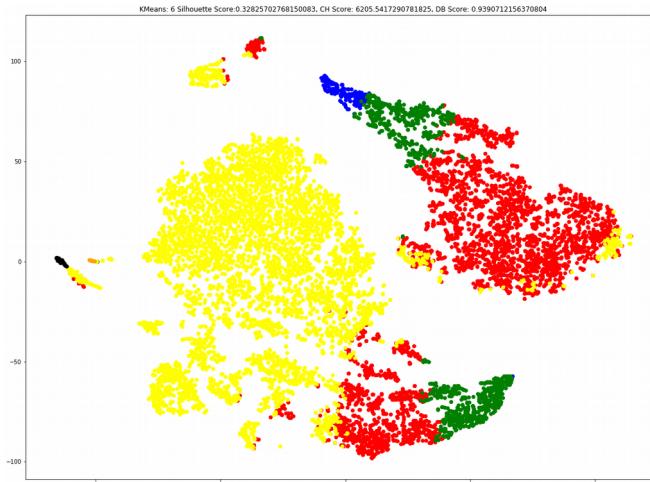
U slučaju 3 klastera, klasteri su veličine 730, 9529 i 45 i imaju po 522, 2860 i 45, 208, 6669 i 0 redova iz svakog fajla, 13, 1 i 1710 nula kolona i 217, 13 i 49 ne-nula kolona. Treći klaster se sastoji od 730 čelija većeg klastera iz prethodnog slučaja. Većina čelija iz oba fajla se nalazi u najvećem klasteru (malo više od 83% čelija iz fajla 3195456 i oko 97% čelija iz fajla 3488509).

U slučaju 4 klastera, klasteri su veličine 9618, 640, 13 i 33 i imaju po 2893, 489, 13 i 32, 6725, 151, 0 i 1 redova iz svakog fajla, 1, 14, 4976 i 2358 nula kolona i 13, 230, 142 i 48 ne-nula kolona. Četvrti klaster se sastoji od 29 i 4 čelija najmanjeg i najvećeg klastera iz prethodnog slučaja. Takođe, 90 čelija iz srednjeg i 3 iz najmanjeg su se pripojile najvećem klasteru. Većina čelija iz oba fajla se nalazi u najvećem klasteru (malo više od 84% čelija iz fajla 3195456 i oko 98% čelija iz fajla 3488509). Najmanja dva klastera se skoro skroz sastoje od čelija iz fajla 3195456, a malo više od  $\frac{3}{4}$  klastera veličine 640 su čelije iz fajla 3195456.

Ocene su jako dobre za broj klastera do 5 (senke za 2, 3 i 4 klastera su 0.92, 0.69 i 0.71). Nakon 5, ocene polako padaju (senke za 5, 6 i 7 klastera su 0.54, 0.33 i 0.34).



Slika 21: Grupa fajlova 3195456 i 3488509, kmeans,  
2 klastera

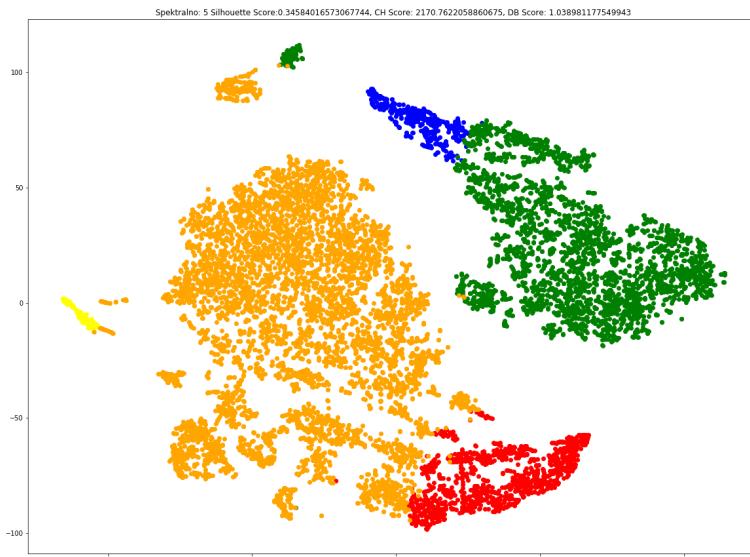


Slika 22: Grupa fajlova 3195456 i 3488509,  
kmeans, 6 klastera

### Spektralno klasterovanje:

U slučaju 2 klastera, klasteri su veličine 6945 i 3359 i imaju po 368 i 3359, 6877 i 0 redova iz svakog fajla, 2 i 6 nula kolona i 10 i 92 ne-nula kolona. U slučaju 3 klastera, klasteri su veličine 6840, 3359 i 105 i imaju po 35, 3359 i 33, 6805, 0 i 72 redova iz svakog fajla, 2, 6 i 569 nula kolona i 30, 92 i 23 ne-nula kolona. Treći klaster se sastoji od 105 celija iz većeg klastera iz prethodnog slučaja.

Svi dalji slučajevi su slični po ocenama i po tome što su klasteri homogeni u smislu pripadnosti jednom fajlu. Ta odlika se ne odražava dobro na ocene. Ocene su slične za sve brojeve klastera i relativno loše (senka je uvek između 0.24-0.35). Najbolja ocena je za 5 klastera (senka je 0.35).



Slika 23: Grupa fajlova 3195456 i 3488509, spektralno, 5  
klastera

### Grupa 3852752, 3852753, 3852754 i 3852755

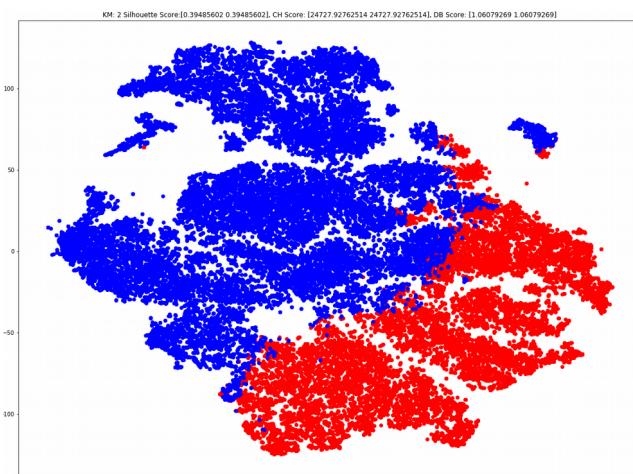
Ova grupa se sastoji od spojenih tabela koje sadrže podatke iz 12. (3852752), 13. (3852753), 14.

(3852754) i 15. dana (3852755) nakon začeća. Ima 37966 redova, 3 nula kolona i 22 ne-nula kolone.

## Kmeans:

U slučaju 2 klastera, klasteri su veličine 14711 i 23255 i imaju po 3769 i 7413; 1861 i 3305; 3069 i 7118; 6012 i 5419 čelija iz svakog fajla, 7 i 2 nula kolone i 163 i 22 ne-nula kolone. U slučaju 3 klastera, klasteri su veličine 11304, 22008 i 4654 i imaju po 4496, 6686 i 0; 2089, 3066 i 0; 2973, 6946 i 268; 1746, 5310 i 4375 redova iz svakog fajla, 9, 2 i 58 nula kolona i 190, 22 i 203 ne-nula kolona. Može se primetiti da najmanji klaster ima samo čelije iz fajlova 3852754 i 3852755 koji sadrže podatke uzete 14 i 15 dana nakon začeća. Čelije fajlova 3852752 i 3852753 su srazmerno podeljene između najvećeg i srednjeg klastera (razmera je približno 2:3).

Ocene su loše za sve brojeve klastera (senke je između 0.18 i 0.4). Najbolje ocene su za 2 klastera (senka je 0.39).

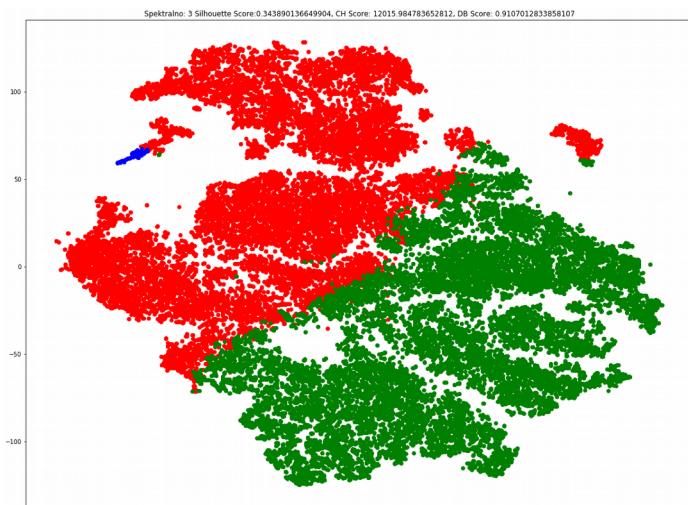


Slika 24: Grupa fajlova 3852752, 3852753, 3852754 i 3852755, kmeans, 2 klastera

## Spektralno klasterovanje:

U slučaju 2 klastera, klasteri su veličine 20722 i 17244 i imaju po 6675 i 4507, 2811 i 2355, 4020 i 6167, 7216 i 4215 redova iz svakog fajla, 4 i 3 nula kolona i 109 i 22 ne-nula kolona. Zanimljivo je da veći klaster ima više specifičnih kolona. U slučaju 3 klastera, klasteri su veličine 18772, 124 i 19070 i imaju po 5262, 0 i 5920; 2707, 0 i 2459; 6479, 44 i 3664; 4324, 80 i 7027 redova iz svakog fajla, 3, 808 i 6 nula kolona i 22, 215 i 115 ne-nula kolona. Treći klaster se sastoji od 124 čelije manjeg klastera iz slučaja sa 2 klastera. U veći klaster je prešlo 6 čelija manjeg, a u manji klaster je prešlo 1658 čelija većeg klastera iz slučaja sa 2 klastera. U najmanjem klasteru se nalaze samo čelije fajlova 3852754 i 3852755 koji sadrže podatke prikupljene 12. i 13. dana nakon začeća. Čelije fajlova 3852752 i 3852753 su približno slično podeljeni po klasterima kao i čelije 3852754 i 3852755.

Ocene su loše za sve brojeve klastera, čak malo lošije od ocena za kmeans (senka je između 0.18 i 0.34). Najbolje ocene su za 3 klastera (senka je 0.34).



*Slika 25: Grupa fajlova 3852752, 3852753, 3852754 i 3852755, spektralno, 3 klastera*

## **Grupa 3140915, 3140916, 3140917 i 3140918**

Ova grupa se sastoji od spojenih tabela koje sadrže podatke iz 12. (3140915), 14. (3140916) i 17. dana (3140917 i 3140918) nakon začeća. Ima 32905 redova, 3 nula kolona i 0 ne-nula kolone.

### **Kmeans:**

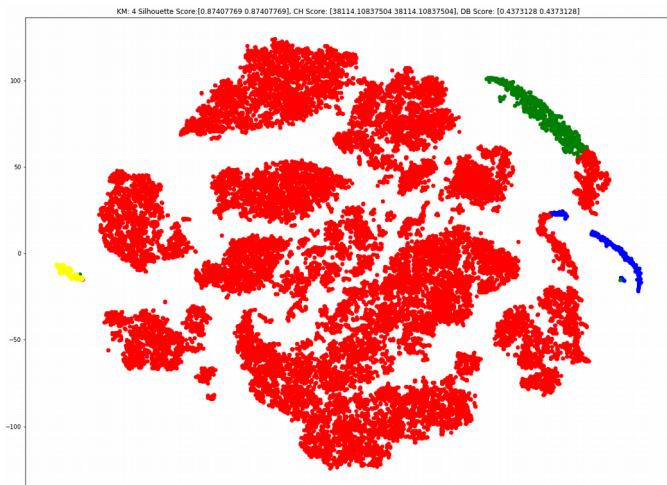
U slučaju 2 klastera, klasteri su veličine 32451 i 454 i imaju po 12682 i 0; 9006 i 3; 5337 i 234; 5426 i 217 celija iz svakog fajla, 0 i 140 nula kolone i 0 i 179 ne-nula kolone. U slučaju 3 klastera, klasteri su veličine 31373, 445 i 1087 i imaju po 12448, 0 i 234; 8945, 3 i 61; 4974, 227 i 370; 5006, 215 i 422 redova iz svakog fajla, 0, 146 i 29 nula kolona i 0, 179 i 6 ne-nula kolona. Može se primetiti da najmanji klaster ima samo celije iz fajlova 3140917 i 3140918 koji sadrže podatke uzete 17 dana nakon začeća. Celije fajlova 3140915 i 3140916 su srazmerno podeljene između najvećeg i srednjeg klastera (u veći klaster je otišlo 98-99% celija). Treći klaster sadrži 1078 celija većeg i 9 klastera iz manjeg klastera iz slučaja sa 2 klastera.

U slučaju 4 klastera, klasteri su veličine 31361, 445, 920 i 179 i imaju po 12420, 0, 83 i 179; 8950, 3, 56 i 0; 4978, 227, 366 i 0; 5013, 215, 415 i 0 redova iz svakog fajla, 0, 146, 35 i 345 nula kolona i 0, 179, 7 i 29 ne-nula kolona. Može se primetiti da najmanji klaster ima samo celije iz fajla 3140915 koji sadrže podatke uzete 12 dana nakon začeća. Klaster veličine 920 se većinski sastoji od celija fajlova 3140917 i 3140918 (izuzetak su 139 celija), dok se klaster veličine 445 skoro skroz sastoji od celija iz istih fajlova (izuzetak su 3 celije iz fajla 3140916). Veći klaster sadrži 18 celija srednjeg klastera iz slučaja sa 3 klastera. Četvrti klaster sadrži 149 celija srednjeg i 30 celija većeg klastera iz slučaja sa 3 klastera.

U slučaju 5 klastera, klasteri su veličine 31221, 918, 179, 265 i 322 i imaju po 12420, 83, 179, 0 i 0; 8916, 56, 0, 0 i 37; 4924, 366, 0, 130 i 151; 4961, 413, 0, 135 i 134 redova iz svakog fajla, 0, 35, 345, 228 i 202 nula kolona i 0, 7, 29, 285 i 105 ne-nula kolona. Najmanji klaster je isti kao najmanji klaster iz slučaja sa 4 klastera i ima samo celije iz fajla 3140915 koji sadrže podatke uzete 12 dana nakon začeća. Klasteri veličine 918 i 322 se većinski sastoje od celija fajlova 3140917 i

3140918 (izuzetak su 139 i 37 celija), dok se klaster veličine 265 skroz sastoji od celija iz istih fajlova. Četvrti klaster sadrži 140, 180 i 2 celije klastera veličina 31361, 445 i 920 iz slučaja sa 3 klastera.

Ocene su prilično dobre za sve brojeve klastera osim za 7 (senke je između 0.8 i 0.89 osim za 7 klastera gde je 0.4). Najbolje ocene su za 2 klastera (senka je 0.89).



Slika 26: Grupa fajlova 3140915, 3140916, 3140917 i 3140918, kmeans, 4 klastera

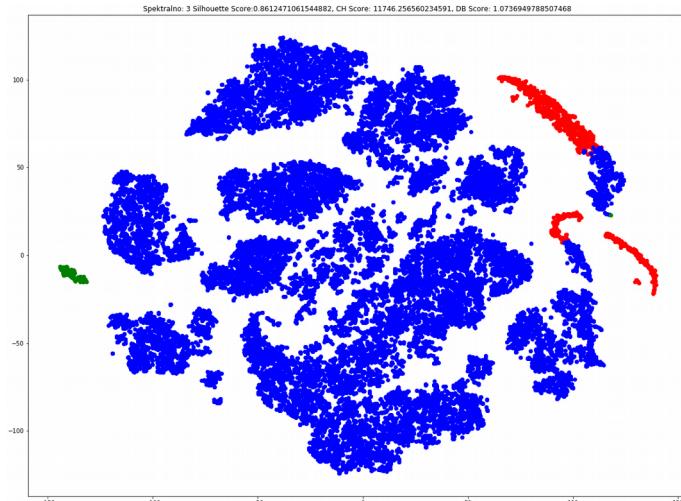
## Spektralno klasterovanje:

U slučaju 2 klastera, klasteri su veličine 31362 i 1543 i imaju po 12600 i 82, 8899 i 110, 4906 i 665, 4957 i 686 redova iz svakog fajla, 0 i 27 nula kolona i 0 i 2 ne-nula kolone. U slučaju 3 klastera, klasteri su veličine 1545, 31177 i 183 i imaju po 82, 12418 i 182; 110, 8898 i 1; 666, 4905 i 0; 687, 4956 i 0 redova iz svakog fajla, 27, 0 338 nula kolona i 2, 0 i 28 ne-nula kolona. Treći klaster se sastoji od 183 celije većeg klastera iz slučaja sa 2 klastera. Takođe, 2 celije većeg su prešle u manji klaster iz slučaja sa 2 klastera. Najmanji klaster se sastoji većinski od celija fajla 3140915 (samo je jedna celija iz 3140916), dok se srednji klaster najviše sastoji od celija iz fajlova 3140917 i 3140918 koji sadrže podatke uzete 17. dana nakon začeća. Celije fajlova 3140917 i 3140918 su u sličnoj razmeri podeljene između klastera. Najveći klaster sadrži sličan procenat celija iz fajlova 3140915 i 3140916 (97-98%).

U slučaju 4 klastera, klasteri su veličine 30951, 1285, 183 i 486 i imaju po 12409, 91, 182 i 0; 8740, 258, 1 i 10; 4891, 432, 0 i 248; 4911, 504, 0 i 228 redova iz svakog fajla, 0, 25, 338 i 131 nula kolona i 0, 6, 28 i 141 ne-nula kolona. Četvrti klaster se sastoji od 486 celija srednjeg klastera iz slučaja sa 3 klastera. Takođe, 138 celija srednjeg su prešle u veći, a 364 celije većeg su prešle u srednji klaster iz slučaja sa 3 klastera. Najmanji klaster je isti kao najmanji iz slučaja sa 3 klastera. Klaster veličine 486 se sastoji većinski od celija fajlova 3140917 i 3140918 (samo je 10 celija iz 3140916), koji sadrže podatke uzete 17. dana nakon začeća. Celije fajlova 3140917 i 3140918 su u sličnoj razmeri podeljene između klastera. Najveći klaster sadrži sličan procenat celija iz fajlova 3140915 i 3140916 (97-98%).

Ocene su prilično dobre za brojeve klastera do 4, iako su za nijansu slabije od ocena kmeans-a za iste brojeve klastera (senka za 2, 3 i 4 klastera su 0.85, 0.86 i 0.85). Ocene za 5 klastera i više su

dosta slabije (senka za 5, 6 i 7 su 0.04, 0.14 i 0.16). Najbolje ocene su za 3 klastera (senka je 0.86).



Slika 27: Grupa fajlova 3140915, 3140916, 3140917 i 3140918, spektralno, 3 klastera

## Zaključak

Najbolje ocene imaju fajlovi sa čelijama uzetim 17 dana nakon začeća. To je i očekivano, s obzirom da se pre toga malo čelija diferencira. Izuzetno dobre ocene ima kmeans klasterovanje svih podataka koji su uzeti 17. dana nakon začeća iz [1]. Takođe je primetno da ocene u svim fajlovima opadaju sa povećanjem broja klastera. To se može objasniti ili nezavršenom diferencijacijom ili nedovoljnom razlikom između podtipova čelija.

Klasterovanje kmeans-om nad spojenim tabelama je dalo dobre rezultate sem za fajlove iz [3]. To je možda uzrokovano činjenicom da su podaci uzimani dan za danom (12, 13, 14 i 15 dana) čime je razlika u podacima nedovoljna da se napravi značajna podela, što nije slučaj za fajlove iz [1] gde su podaci uzimani sa većim korakom (12, 14 i 17 dana). Klasteri napravljeni nad spojenim tabelama koje su sadržale podatke iz istih stupnjeva razvoja embriona su bili heterogeniji po udelu pripadnosti fajlu od klastera napravljenih nad spojenim tabelama koje su sadržale podatke iz različitih stupnjeva razvoja embriona.

Kmeans je dao bolje rezultate od spektralnog klasterovanja. Poređenjem ocena ova dva algoritma, može se primetiti da je za isti broj klastera bolja podela podataka ona koja u jedan klaster stavlja veliku većinu čelija, dok su ostali klasteri dosta manji. Ovo se može objasniti time što većina čelija nije počela diferenciranje ili je u stupnju diferencijacije u kom se ne razlikuje dovoljno od ostalih čelija. Najbolje ocene se dobijaju za 2 ili 3 klastera.

Što se tiče daljeg istraživanja bilo bi korisno uporediti klastere fajlova koji sadrže čelije iz različitih dana nakon začeća. Time bi se mogla detaljnije istražiti diferencijacija čelija.

	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>
<b>2699154</b>	0.84	0.5	0.5	0.51	0.43	0.2
<b>2699155</b>	0.96	0.94	0.9	0.89	0.3	0.3
<b>2699156</b>	0.97	0.96	0.96	0.96	0.94	0.94
<b>2699157</b>	0.91	0.87	0.87	0.87	0.82	0.79
<b>3140915</b>	0.92	0.9	0.53	0.37	0.37	0.29
<b>3140916</b>	0.79	0.8	0.39	0.39	0.36	0.37
<b>3140917</b>	0.82	0.83	0.81	0.79	0.73	0.4
<b>3140918</b>	0.83	0.82	0.81	0.77	0.74	0.72
<b>3140919</b>	0.51	0.54	0.42	0.22	0.2	0.21
<b>3140920</b>	0.52	0.54	0.42	0.26	0.2	0.2
<b>3195456</b>	0.89	0.71	0.74	0.61	0.61	0.48
<b>3488509</b>	0.64	0.65	0.5	0.4	0.32	0.32
<b>3852752</b>	0.38	0.3	0.3	0.23	0.24	0.2
<b>3852753</b>	0.39	0.4	0.28	0.2	0.17	0.13
<b>3852754</b>	0.5	0.34	0.21	0.22	0.2	0.16
<b>3852755</b>	0.43	0.35	0.33	0.34	0.3	0.19

Tabela 1: senke za kmeans

	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>
<b>2699154</b>	0.49	0.48	0.50	0.34	0.25	0.15
<b>2699155</b>	0.95	0.36	0.38	0.39	0.31	0.27
<b>2699156</b>	0.96	0.46	0.50	0.50	0.14	0.14
<b>2699157</b>	0.90	0.81	0.16	0.17	0.17	0.15
<b>3140915</b>	0.92	0.90	0.29	0.40	0.31	0.30
<b>3140916</b>	0.79	0.80	0.36	0.25	0.24	0.08
<b>3140917</b>	0.78	0.82	0.34	0.35	0.19	0.19
<b>3140918</b>	0.78	0.81	0.34	0.35	0.35	0.21
<b>3140919</b>	0.63	0.49	0.20	0.19	0.17	0.18
<b>3140920</b>	0.63	0.49	0.20	0.19	0.17	0.18
<b>3195456</b>	0.87	0.69	0.31	0.38	0.42	0.33
<b>3488509</b>	0.61	0.63	0.38	0.39	0.41	0.23
<b>3852752</b>	0.35	0.35	0.29	0.28	0.23	0.18
<b>3852753</b>	0.34	0.23	0.19	0.20	0.21	0.14
<b>3852754</b>	0.47	0.25	0.24	0.25	0.21	0.19
<b>3852755</b>	0.35	0.35	0.23	0.17	0.17	0.17

Tabela 2: senke za spektralno

## Literatura

- [1] Xin-Xin Yu, Wei-Lin Qiu, Cheng-Ran Xu, Liu Yang, Yu Zhang, Mao-Yang He, Lin-Chen Li: Defining multistep cell fate decision pathways during pancreatic development at single-cell resolution
- [2] Lauren E. Byrnes, Daniel M. Wong, Meena Subramaniam, Nathaniel P. Meyer, Caroline L. Gilchrist, Sarah M. Knox, Aaron D. Tward, Chun J. Ye, Julie B. Sneddon: Lineage dynamics of murine pancreatic development at single-cell resolution
- [3] Aimée Bastidas-Ponce, Sophie Tritschler, Leander Dony, Katharina Scheibner, Marta Tarquis-Medina, Ciro Salinno, Silvia Schirge, Ingo Burtscher, Anika Böttcher, Fabian J. Theis, Heiko Lickert, Mostafa Bakhti: Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis

# Dodatak - kodovi

## Preprocessing

August 27, 2020

```
[31]: import pandas as pd
import numpy as np
from functools import reduce

[32]: common_mouse_list_data = pd.read_csv("common_mouse_list.csv")
common_mouse_list_filtered = common_mouse_list_data.loc[:, ["ENSMUSG_ID", "mm10"]]

sum_num_of_rows = 0
common_mouse_list_filtered_special = common_mouse_list_filtered

list_of_filepaths = ["GSM2699154/GSM2699154_E14_B1.csv", "GSM2699155/GSM2699155_E14_B2.csv",
                     "GSM2699156/GSM2699156_E12_B2.csv", "GSM2699157/GSM2699157_E17_B2.csv",
                     "GSM3140915/GSM3140915_E12_v2.csv", "GSM3140916/GSM3140916_E14_v2.csv",
                     "GSM3140917/GSM3140917_E17_1_v2.csv", "GSM3140918/GSM3140918_E17_2_v2.csv",
                     "GSM3140919/GSM3140919_Fev_Cre_noreporter.csv", "GSM3140920/GSM3140920_Fev_Cre_reporter.csv",
                     "GSM3195456/GSM3195456_10XGenomics_Pancreas_E14-5_1_matrix.csv",
                     "GSM3488509/GSM3488509_10XGenomics_Pancreas_E14-5_2_matrix.csv",
                     "GSM3852752/matrix.csv",
                     "GSM3852753/matrix.csv", "GSM3852754/matrix.csv", "GSM3852755/matrix.csv"]

common_mouse_list_filtered_special.head()
```

	ENSMUSG_ID	mm10
0	ENSMUSG00000101435	mm10_Gm28772
1	ENSMUSG00000044244	mm10_I120rb
2	ENSMUSG00000069094	mm10_Pde7a
3	ENSMUSG00000105704	mm10_Gm43055
4	ENSMUSG00000033871	mm10_Ppargc1b

```

: def update_cols_intersect(num_of_file):

:     if num_of_file == 0:
:         cols_intersect = pd.DataFrame([np.sum(aggregate_cols.iloc[x] >0)
:                                         for x in range(0, aggregate_cols.shape[0])])..
:     ↪transpose()
:         cols_intersect.columns = gene_cols

:     else:
:         cols_intersect = pd.read_csv("presek_kolona.csv")
:         cols_intersect = cols_intersect.drop("Unnamed: 0", axis =1)
:         add_cols_intersect = pd.DataFrame([np.sum(aggregate_cols.iloc[i, :] >0)
:                                              for i in range(0, aggregate_cols.shape[0])]).transpose()
:         add_cols_intersect.columns = gene_cols
:         x = cols_intersect.columns.intersection(add_cols_intersect.columns)
:         for cols in set(cols_intersect.columns):
:             if cols not in x:
:                 cols_intersect = cols_intersect.drop(cols, axis = 1)
:         for cols in set(add_cols_intersect.columns):
:             if cols not in x:
:                 add_cols_intersect = add_cols_intersect.drop(cols, axis = 1)

:         cols_intersect = cols_intersect.add(add_cols_intersect)

:         cols_intersect.to_csv("presek_kolona.csv")

: def prep(filepath, num_of_file):
:     sample_data = pd.read_csv(filepath)
:     sample_id = filepath[3:10] + '_'

:     sample_data_filtered = common_mouse_list_filtered.set_index("ENSMUSG_ID").
:     ↪join(sample_data.set_index("Index"),
:           ↪how = "inner")

:     gene_cols = sample_data_filtered.index.tolist()
:     aggregate_cols = sample_data_filtered.drop("mm10", axis =1).reset_index().
:     ↪drop("index", axis =1)

:     update_cols_intersect(num_of_file)

:     sample_data_filtered.drop("mm10", axis =1, inplace =True)
:     sample_data_filtered_transposed = sample_data_filtered.transpose()
:     num_rows = sample_data_filtered_transposed.shape[0]
:     sample_data_filtered_transposed = sample_data_filtered_transposed.
:     ↪set_index(pd.Series(

```

```

        i += 1
        continue
    columns.append('M' + column[-5:] + common_mouse_list_new.loc[column, u
→'long'])

sample.columns = columns

sample = sample.reset_index()
sample = sample.set_index('id')
sample = sample.drop('index', axis = 1)

sample.to_csv(prep_file[:28] + '3.csv')

```

```
[ ]: for filepath in list_of_filepaths:
    rename(filepath)
```

```

        [sample_id + str(x) for x in range(1, num_rows+1)])
    sample_data_filtered_transposed.to_csv(filepath[:11] + sample_id + u
→"prep_file.csv")

    return sample_data_filtered.shape[1]

```

```
[ ]: sum_num_of_rows = 0
i = 0
for filepath in list_of_filepaths:
    sum_num_of_rows += prep(filepath, i)
    i += 1
```

```
[48]: def filter_cols(filepath):

    prep_file = filepath[:11] + filepath[3:10] + '_prep_file.csv'
    sample_data = pd.read_csv(prep_file)
    sample_data = sample_data.set_index("Unnamed: 0")
    del sample_data.index.name

    for cols in sample_data.columns:
        if cols not in cols_intersect:
            sample_data = sample_data.drop(cols, axis = 1)

    for i in sample_data.index:
        if np.sum(sample_data.loc[i, :]) < 1000 or np.sum(sample_data.loc[i, :u
→> 0]) < 500:
            sample_data = sample_data.drop(i, axis =0)

    sample_data.to_csv(prep_file[:28] + '2.csv')
```

```
[ ]: percentage = sum_num_of_rows / 100
cols_intersect = pd.read_csv("presek_kolona.csv")
cols_intersect = zip(cols_intersect.columns, cols_intersect.iloc[0])
cols_intersect = set(map(lambda x : x[0], filter(lambda x : x[1] > percentage,u
→cols_intersect)))
```

```
[ ]: for filepath in list_of_filepaths:
    filter_cols(filepath)
```

```
[81]: def rename(filepath):
    prep_file = filepath[:11] + filepath[3:10] + '_prep_file2.csv'
    sample = pd.read_csv(prep_file)
    columns = ['id']
    i = 0
    for column in sample.columns:
        if i == 0:
```

# TSNE

August 28, 2020

```
[1]: import pandas as pd
import numpy as np
from sklearn.manifold import TSNE
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt

[2]: list_of_filepaths = ["GSM2699154/2699154_prep_file3.csv", "GSM2699155/
→2699155_prep_file3.csv",
"GSM2699156/2699156_prep_file3.csv", "GSM2699157/
→2699157_prep_file3.csv",
"GSM3140915/3140915_prep_file3.csv", "GSM3140916/
→3140916_prep_file3.csv",
"GSM3140917/3140917_prep_file3.csv", "GSM3140918/
→3140918_prep_file3.csv",
"GSM3140919/3140919_prep_file3.csv", "GSM3140920/
→3140920_prep_file3.csv",
"GSM3195456/3195456_prep_file3.csv", "GSM3488509/
→3488509_prep_file3.csv",
"GSM3852752/3852752_prep_file3.csv", "GSM3852753/
→3852753_prep_file3.csv",
"GSM3852754/3852754_prep_file3.csv", "GSM3852755/
→3852755_prep_file3.csv"]
colors = ["red", "blue", "green", "yellow", "orange", "purple", "black", □
→"magenta", "brown", "cyan"]

[16]: for l in list_of_filepaths:
    data = pd.read_csv(l, index_col = 0)
    data.index.name = None

    pca = PCA()
    data_pca = pd.DataFrame(pca.fit_transform(data))
    for i in range(1, len(data.columns), 50):
        if pca.explained_variance_ratio_[:i].sum() >= 0.95:
            x = i
            break

    tsne_data = TSNE(n_iter = 2000).fit_transform(data_pca[data_pca.columns[:x]])
```

```

tsne_data = pd.DataFrame(tsne_data)
tsne_data.to_csv(l[19] + "tsne_file.csv")

[3]: group_files = [list_of_filepaths[:4], list_of_filepaths[4:10], ▾
                   ↵list_of_filepaths[11:19], list_of_filepaths[20:], ▾
                   ↵list_of_filepaths[10], list_of_filepaths[19]]]

[ ]: num_group = 1
for group in group_files:

    border_index = []
    border_index.append(0)
    data = pd.DataFrame()
    for file in group:
        df = pd.read_csv(file, index_col = 0)
        df.index.name = None
        data = data.append(df)
        border_index.append(data.shape[0])
    data.to_csv("grupe/" + str(num_group) + ".grupa/prep_group.csv")

    pca = PCA()
    data_pca = pd.DataFrame(pca.fit_transform(data))
    for i in range(1, len(data.columns), 50):
        if pca.explained_variance_ratio_[:i].sum() >= 0.95:
            x = i
            break

    tsne_data = TSNE(n_iter = 2000).fit_transform(data_pca[data_pca.columns[:x]])
    tsne_data = pd.DataFrame(tsne_data)
    tsne_data.to_csv("grupe/" + str(num_group) + ".grupa/" + "tnse_file.csv")

    fig = plt.figure(figsize= (20,15))
    for j in range(1, len(border_index)):
        plt.scatter(tsne_data[tsne_data.columns[0]].iloc[border_index[j-1]:border_index[j]], ▾
                    ↵tsne_data[tsne_data.columns[1]].iloc[border_index[j-1]:border_index[j]], c=colors[j-1], ▾
                    ↵label=group[j-1][11:18])
    plt.title('Group ' +str(num_group))
    plt.legend()
    fig.savefig("grupe/" + str(num_group) + ".grupa/plot")
    num_group += 1

```

# Clustering

August 28, 2020

```
[1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.cluster import SpectralClustering
from sklearn.metrics import silhouette_score
from sklearn.metrics import calinski_harabasz_score
from sklearn.metrics import davies_bouldin_score

[2]: list_of_filepaths = ["GSM2699154/2699154_prep_file3.csv", "GSM2699155/
→2699155_prep_file3.csv",
    "GSM2699156/2699156_prep_file3.csv", "GSM2699157/
→2699157_prep_file3.csv",
    "GSM3140915/3140915_prep_file3.csv", "GSM3140916/
→3140916_prep_file3.csv",
    "GSM3140917/3140917_prep_file3.csv", "GSM3140918/
→3140918_prep_file3.csv",
    "GSM3140919/3140919_prep_file3.csv", "GSM3140920/
→3140920_prep_file3.csv",
    "GSM3195456/3195456_prep_file3.csv", "GSM3488509/
→3488509_prep_file3.csv",
    "GSM3852752/3852752_prep_file3.csv", "GSM3852753/
→3852753_prep_file3.csv",
    "GSM3852754/3852754_prep_file3.csv", "GSM3852755/
→3852755_prep_file3.csv"]

colors = ["red", "blue", "green", "yellow", "orange", "black", "brown"]
```

## 1 KMeans

```
[3]: def km(filename, num_of_clus):

    labels.index = data.index
    tsne_data = pd.read_csv(filename[:19] + "tsne_file.csv", index_col = 0)
```

```

model = KMeans(num_of_clus)
model.fit(data)

tsne_data["labels"] = model.labels_
data["labels"] = model.labels_
fig = plt.figure(figsize= (20,15))
for j in range(0, num_of_clus):
    plt.scatter(tsne_data[tsne_data.columns[0]][tsne_data['labels'] == j], tsne_data[tsne_data.columns[1]][tsne_data['labels'] == j], c=colors[j])
plt.title('KMeans: ' +str(num_of_clus) +' Silhouette Score:' + str(silhouette_score(data, tsne_data['labels']))+
', CH Score: ' + str(calinski_harabasz_score(data, tsne_data['labels']))+
', DB Score: ' + str(davies_bouldin_score(data, tsne_data['labels'])))
fig.savefig(filename[:11]+ "KM/" + filename[11:19] + "KM_" + str(num_of_clus))
labels[str(num_of_clus)] = model.labels_

for i in data.columns:
    if np.sum(data.loc[:, i] > 0) < 1:
        data.drop(i, axis =1, inplace =True)
desc = pd.DataFrame()
for j in range(0, num_of_clus):
    df = data.loc[data["labels"] == j].describe().rename(index = lambda x: str(j) + ".cluster_" + x)
    desc = desc.append(df)
desc.to_csv(filename[:11]+ "KM/KM_" + str(num_of_clus) + ".csv")

```

```

[ ]: for l in list_of_filepaths:
    data = pd.read_csv(l, index_col = 0)
    labels = pd.DataFrame(columns = ["2", "3", "4", "5", "6", "7"], index = data.index)
    for i in range(2, 8):
        km(l, i)
    labels.to_csv(l[:11]+ "KM/" + l[11:19]+ "labels_KM.csv")

```

## 1.1 Spectral clustering

```

[5]: def spec(filename, num_of_clus):

    model = SpectralClustering(num_of_clus, affinity='nearest_neighbors')
    tsne_data = pd.read_csv(filename[:19] + "tsne_file.csv", index_col = 0)

```

```

model.fit(data.values)
tsne_data["labels"] = model.labels_
data["labels"] = model.labels_

fig = plt.figure(figsize= (20,15))
for j in range(0, num_of_clus):
    plt.scatter(tsne_data[tsne_data.columns[0]][tsne_data['labels'] == j],
                tsne_data[tsne_data.columns[1]][tsne_data['labels'] == j], c=c)
plt.title('Spectral clus: ' +str(num_of_clus) +' Silhouette Score:' +str(silhouette_score(data, tsne_data['labels']))+
          ', CH Score: ' + str(calinski_harabasz_score(data, tsne_data['labels']))+
          ', DB Score: ' + str(davies_bouldin_score(data, tsne_data['labels'])))
fig.savefig(filename[:11]+ "Spec/" + filename[11:19] + "Spec_" +str(num_of_clus))
labels[str(num_of_clus)] = model.labels_

for i in data.columns:
    if np.sum(data.loc[:, i] > 0) < 1:
        data.drop(i, axis =1, inplace =True)
desc = pd.DataFrame()
for j in range(0, num_of_clus):
    df = data.loc[data["labels"] == j].describe().rename(index = lambda x:str(j) + ".cluster_" + x)
    desc = desc.append(df)
desc.to_csv(filename[:11]+ "Spec/Spec_" + str(num_of_clus) + ".csv")

```

```

[ ]: for l in list_of_filepaths:
    data = pd.read_csv(l, index_col = 0)
    labels = pd.DataFrame(columns = ["2", "3", "4", "5", "6", "7"], index = data.index)
    for i in range(2, 8):
        spec(l, i)
    labels.to_csv(l[:11]+ "Spec/" + l[11:19]+ "labels_spec.csv")

```

## 1 Klasterovanje po grupama

```
[ ]: import numpy as np
import pandas as pd

group_files = [[ "GSM2699154/2699154_prep_file3.csv", "GSM2699155/
→2699155_prep_file3.csv",
    "GSM3140916/3140916_prep_file3.csv"],
    ["GSM2699157/2699157_prep_file3.csv", "GSM3140917/
→3140917_prep_file3.csv",
        "GSM3140918/3140918_prep_file3.csv"],
    ["GSM3195456/3195456_prep_file3.csv", "GSM3488509/
→3488509_prep_file3.csv"],
    ["GSM3852752/3852752_prep_file3.csv", "GSM3852753/
→3852753_prep_file3.csv",
        "GSM3852754/3852754_prep_file3.csv", "GSM3852755/
→3852755_prep_file3.csv"],
    ["GSM3140915/3140915_prep_file3.csv", "GSM3140916/
→3140916_prep_file3.csv",
        "GSM3140917/3140917_prep_file3.csv", "GSM3140918/
→3140918_prep_file3.csv"]]

group_filepaths = ["grupe/14.dan_1/", "grupe/17.dan/", "grupe/14.dan_2/",
    "grupe/mesavina1/", "grupe/mesavina2/"]
```

### 1.1 Pravljenje fajlova

```
[ ]: for i in range(0, len(group_files)):
    df = pd.DataFrame()
    for f in group_files[i]:
        df = df.append(pd.read_csv(f, index_col = 0))
    df.to_csv(group_filepaths[i]+"group.csv")
```

## 1.2 Pravljenje tsne

```
[ ]: from sklearn.manifold import TSNE
      from sklearn.decomposition import PCA

      for i in range(0, len(group_filepaths)):
          df = pd.read_csv(group_filepaths[i]+"group.csv", index_col = 0)
          pca = PCA()
          data_pca = pd.DataFrame(pca.fit_transform(df))
          for l in range(1, len(df.columns), 50):
              if pca.explained_variance_ratio_[:l].sum() >= 0.95:
                  x = l
                  break
          tsne_data = TSNE(n_iter = 2000).fit_transform(data_pca[data_pca.columns[:x]])
          tsne_data = pd.DataFrame(tsne_data)
          tsne_data.set_index(df.index, inplace = True)
          tsne_data.to_csv(group_filepaths[i] + "tsne_file.csv")
```

## 1.3 Klasterovanje

```
[ ]: import matplotlib.pyplot as plt
      from sklearn.cluster import SpectralClustering
      from sklearn.cluster import KMeans
      from sklearn.metrics import silhouette_score, calinski_harabasz_score,
      ↪ davies_bouldin_score

      """ocene_ss = pd.DataFrame(index = list(map(lambda x : x[6:-1], ↪
      ↪ group_filepaths)),
      ↪ columns = ["2", "3", "4", "5", "6", "7"])
      ocene_db = pd.DataFrame(index = list(map(lambda x : x[6:-1], group_filepaths)),
      ↪ columns = ["2", "3", "4", "5", "6", "7"])
      ocene_ch = pd.DataFrame(index = list(map(lambda x : x[6:-1], group_filepaths)),
      ↪ columns = ["2", "3", "4", "5", "6", "7"])"""

      ocene_ss = pd.read_csv("Ocene/km_grupe_ss.csv", index_col = 0).append(
          pd.DataFrame(index = list(map(lambda x : x[6:-1], group_filepaths)), columns =
      ↪ = ["2", "3", "4", "5", "6", "7"]))
      ocene_db = pd.read_csv("Ocene/km_grupe_db.csv", index_col = 0).append(
          pd.DataFrame(index = list(map(lambda x : x[6:-1], group_filepaths)), columns =
      ↪ = ["2", "3", "4", "5", "6", "7"]))
      ocene_ch = pd.read_csv("Ocene/km_grupe_ch.csv", index_col = 0).append(
          pd.DataFrame(index = list(map(lambda x : x[6:-1], group_filepaths)), columns =
      ↪ = ["2", "3", "4", "5", "6", "7"]))
```

```

#ocene_ss = pd.read_csv("Ocene/spec_grupe_ss.csv", index_col = 0).append(
#    pd.DataFrame(index = list(map(lambda x : x[6:-1], ↵
#        ↵group_filepaths)),columns = ["2", "3", "4", "5", "6", "7"]))
#ocene_db = pd.read_csv("Ocene/spec_grupe_db.csv", index_col = 0).append(
#    pd.DataFrame(index = list(map(lambda x : x[6:-1], ↵
#        ↵group_filepaths)),columns = ["2", "3", "4", "5", "6", "7"]))
#ocene_ch = pd.read_csv("Ocene/spec_grupe_ch.csv", index_col = 0).append(
#    pd.DataFrame(index = list(map(lambda x : x[6:-1], ↵
#        ↵group_filepaths)),columns = ["2", "3", "4", "5", "6", "7"]))

colors = ["red", "blue", "green", "yellow", "orange", "black", "purple"]

for i in range(0, len(group_files)):

    df = pd.read_csv(group_filepaths[i]+"group.csv", index_col = 0)
    tsne_data = pd.read_csv(group_filepaths[i] + "tsne_file.csv", index_col = 0)

    labels = pd.DataFrame(index = df.index, columns = ["2", "3", "4", "5", "6", ↵
    ↵"7"])
    for num_of_clus in range(2, 8):

        model = KMeans(num_of_clus)
        model.fit(df)
        #model = SpectralClustering(num_of_clus, affinity='nearest_neighbors')
        #model.fit(df.values)

        fig = plt.figure(figsize= (20,15))
        for j in range(0, num_of_clus):
            plt.scatter(tsne_data.loc[model.labels_ == j][tsne_data.columns[0]], ↵
                        tsne_data.loc[model.labels_ == j][tsne_data.columns[1]], ↵
                        ↵c=colors[j])

        ocene_ss.at[group_filepaths[i][6:-1], str(num_of_clus)] = ↵
        ↵silhouette_score(df, model.labels_)
        ocene_db.at[group_filepaths[i][6:-1], str(num_of_clus)] = ↵
        ↵davies_bouldin_score(df, model.labels_)
        ocene_ch.at[group_filepaths[i][6:-1], str(num_of_clus)] = ↵
        ↵calinski_harabasz_score(df, model.labels_)

        plt.title('KM: ' +str(num_of_clus) +
                  ' Silhouette Score: ' +str(ocene_ss.at[group_filepaths[i][6:-
                  1], str(num_of_clus)])+
                  ', CH Score: ' + str(ocene_ch.at[group_filepaths[i][6:-1], ↵
                  ↵str(num_of_clus)])+
                  ', DB Score: ' + str(ocene_db.at[group_filepaths[i][6:-1], ↵
                  ↵str(num_of_clus)]))

        #plt.title('Spektralno: ' +str(num_of_clus) +
        #          '# Silhouette Score: ' +str(ocene_ss.at[group_filepaths[i][6:-
        #          1], str(num_of_clus)])+
        #          '# , CH Score: ' + str(ocene_ch.at[group_filepaths[i][6:-1], ↵
        #          ↵str(num_of_clus)])+
        #          '# , DB Score: ' + str(ocene_db.at[group_filepaths[i][6:-1], ↵
        #          ↵str(num_of_clus)]))

        plt.show()

fig.savefig(group_filepaths[i]+ "KM/KM_" + str(num_of_clus))
#fig.savefig(group_filepaths[i]+ "Spec/Spec_" + str(num_of_clus))

labels[str(num_of_clus)] = model.labels_

labels.to_csv(group_filepaths[i]+ "KM/labels.csv")
#labels.to_csv(group_filepaths[i]+ "Spec/labels.csv")

ocene_ss.to_csv("Ocene/km_grupe_ss.csv")
ocene_db.to_csv("Ocene/km_grupe_db.csv")
ocene_ch.to_csv("Ocene/km_grupe_ch.csv")

#ocene_ss.to_csv("Ocene/spec_grupe_ss.csv")
#ocene_db.to_csv("Ocene/spec_grupe_db.csv")
#ocene_ch.to_csv("Ocene/spec_grupe_ch.csv")

```