Gašper Kočjaž (63120162) **Naloga 1** 25. oktober 2019

Uvod. V prvi domači nalogi sem implementiral algoritem za hierarhično razvrščanje v skupine in analiziral glasovanje držav na tekmovanju za Pesem Evrovizije. Za računanje razdalj med skupinami sem uporabil Evklidsko razdaljo in rezultate prikazal z tekstovnim in grafičnim dendrogramom. Cilj naloge je določiti skupine držav, ki glasujejo podobno in ugotoviti katerim državam namenjajo nadpovprečne/podpovprečne ocene. Prav tako sem rešil dodatno nalogo, kjer primerjam pristranskost ocenjevanja žirije in televotinga.

Računanje razdalj. V domači nalogi sem uporabil postopek hierarhičnega razvrščanja v skupine, ki uporablja evklidsko razdaljo 1 kot mero različnosti oz. razdalje. Razdaljo med dvema posameznima profiloma glasovanja (posamezne države) sem izračunal z evklidsko razdaljo, pri čemer sem moral upoštevati manjkajoče podatke. Manjkajočih glasov ne smemo interpretirati kot oceno 0 točk, zato pare istoležnih glasov z manjkajočimi podatki pri računanju razdalje preskočimo. Kjer sta bila oba istoležna glasova znotraj profilov definirana, sem ju upošteval za izračun evklidske razdalje - v nasprotnem primeru, če eden izmed glasov ali oba glasova nista definirana ju ne upoštevam pri izračunu razdalje. Na koncu sem opravil še normalizacijo razdalje, tako da sem izvedel deljenje z številom definiranih/upoštevanih parov glasov (zmanjšamo vpliv velikega števila glasov čez leta) in nato še množenje z številom vseh parov glasov oz. originalno dolžino profila glasovanja (potrebno zaradi predhodnjega deljenja) 2.

Razdaljo med skupinami sem določil kot povprečno razdaljo med profili posameznih parov držav dveh skupin oz. angl. *average linkage*. Zaradi velikega števila manjkajočih podatkov je bilo možno, da je razdalja med posameznima državama enaka 0 (profila nimata nobenih definiranih istoležnih ocen za izračun razdalje) in zato izračunana povprečna razdalja med skupinami nižja kot dejanska - zato sem metodo računanja razdalj dopolnil tako, da je ignorirala take pare držav pri računanju povprečne razdalje med skupinama.

Podatki. Podatke o glasovanju v finalnem delu tekmovanja za Pesem Evrovizije smo dobili na portalu *data.world* (`https://data.world/datagraver/eurovision-song-contest-scores-1975-2019/`). Datoteka v *.csv* formatu vsebuje podatke o tekmovanjih iz obdobja 1975-2019, tako za glasovanje žirije in ocene iz televotinga (ki je na voljo za obdobje 2016-2019). Vse skupaj je glasovalo 51 različnih držav, 50 pa jih je tudi nastopilo na tekmovanju s svojo pesmijo. Eden izmed korakov predobdelave podatkov je bila tudi standardizacija imen držav, npr. v primeru Makedonije (*North Macedonia* & *F.Y.R. Macedonia*) in Srbije (*Serbia & Montenegro* & *Serbia*). Posamezna vrstica *.csv* datoteke vsebuje podatke o letu tekmovanja, tipu glasovanja (žirija/televoting),

$$d\left(a,b\right)=\sqrt{\sum_{i=1}^{n}\left(a_i-b_i\right)^2} \tag{1}$$

Slika 1: Osnovna formula za evklidsko razdaljo.

$$d(a,b) = \sqrt{(N/n) * \sum_{i=1}^{n} (a_i - b_i)^2} \tag{2}$$

Slika 2: Prilagojena formula za evklidsko razdaljo z normalizacijo razdalje: N = dolžina celotnega profila glasovanja, n = število definiranih parov glasov.

ime države, ki dodeljuje oceno, ime države, ki prejme oceno in sama ocena. Za uporabo v algoritmu hierarhičnega razvrščanja sem podatke preoblikoval tako, da sem v slovar *self.data* shranil celoten profil glasovanja posamezne države v obliki vektorja oz. seznama ocen: *država ↦ [ocena_za_državo1_letoX, ocena_za_državo2_letoX, ..., ocena_za_državo1_letoX+1, ...]*.

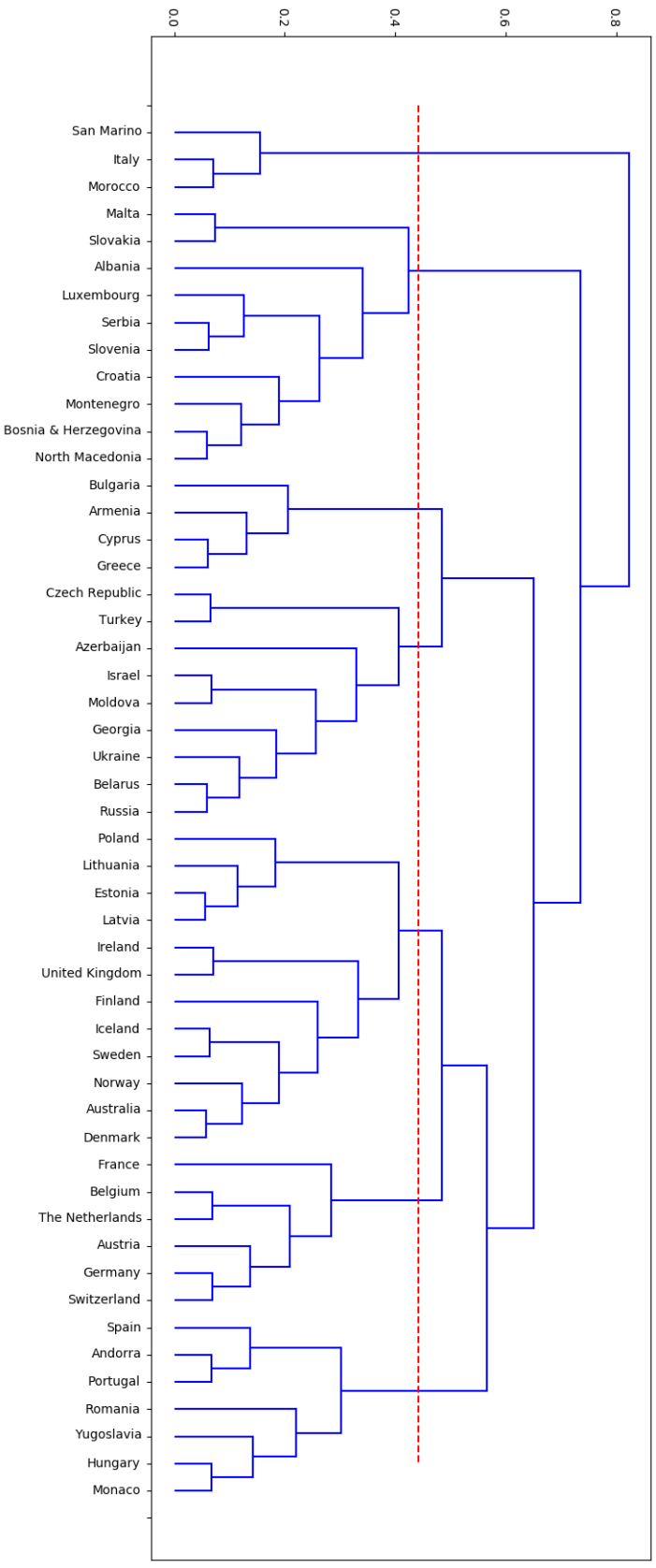
Profil glasovanja posamezne države sem zgradil tako, da sem prvič prebral vhodno datoteko zato, da sem ugotovil katere države sodelujejo na tekmovanju in glasovanju, ter koliko let poteka glasovanje - s tem sem določil velikost in obliko seznama, ki ga hranim za posamezne države v slovarju *self.data*. Ob drugem branju vhodne datoteke pa sem bral ocene in jih shranjeval na ustrezna mesta znotraj profila države. Pri oblikovanju ustrezne podatkovne strukture sem upošteval številne manjkajoče podatke (npr. država se nekega leta ni udeležila tekmovanja/glasovanja), ki sem jih v profilu ocen označil za nedefinirane z podatkovnim tipom *None*.

Podatki o poteku gručenja so shranjeni v gnezdenem seznamu držav oz. skupin *self.clusters*, ki ga uporabljam pri izgradnji tekstovnega dendrograma. Pri izgradnji slikovnega dendrograma, pa si pomagam s strukturo *self.cluster_history*, ki poleg držav vsebuje tudi razdaljo, pri kateri so bile združene in pa pozicijo sredine skupine na x osi grafa.

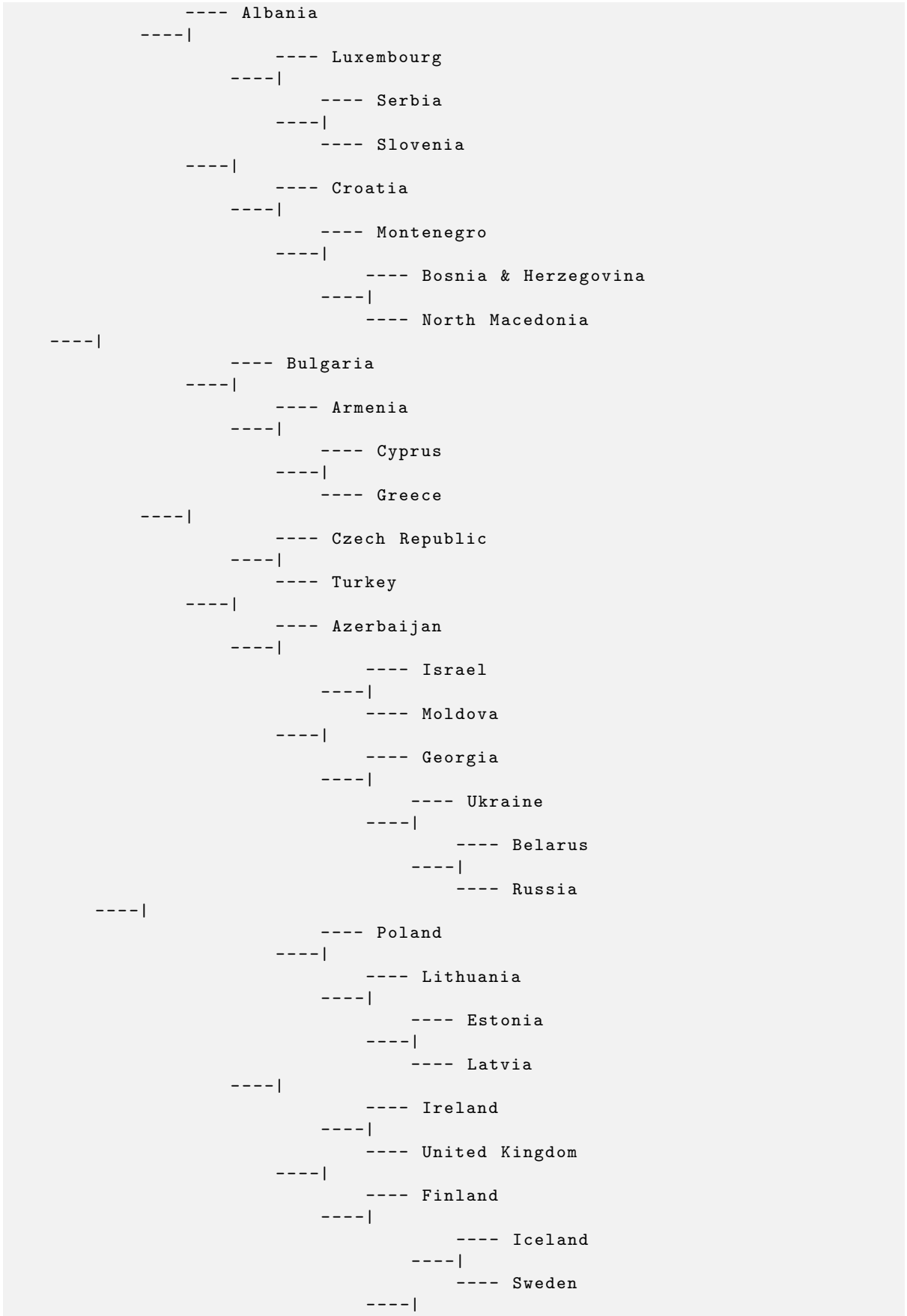
Če je naloga zasnovana tako, da vključuje analizo izbranih podatkov, v tem razdelku opišeš, kakšni so ti podatki in navedeš nekaj osnovnih statističnih lastnosti teh podatkov. Slednje vključujejo velikost podatkov (na primer število primerov, število in vrsto atributov), delež manjkajočih podatkov, opis in porazdelitev vrednosti ciljnih spremenljivk, in podobno. Če si podatke pridobil sam, tu opišeš, na kakšen način, kje in kako.

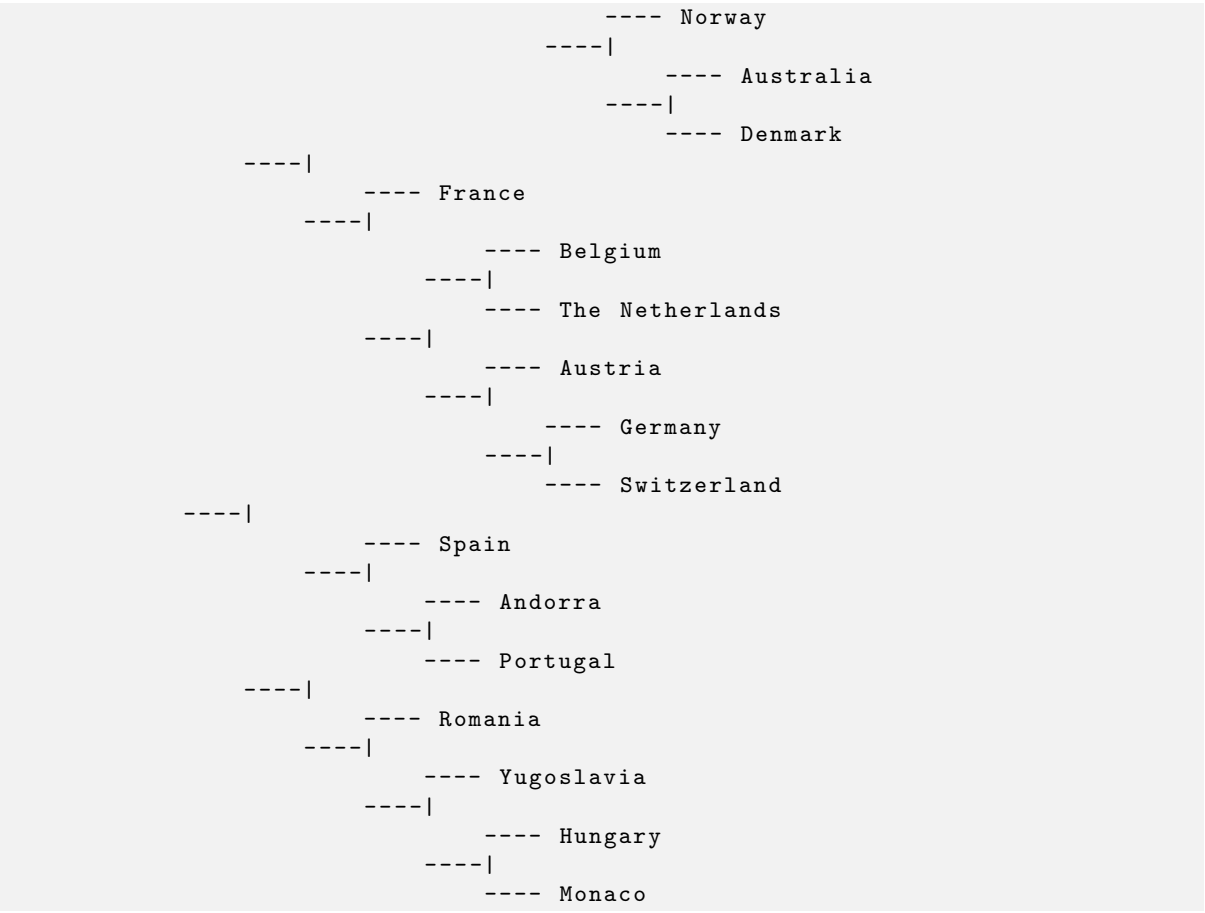
Dendrogram Rezultat hierarhičnega razvrščanja predstavimo z dendrogramom, ki ponazarja gručenje držav oz. skupin v obliki drevesa in vizualizira razdalje med njimi. Razvil sem metodo, ki implementira tekstovni dendrogram v metodi *out()*. Poleg tega sem v metodi *plot_tree()* izrisal tudi grafični dendrogram 3 s pomočjo orodja za risanje grafov, *pyplot* iz knjižnice *matplotlib*. Gradnja grafičnega dendrograma poteka s pomočjo branja informacij o razdalji in lokaciji clustra na x osi dendrograma iz slovarja *self.cluster_history*.

```
        ---- San Marino
    ----|
            ---- Italy
        ----|
            ---- Morocco
----|
                ---- Malta
            ----|
                ---- Slovakia
        ----|
```



Slika 3: Grafični dendrogram.


Albania
Luxembourg
Serbia
Slovenia
Croatia
Montenegro
Bosnia & Herzegovina
North Macedonia
Bulgaria
Armenia
Cyprus
Greece
Czech Republic
Turkey
Azerbaijan
Israel
Moldova
Georgia
Ukraine
Belarus
Russia
Poland
Lithuania
Estonia
Latvia
Ireland
United Kingdom
Finland
Iceland
Sweden



Skupine in njihove preferenčne izbire. Skupine (prikazane v tabeli 1) sem določil z uporabo algoritma za hierarhično razvrščanje, ki vrne skupine v obliki gnezdenega seznama. Na podlagi grafične vizualizacije sem določil rez, ki loči države v ločene skupine. Pri tem sem poskušal rez postaviti tako, da je bila razlika v razdalji med potencionalnimi grupiranji čim večja in, da je bil rezultat grupiranja čim bolj smiselen glede na medsebojni odnos držav znotraj skupine (upoštevanje sosednosti oz. geografije, kulture, jezikov, itd.). Prav tako sem pri oblikovanju skupin upošteval njihovo velikost in skupno število različnih skupin, saj želimo enakomerno grupiranje brez osamelcev.

Preferenčne izbire skupin sem določil z uporabo funkcije *self.compare_cluster_to_others*, ki predstavi določeno skupino z svojim vektorjem glasovanja tako, da za posamezno oceno v profilu glasovanja določi povprečje istoležnih ocen (ki so definirane) vseh držav iz skupine. Tako dobljeni vektor glasovanja primerjam z vektorjem vseh preostalih držav, ki niso del te skupine in shranim skupno razliko v ocenah posameznih držav. Če je skupina glasovala dobro za neko državo je skupna razlika pozitivna in visoka, v nasprotnem primeru je razlika negativna (in je njena absolutna vrednost visoka). Če država od skupine ne prejema ocen, ki bi statistično odstopale od ocen preostalih skupin, je razlika blizu 0.

Spodnji tabeli vključujeta nadpovprečno ocenjene države 2 in podpovprečno ocenjene države 3 za vsako skupino.

Tabela 1: Razporeditev v skupine, ki jih vrne algoritem hierarhičnega razvrščanja.

# skupine	države v skupini
1	San Marino, Italy, Morocco
2	Malta, Slovakia, Albania, Luxembourg, Serbia, Slovenia, Croatia, - Montenegro, Bosnia & Herzegovina, North Macedonia
3	Bulgaria, Armenia, Cyprus, Greece
4	Czech Republic, Turkey, Azerbaijan, Israel, Moldova, Georgia, - Ukraine, Belarus, Russia
5	Poland, Lithuania, Estonia, Latvia, Ireland, United Kingdom, - Finland, Iceland, Sweden, Norway, Australia, Denmark
6	France, Belgium, The Netherlands, Austria, Germany, Switzerland
7	Spain, Andorra, Portugal, Romania, Yugoslavia, Hungary, Monaco

Tabela 2: Najbolje ocenjene države po skupinah.

# skupine	najbolje ocenjene
1	Monaco, Morocco, Albania, Moldova, San Marino
2	Croatia, Bosnia & Herzegovina, Serbia, North Macedonia, Malta
3	Greece, Cyprus, Spain, Russia, Armenia
4	Ukraine, Azerbaijan, Russia, Yugoslavia, Armenia
5	Sweden, Denmark, Norway, Estonia, Iceland
6	Turkey, Israel, The Netherlands, Portugal, Germany
7	Italy, Germany, Romania, Portugal, Spain

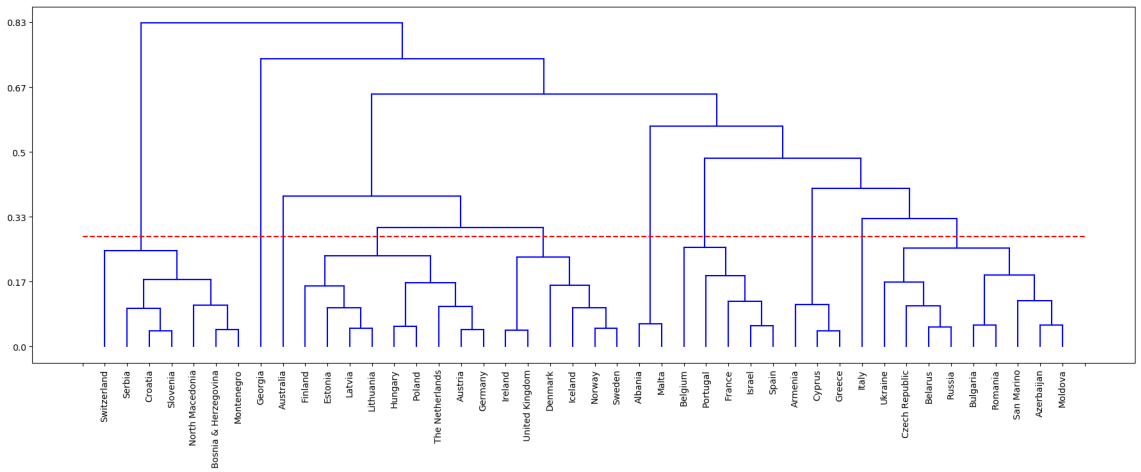
Tabela 3: Najslabše ocenjene države po skupinah.

# skupine	najslabše ocenjene
1	Sweden, Russia, Denmark, Israel, Germany
2	Denmark, Sweden, Luxembourg, Austria, Norway
3	Sweden, Germany, United Kingdom, Russia, Denmark
4	Sweden, Israel, Turkey, Denmark, Cyprus
5	Greece, Spain, Turkey, Italy, Croatia
6	Russia, Ukraine, Cyprus, Azerbaijan, Malta
7	Norway, Sweden, Bosnia & Herzegovina, Ireland, Yugoslavia

Primerjava žirije in televotinga. Ocenjevanje žirije in televotinga sem primerjal z hierarhičnim razvrščanjem v skupine in analizo podobnosti držav znotraj skupine ter njihovih preferenc pri glasovanju. Ker so podatki o televotingu na voljo samo od leta 2016 naprej, sem za žirijo in televoting zgradil dva ločena dendrograma (televoting 4 žirija 5) nad podatki izključno iz obdobja 2016-2019 - za žirijo ignoriram podatke pred letom 2016, ker sta tako žirija in televoting bolj primerjliva saj primerjamo le podatke v kontekstu istem časovnem obdobju. Po analizi dendrogramov je vidno, da pri televotingu gručenje večinoma poteka med državami, ki so v neposredni geografski bližini in so skupine dosti bolj čiste - redke so države, ki izstopajo/ne ustrezajo v svoji skupini. V nasprotnem primeru pa se pri žiriji v skupinah pogosto pojavljajo manj geografsko in kulturno povezane države (npr. Slovenia v skupini pretežno baltskih držav, Spain v skupini severno evropjskih držav, itd.). Zato sklepam, da žirija ocenjuje bolj nepristransko.

Tudi analiza preferenčnih izbir pokaže za televoting, da skupine pogosto preferirajo države znotraj skupine ali drugače sorodne države. Ko primerjamo absolutno vrednost razlike oz. odstopanja v glasovanju skupine od povprečja vidimo, da so absolutne vrednosti razlik manjše pri žiriji, kar pomeni, da žirija ocenjuje bolj nepristransko (ni držav, ki bi bile posebej priljubljene-/nepriljubljene) medtem, ko so pri televotingu večje razlike zaradi vpliva ljudstva.

V prilogi sem dodal še izpis metode *self.compare_cluster_to_others* (priloga A), ki določi preferenčne države in stopnjo preference v obliki prej omenjene razlike od povprečja.

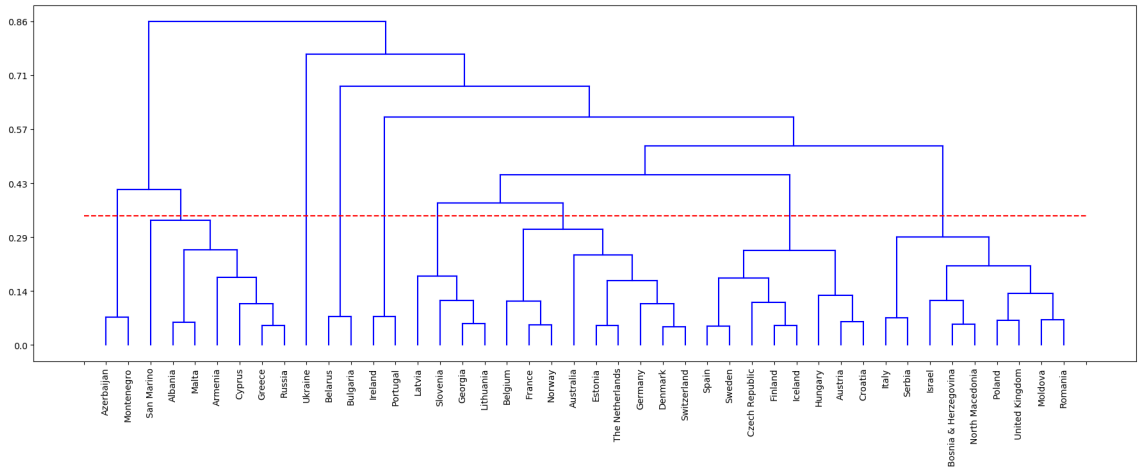


Slika 4: Grafični dendrogram - TELEVOTING.

Izjava o izdelavi domače naloge. Domačo nalogo in pripadajoče programe sem izdelal sam.

Priloge

Primerjava preferenčnih izbir ocenjevanja žirije in televotinga. Spodaj je priložen podroben izpis metode *self.compare_cluster_to_others*, ki za določeno skupino vrne 5 najbolje in



Slika 5: Grafični dendrogram - ŽIRIJA.

5 najslabše ocenjevanih držav.

```
————————[U+FFFD]IRIJA————————
comparing:
['Azerbaijan', 'Montenegro']
{'Serbia': 18.25114709851552, 'Russia': 15.442645074224021, 'Albania': 15.163157894736843, 'Malta': 12.080296896086 37,
{'Sweden': -17.705566801619433, 'The␣Netherlands': -13.271457489878543, 'Australia': -9.926113360323887, 'Austria': -9.
comparing:
['San␣Marino', 'Albania', 'Malta', 'Armenia', 'Cyprus', 'Greece', 'Russia']
{'Cyprus': 17.00214752567694, 'Italy': 13.99765724471607, 'Russia': 10.189915966386554, 'Greece': 9.8859477124183, 'Arm
{'The␣Netherlands': -10.728367710720654, 'Australia': -10.475095492742554, 'Austria': -10.258543417366948, 'Sweden': -6
comparing:
['Ukraine']
{'Belarus': 10.600000000000001, 'France': 8.137195121951219, 'Lithuania': 7.504878048780487, 'Israel': 7.27317073170731
{'Italy': -12.314024390243903, 'Bulgaria': -11.539024390243902, 'Sweden': -10.449390243902439, 'Australia': -9.07134146
comparing:
['Belarus', 'Bulgaria']
{'Malta': 12.871794871794872, 'Austria': 10.005769230769232, 'Lithuania': 8.740384615384615, 'Cyprus': 7.505128205128205
{'Sweden': -9.185897435897438, 'France': -6.891666666666667, 'Australia': -6.674358974358974, 'Italy': -5.44935897435897
comparing:
['Ireland', 'Portugal']
{'Belgium': 16.793589743589745, 'Bulgaria': 11.628205128205128, 'The␣Netherlands': 7.632422402159245, 'Azerbaijan': 6.31
{'Australia': -10.243252361673413, 'Ukraine': -5.857692307692308, 'Malta': -5.714473684210526, 'Russia': -4.98684210526
comparing:
['Latvia', 'Slovenia', 'Georgia', 'Lithuania']
{'Sweden': 6.535285285285285, 'Ukraine': 5.872688477951636, 'Latvia': 5.7368421052631575, 'Belgium': 5.202702702702703,
{'Italy': -6.5383673146831045, 'Australia': -5.486051841315, 'Malta': -4.4557057057057055, 'North␣Macedonia': -4.083333
comparing:
['Belgium', 'France', 'Norway', 'Australia', 'Estonia', 'The␣Netherlands', 'Germany', 'Denmark', 'Switzerland']
{'Sweden': 9.962772890192245, 'The␣Netherlands': 8.2474376114082, 'Germany': 7.67474376114082, 'Norway': 5.1996991978609
{'Azerbaijan': -6.462701612903226, 'Russia': -4.973902329749103, 'Albania': -4.922450309547084, 'Malta': -4.656474014330
comparing:
['Spain', 'Sweden', 'Czech␣Republic', 'Finland', 'Iceland', 'Hungary', 'Austria', 'Croatia']
{'Australia': 6.599988859180035, 'Sweden': 6.161522790934555, 'Czech␣Republic': 5.824598930481283, 'The␣Netherlands': 4.
{'Ukraine': -7.0811051693404625, 'Belgium': -4.261363636363637, 'Estonia': -3.224264705882353, 'Germany': -3.010918003586
comparing:
['Italy', 'Serbia', 'Israel', 'Bosnia␣&␣Herzegovina', 'North␣Macedonia', 'Poland', 'United␣Kingdom', 'Moldova', 'Austra
{'Ukraine': 7.494299613784908, 'Australia': 7.26804176215941, 'Estonia': 4.836397058823529, 'North␣Macedonia': 3.077922
{'Italy': -6.119989814107461, 'Cyprus': -5.709577391562686, 'The␣Netherlands': -5.416313874034461, 'Russia': -3.180555


——————— TELEVOTING ———————
comparing:
['Switzerland', 'Serbia', 'Croatia', 'Slovenia', 'North␣Macedonia', 'Bosnia␣&␣Herzegovina', 'Montenegro']
{'Serbia': 32.11196911196912, 'Croatia': 13.344444444444445, 'Albania': 9.5, 'Italy': 9.444957983193277, 'Slovenia': 8.8
{'Moldova': -6.638888888888889, 'Israel': -5.828338001867414, 'Australia': -5.767927170868347, 'Sweden': -5.641783380018
comparing:
['Finland', 'Estonia', 'Latvia', 'Lithuania', 'Hungary', 'Poland', 'The␣Netherlands', 'Austria', 'Germany']
{'Sweden': 5.326470348647767, 'Belgium': 5.2430555555555545, 'Estonia': 4.1875, 'Austria': 4.038992869875223, 'Denmark'
{'Bulgaria': -5.34375, 'France': -4.971957478005865, 'Serbia': -4.2404284783317046, 'Cyprus': -4.1316898826979465, 'Aze
```

```
comparing:
['Ireland', 'United_Kingdom', 'Denmark', 'Iceland', 'Norway', 'Sweden']
{'Sweden': 14.945645645645646, 'Lithuania': 13.242857142857144, 'Norway': 10.053603603603603, 'Australia': 9.15770308123
{'Italy': -14.028151260504202, 'Ukraine': -8.729365079365078, 'Russia': -8.504761904761903, 'Serbia': -5.95728291316526
comparing:
['Albania', 'Malta']
{'Italy': 25.589574898785425, 'Bulgaria': 10.841666666666667, 'Australia': 9.249561403508771, 'North_Macedonia': 4.26315
{'Moldova': -9.118589743589745, 'Ukraine': -6.616025641025641, 'Serbia': -5.347334682860998, 'Denmark': -5.197233468286
comparing:
['Belgium', 'Portugal', 'France', 'Israel', 'Spain']
{'France': 17.842105263157894, 'Spain': 12.431286549707602, 'The_Netherlands': 6.44942423626634, 'Italy': 5.92215787215
{'Hungary': -6.125825825825825, 'Serbia': -5.505019305019305, 'Lithuania': -5.054054054054054, 'Sweden': -5.03923208923
comparing:
['Armenia', 'Cyprus', 'Greece']
{'Cyprus': 31.64612010796221, 'Greece': 11.773954116059379, 'France': 8.7688113214429, 'Bulgaria': 7.348852901484482, '
{'Denmark': -6.409417514680672, 'Azerbaijan': -5.729018492176387, 'Serbia': -5.487799540431119, 'Poland': -4.7631578947
comparing:
['Ukraine', 'Czech_Republic', 'Belarus', 'Russia', 'Bulgaria', 'Romania', 'San_Marino', 'Azerbaijan', 'Moldova']
{'Azerbaijan': 12.836134453781513, 'Russia': 9.124649859943975, 'Ukraine': 9.067226890756302, 'Moldova': 9.042016806722
{'Italy': -5.567279942279942, 'Sweden': -5.46969696969697, 'Serbia': -4.823953823953824, 'Belgium': -3.8371212121212124
```