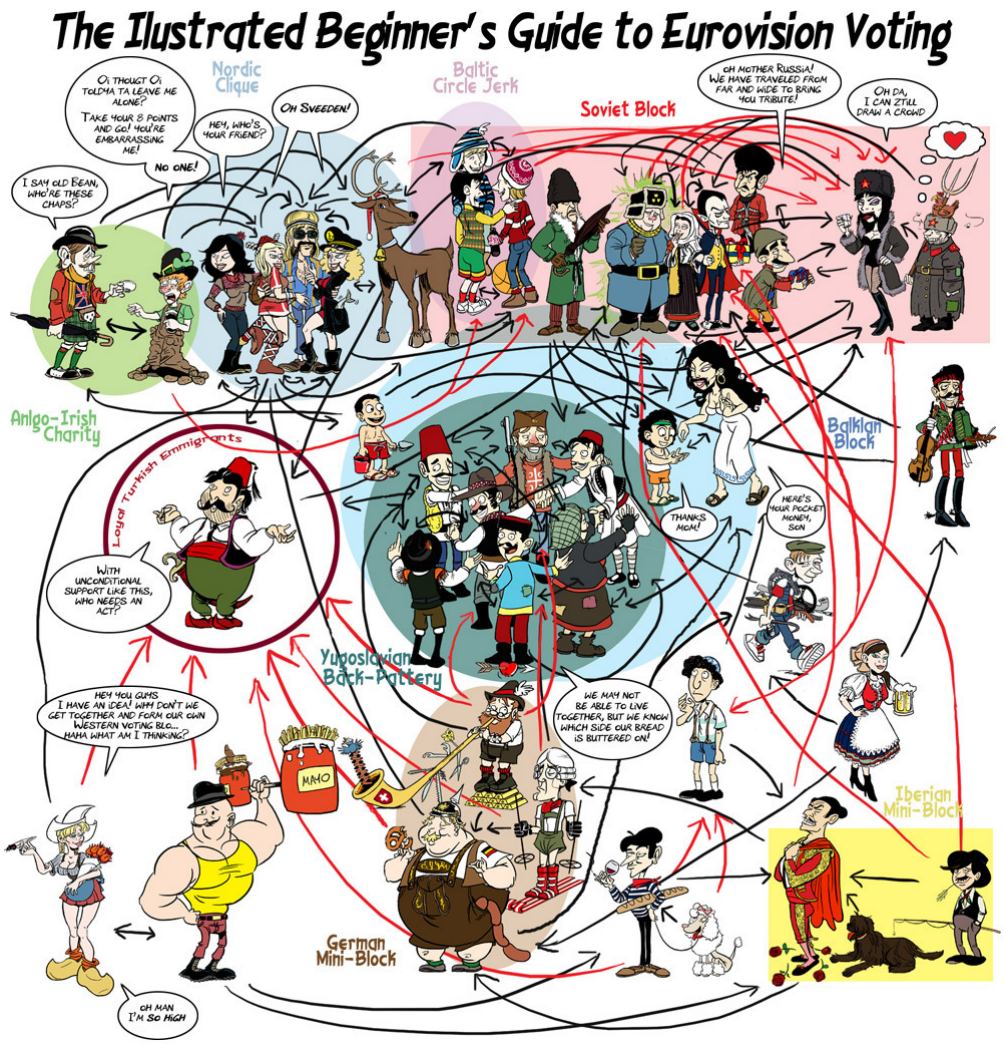
Uvod v odkrivanje znanj iz podatkov

1. domača naloga: glasovanje za Pesem Evrovizije

Glasovanje za Pesem Evrovizije je morda celo zanimivejše od samega poslušanja skladb na tem dogodku. Tisti, ki dolga leta spremljajo glasovanje, sicer pravijo, da ni najbolj objektivno (glej spodnjo ilustracijo). Predstavniki posameznih držav naj bi glasovali pristransko in pri tem favorizirali nastopajoče iz bližnjih ali sorodnih držav.



Je to res? Najenostavnejši način, da to preverimo je, da analiziramo podatke iz preteklih glasovanj. Na voljo so podatki o glasovanju v finalnem delu (dobili smo jih na portalu Kaggle). Primerjaj države med sabo tako, da oceniš razdaljo med njimi glede na njihov profil (vektor) glasovanja. V programski kodi (Python) razvij postopek za hierarhično razvrščanje v skupine in izriši dendrogram držav - izris je lahko tekstovni (obvezno) ali grafični (za dodatne točke), oba moraš razviti sam. Pri delu uporabi predlogo in teste.

Ali razvrščanje poišče smiselne skupine? Katere so te skupine? Za vsako od skupin navedi, katere države izbira preferenčno (jih ima raje) in katere ne, oziroma katerim državam ta skupina dodeljuje nadpovprečne oziroma podpovprečne ocene. Razumevanje in razlaga rezultatov je pomemben del tvoje domače naloge, zato se pri tem delu potrudi.

V nalogi boš moral(a) ustrezno rešiti kar nekaj problemov. Na primer, kako boš zapisal(a) podatke v primerni obliki? Kako boš združil(a) podatke iz posameznih let? Kako boš upošteval mankajoče podatke?

Dodatno (+15%): Grafičen izris dendrograma z lepo razvidnimi razdaljami, ki ste ga v Pythonu razvili sami.

Oddaja: Oddajte eno datoteko (.zip) s celotno kodo projekta (in podatki) ter poročilom v .pdf formatu. Poročilo naj bo sestavljeno le iz naslednjih razdelkov:

1. **Podatki.** Katere podatke ste analizirali? Kako ste iz podatkov izluščili profile glasovanja? (1 odstavek)
2. **Računanje razdalj.** Kako ste računali razdalje med posameznimi profili ter med posameznimi skupinami? Kaj ste naredili z neznanimi vrednostmi? (1 odstavek)
3. **Dendrogram.** Vključite tekstovni diagram (vsi) in po želji tudi grafičnega (dodatna naloga). (1 ali 2 sliki)
4. **Skupine in njihove preferenčne izbire.** Glede na rezultate razvrščanja smiselno določite skupine. Vsaki skupini v tabeli dopišite preferirane države in tiste, za katere ne glasujejo. Opišite, kako ste določili skupine in preferirane države. (1 tabela in 1 odstavek)

Za pisanje poročila uporabite LaTeX predlogo obajavljno na učilnici. V kodi ne uporabljajte obstoječih implementacij za razvrščanje in računanje razdalj (scipy in scikit-learn). Pri vizualizaciji si lahko pomagate z matplotlib, prav tako je dovoljena uporaba csv, math ter numpy.

Poskrbite, da bo vaša programska koda čitljiva in komentirana. Uporabite priloženo ogrodje in teste. V skripti naloga1.py zamenjajte klice *pass* z vašo kodo. Lahko dodate tudi pomožne funkcije, argumentov, imen in klicev obstoječih funkcij pa ne spreminjajte, saj se del naloge preverja avtomatsko. Zato da lahko pred oddajo preverite ali naloga pravilno uporablja ogrodje, smo pripravili skripto, ki jo poženete z naslednjim ukazom:

```
python test_naloga1.py
```

Skripta preveri tudi pravilnost funkcij za razvrščanje ter merjenje razdalj.

Pomoč

Branje datoteke

Če datoteko z ocenami odpirate v Pythonu 3, vam bo Python najverjetneje javil težave pri dekodiranju besedila. Priporočam, da datoteko odprete s parametrom encoding="latin1":

```
f = open("eurovision-final.csv", "rt", encoding="latin1")
```

Nato jo boste lahko prebrali z modulom csv:

```
import csv
for l in csv.reader(f):
    print(l)
```

Primer tekstovnega dendrograma

Preprost tekstovni dendrogram, iz katerega jasno vidimo hierarhijo, lahko s preprosto rekurzivno funkcijo izgleda takole:

```
            ---- Branka
        ----|
                ---- Edo
            ----|
                ---- Polona
    ----|
            ---- Helena
        ----|
            ---- Zala
----|
            ---- Nika
        ----|
                ---- Albert
            ----|
                ---- Ivan
    ----|
                    ---- Franci
                ----|
                        ---- Cene
                    ----|
                        ---- Leon
            ----|
                    ---- Jana
                ----|
                        ---- Dea
                    ----|
                        ---- Metka
        ----|
                ---- Rajko
            ----|
                ---- Stane
```

Status oddaje naloge

Število oddaj	To je vaš 1 poskus.
Status oddaje naloge	Neoddano

Stanje ocen	Neocenjeno
Rok za oddajo	torek, 23. oktober 2018, 08:00
Preostali čas	17 dni 16 ure
Zadnja sprememba	-
Komentar oddaje	Komentarji (0)

Oddaj nalogo

You have not made a submission yet

◄ PyCharm

Skok na...