

I-SUNS: Zadanie č.4

UČENIE BEZ UČITEĽA

Vo vybranom programovacom jazyku implementujte program, ktorý bude zhlukovať počítačové hry na platforme [Steam](#).

Čas odovzdania je určený časom vloženia do AIS. Deadline pre získanie 8 bodov je pred vaším cvičením v 10. týždni semestra (**26/27.11.2020**).

Dáta

K dispozícii sú súbory s dátami, ktoré obsahujú údaje o hrách. Pre vás sú minimom na vypracovanie tohoto zadania:

- steam.csv
- steam_spy_tag_data.csv

Ostatné súbory sú dobrovoľné. Popis nejednoznačných stĺpcov je na konci zadania.

Úlohy

1. **Načítajte dáta a pripravte ich na spracovanie.** Tentokrát sú dáta rozdelené vo viacerých súboroch (mali by sa dať spojiť pomocou ID hry *steam_app_id*). Mnoho zo stĺpcov je v textovej forme; skúste sa s tým vysporiadať (minimálne spracujte developera/vydavateľa a žánre). Spravte EDA (Exploratory data analysis). Dáta nezapodnajte normalizovať. **2b**
2. **Spravte nad dátami clustering.** Vyberte si jeden model s vopred určeným počtom zhlukov **1b** a jeden bez vopred určeného počtu zhlukov **1b**. Vyberte také nastavenia modelov, aby ste vedeli výsledky ďalej analyzovať.
3. **Vizualizujte výsledky.** Zmenšite dimenziu vstupných dát (napr. PCA, TsNE). Zobrazte hry v 2D/3D priestore a „vyfarbite“ ich podľa zhlukov **2b**
4. **Analyzujte výsledky** - Odpovedzte napríklad na otázky **2b**:
 - Ktorá hra je vhodným reprezentantom pre daný zhluk? (ukážte na pár zhlukov).

- Koľko vzniká outlierov/šumu pri vašom zhlukovaní? Aké sú to hry; prečo si myslíme, že je tomu tak?
- Dá sa ku (aspoň niektorým zhlukom) priradiť slovný názor/kategória?
- Myslíte, že takéto zhlukovanie by pomohlo pri vytvorení napr. doporučovacieho systému? (Môžete čerpať z vlastných skúseností).

Nepovinné úlohy

- Natrénujte a analyzujte Samoorganizujúcu sa mapu. **1-2b**
- Vhodne kódujte požiadavky na hardvér a využite ich vo svojich analýzach. **1b**
- Vráťte sa ku starým zadaniam a pokúste sa im zvýšiť úspešnosť modelmi na zníženia dimenzie (vysoko korelované príznaky sa dajú zlúčiť do jedného všeobecného) **1-2b**

Poznámky, spresnenia, odkazy

- Zadanie má dve časti:
 1. vytvorenie kódu a spracovanie vzoriek
 2. napísanie dokumentácie

Aby bolo zadanie považované za odovzdané je potreba spraviť obe. Nepodceňte dokumentáciu - je potreba sa v nej vyjadriť ku vstupom a výstupom a každej podúlohe v zadaní (aspoň niekoľkými slovami - metódy, výsledky, ak sa to hodí aj obrázky). Takisto (keďže zadanie neodovzdávate osobne) je dobré svoj postup odôvodňovať.

- Je potrebné, aby bolo zadanie pred časom vášho cvičenia nahraté v AIS v prislúchajúcom mieste odovzdania.
- Dobre čítajte dokumentáciu metód, ktoré používate.
- Nie ste hodnotení na základe úspešnosti vašich modelov, ale pri zlých výsledkoch je očakávaná aspoň snaha ich zlepšiť a pochopenie, prečo tomu tak bolo.

Upresnenie stĺpcov

- *appid* - Unique identifier for each title.
- *name* - Title of app (game).

- *release_date* - Release date in format YYYY-MM-DD.
- *english* - Language support: 1 if is in English.
- *developer* - Name (or names) of developer(s). Semicolon delimited if multiple.
- *publisher* - Name (or names) of publisher(s). Semicolon delimited if multiple.
- *platforms* - Semicolon delimited list of supported platforms. At most includes: windows;mac;linux.
- *required_age* - Minimum required age according to PEGI UK standards. Many with 0 are unrated or unsupplied.
- *categories* - Semicolon delimited list of game categories, e.g. single-player;multi-player.
- *genres* - Semicolon delimited list of game genres, e.g. action;adventure.
- *achievements* - Number of in-games achievements, if any.
- *positive_ratings* - Number of positive ratings, from SteamSpy.
- *negative_ratings* - Number of negative ratings, from SteamSpy.
- *average_playtime* - Average user playtime, from SteamSpy.
- *owners* - Estimated number of owners. Contains lower and upper bound (like 20000-50000). May wish to take mid-point or lower bound. Included both to give options.
- *price* - Current full price of title in GBP, (pounds sterling).