Zadanie č. 2

Pripravené dáta z úlohy č.1 môžu obsahovať veľa rovnakých vzoriek. Cieľom tohoto zadania bude vyhľadať rovnaké alebo podobné vzorky, ktoré sa nachádzajú medzi množinami. Tieto vzorky následne odstránime a tak sa vyheme ich duplicite (Odstránte ich z trénovacej množiny).

Úloha č.1.

Zistite prienik medzi trénovacími, validačnými a testovacími dátami. Myslite na to, že obrazy nemusia byť úplne rovnaké ale aj veľmi podobné. (P.S.: Máte veľké množstvo dát. Dajte si pozor aj na rýchlosť spracovania. Pokiaľ nenájdete úplné riešenie pracujte s menšou mnoužinou dát.)

Úloha č.2.

Nájdite podobné skupiny znakov pomocou zhlukovacích algoritmov K-means, DBSCAN a SOM. Po nájdení skupiny podobných znakov (zhlukov), vykreslite priemerný obraz a obraz najbližší priemernému obrazu. Pre každý nájdený zhluk vypíšte počet prvkov v zhluku a štatistickú odchýlku celého zhluku. DBSCAN musí vyhľadať aspoň 10 zhlukov pričom každý zhluk musí obsahovať apsoň 30 prvkov. Tieto algoritmy spustite aspoň na vašich validačných a testovacích dátach, ktoré budú mať aspoň 10 000 prvkov.

BONUS 1 : Nájdite podobné znaky v zahashovaných obrázkoch. HINT: locally sensitive hashing, perceptual hasning, difference hasshing. Existuju python kniznice napr. [tato](https://pypi.python.org/pypi/ImageHash) a pre matlab [tato](http://users.ece.utexas.edu/~bevans/projects/hashing/toolbox/index.html). (2b)