Twitter felhasználók kirándulási szokásai

Szeiler Pál
Témavezető: Béres Ferenc, Molnár András József, Benczúr András

Az adatok tisztítása

A legfontosabb adat a felhasználók leírásai
A félév végén csak az angol nelvű leírásokra hagyatkoztunk
Tisztítás:

- stopwordok
- emojik
- tokenizálás

Használt algoritmusok

Két főbb probléma:

- szavak átalakítása vektorokká
- dimenziócsökkentés

Amiket próbáltunk:

- Word2Vec
- PCA, KPCA, T-SNE

Word2Vec

- neurális hálókon alapszik
- CBOW, SG modell
- one-hot-encoding $\rightarrow$ beágyazott vektorok
- fontos paraméterek: beágyazott vektorok mérete, window size

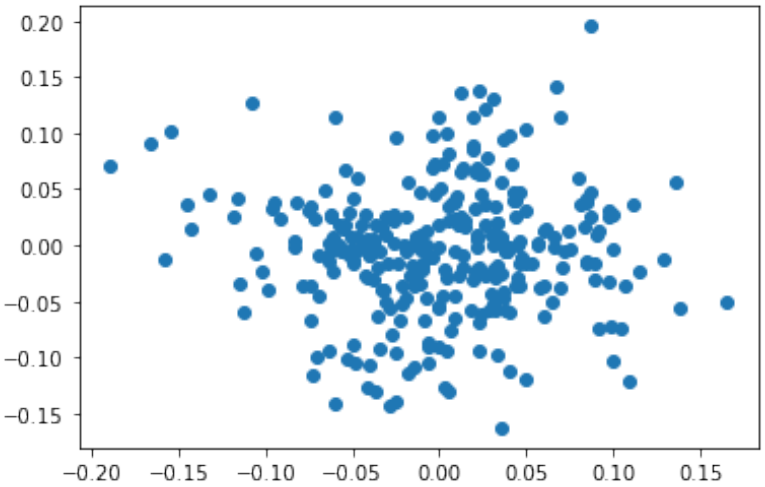
PCA, KPCA

PCA:

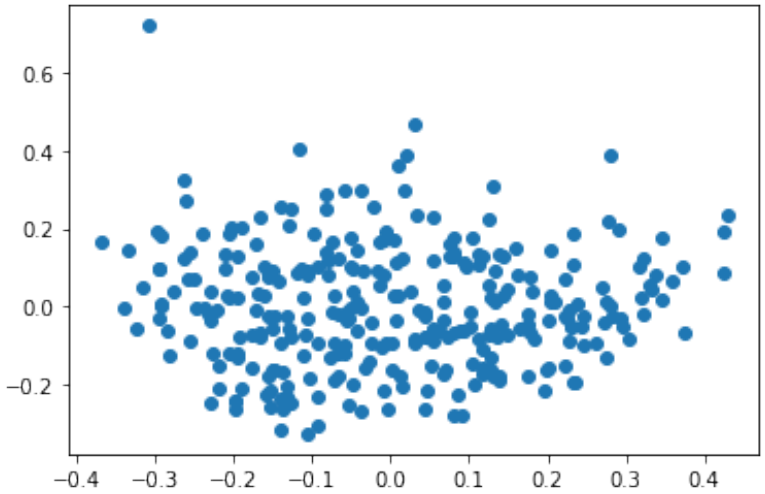
- vesszük az adathalmazra legjobban illeszkedő vektort
- i. lépés: vesszük az adathalmazra legjobban illeszkedő vektort, mely merőleges az első $i-1$ kiválasztott vektorra
- a kapott vektorok a főkomponensek
- első k főkomponens $\rightarrow$ az adat egy k dimenziós térbe transzformálható

KPCA:

- ötlet: az adat transzformálása egy m dimenziós térbe, ahol lineárisan szeparálható
- ehhez kernel: $\mathcal{K}_{i,j} = \langle \phi(x_i), \phi(x_j) \rangle$
- gyakori választás a Gauss vagy a polinomiális kernel



(a) PCA



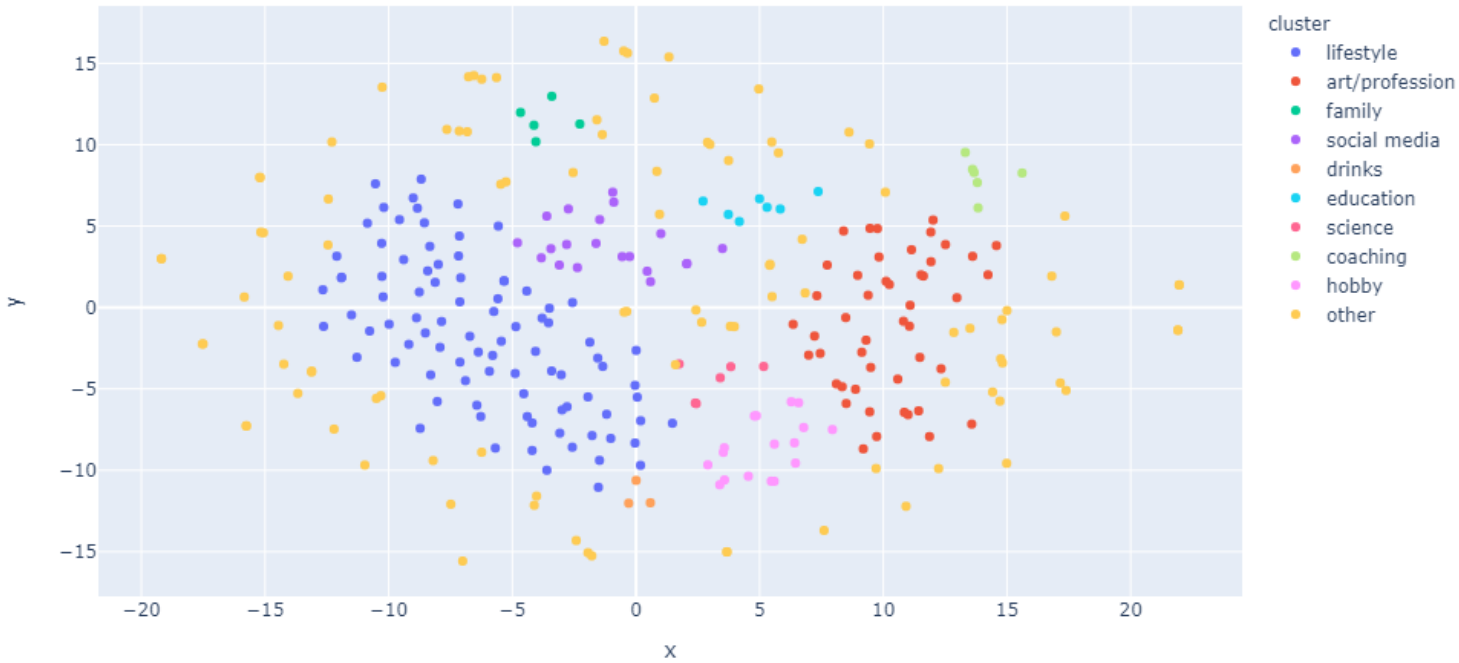
(b) Kernel-PCA .

T-SNE

- adatpontpárokhoz valószínűség normális eloszlás szerint
- a kétdimenziós tér pontpárjaihoz szintén valószínűség t eloszlás szerint
- minimalizáljuk a KL divergenciát

Klaszterezés 2 dimenzióban

- a leggyakoribb szavakat klasztereztük
- dimenziócsökkentéshez T-SNE
- klaszterezéshez DBSCAN
- cél: kevés outlier, elég sok klaszter



Felhasználók kategorizálása és klaszterezése

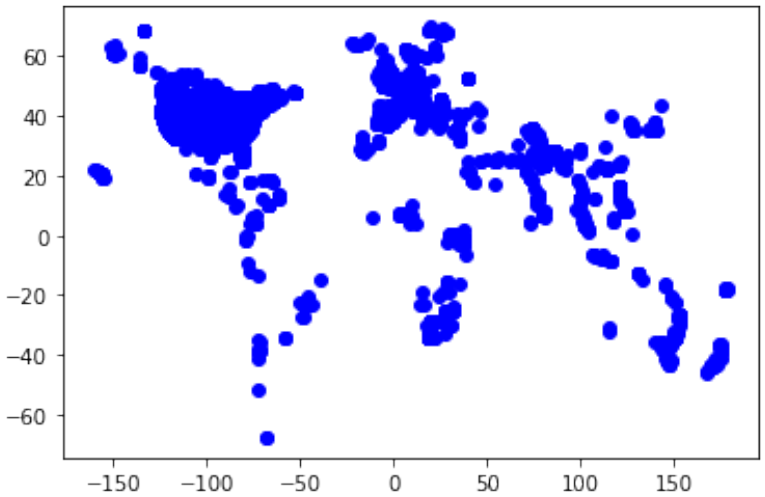
A kategóriák a leggyakoribb szavak klaszterei
Minden felhasználóhoz egy vektor:

$$v_{user} = \frac{\sum_i u_i}{|description(user)|}$$

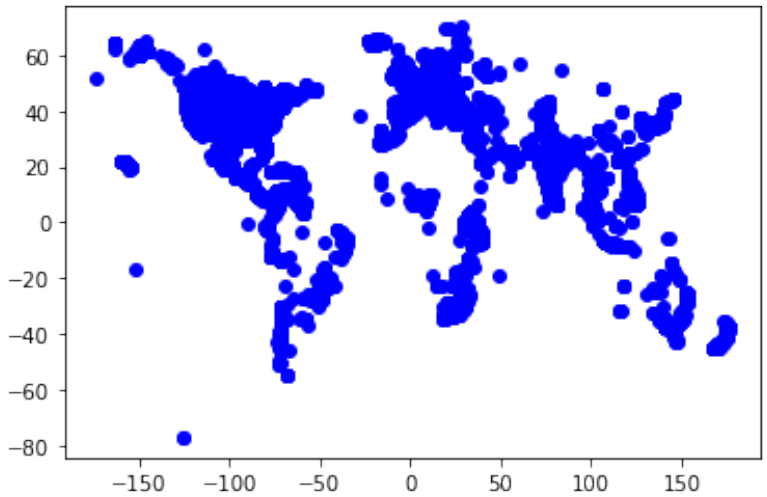
- klaszterközéppontoktól való távolság
- klaszterektől való minimális távolság

Összekapcsolás az előző félévvel

Az előző félévben végzett távolságméréseket megnéztük klaszterenként
Nagy különbségeket nem tapasztaltunk
Összehasonlítottuk, hogy honnan tweetelnek a különböző klaszterbe sorolt userek
Itt már jelentős eltérések vannak



(c) Education



(d) Art/profession

További tervek

- A különböző klaszterekbe tartozó felhasználók preferenciáinak megismerése
- Ajánlórendszer építése