Twitter felhasználók kirándulási szokásai

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

1 Bevezetés

Az előző félévben Twitter felhasználók utazásait vizsgáltam. Ebben a félévben csak a felhasználók leírásaiból kinyerhető adatokkal dolgoztam, a legfontosabb kérdés volt, hogy lehetséges-e klaszterekbe osztani a felhasználókat vagy vannak-e esetleg olyan kategóriák, melyekbe a felhasználókat be lehet sorolni csupán a leírójuk alapján. Majd az előző féléves eredményeket felhasználva vizsgáltuk a különböző kategóriákban az utazási szokásokat.

2 Az adat előkészítése

Ahhoz, hogy a leírásokat használni tudjuk, szükséges volt néhány feldolgozó lépés elvégzése. Az nltk könyvtár segítségével a szövegekből kivettük a speciális karaktereket, az olyan gyakori szavakat, mint például a "the" és tokenizáltuk a leírókat. A félév során dolgoztunk a teljes adathalmazzal is, valamint a csupán angol nyelvű leírókkal. A tokenizálás után a leírásokból betanítottunk egy Word2Vec modellt. Sajnos előre betanított szó beágyazást nem tudtunk használni a sok új szó miatt. Mivel a felhasználók több mint fele az Egyesült Államokban vagy Angliában él (1. ábra), így a végén csak az angol nyelvű leírásokkal foglalkoztunk. Érdekesség, hogyha nem csak az angol nyelvű felhasználókat vesszük figyelembe, akkor a Word2Vec modell vektorainak 2 dimenzióra való redukálása után az idegen nyelvű szavak egy önálló klasztert alkotnak messze az angol szavaktól. A következő fejezetekben részletesen ismertetem a félév során használt eljárásokat.

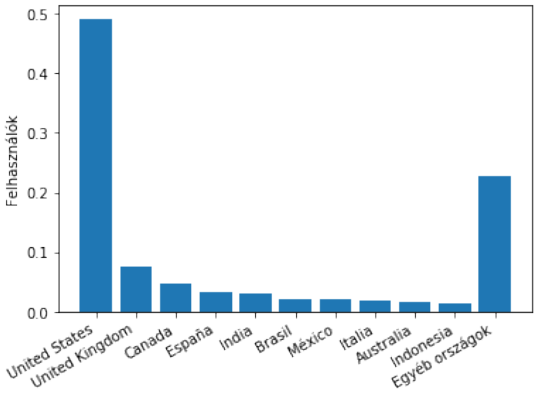


Figure 1: Az adatban előforduló felhasználók országonkénti eloszlása.

3 Word2vec

Bevett módszer a szavak vektorként való ábrázolása, ehhez használtuk a Word2Vecet, mely egy neurális hálókon alapuló eljárás. A Word2Vecnek két lehetséges megvalósítása van:

A **Continuous-bag-of-words** (CBOW) modellben egy szó megtippelése a feladat, ismerve a környezetét. A környezetében lévő szavak sorrendje nem számít. A **Skip-gram** modellben egy szó alapján megmondjuk a szövegkörnyezetét. Mindkét esetben egy szó szövegkörnyezete az előtte és utána lévő előre meghatározott számú szó. Ezt ablaknak is hívják (window).

A módszer bemenetként a szöveg 0-1 értékekkel elkódolt vektoros ábrázolását kapja meg. A neurális háló két rétegből áll: az első egy rejtett réteg, annyi neuront tartalmaz, ahány dimenziós vektorokat szeretnénk készíteni, az output rétegben a neuronok száma megegyezik az input méretével, azaz az adatban található összes szó számával.

Az általunk épített Word2Vec modellben 100 dimenziós beágyazott vektorokat készített. Próbáltunk más dimenziót is, de nem volt nagy különbség. A modellbe nem vettük bele a ritka szavakat, amelyek kevesebb mint tízszer fordulnak elő az adatban. A szavak után minden felhasználóhoz is rendeltünk egy vektort, a leírójukban szereplő szavak vektorait összeadtuk majd koordinátánként átlagoltunk.

4 Dimenziócsökkentés

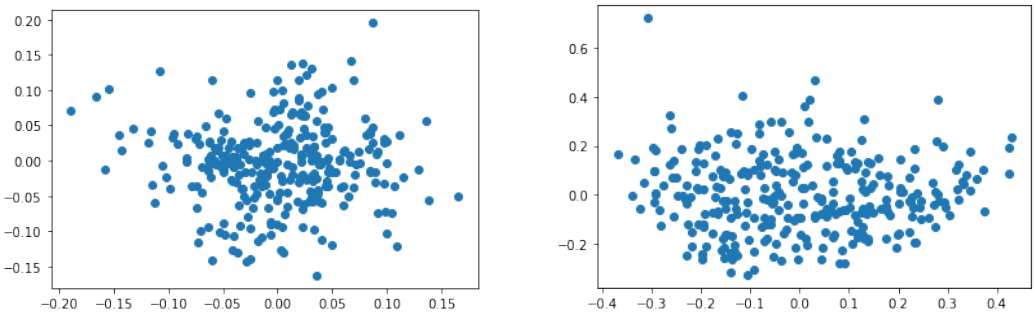
Nagy dimenziós adathalmazok esetében megfigyelhető a "dimenzió átka" jelenség, melyszerint magas dimenzióban az adatpontok távol vannak egymástól. Emellett ilyen adathalmazokat ábrázolni sem lehet, ezért szükség van olyan eljárásokra, amelyek kisebb dimenziós terekbe -a mi esetünkben a síkba- képzik az adatot megtartva az adatpontok közötti fontosabb kapcsolatokat. A félév során három dimenziócsökkentő algoritmust próbáltunk ki, ezeket mutatom be röviden. A 2 és a 3 ábrákon a 300 leggyakoribb szót vetítettük két dimenzióba, ezeken látható a módszerek közötti különbség.

4.1 PCA

Az adathalmazra először is a legjobban illeszkedő vektort illesztjük, majd minden további vektort úgy választunk meg, hogy az előzőre merőleges, az adatra pedig a lehető legjobban illeszkedjen. Az eljárás végén a vektorokat normáljuk. Ezeket a vektorokat főkomponenseknek hívjuk. Az első k komponens megtartásával az adatot egy k dimenziós térbe transzformálhatjuk. Mivel ez egy lineáris transzformáció, nem minden adathalmazra használható jól. A mi esetünkben sem ez adta a legjobb eredményt.

4.2 Kernel-PCA

A K-PCA ötlete a kernel módszereken alapszik, azaz ha az adat nem szeparálható lineárisan az n dimenziós térben, akkor felvetítve egy alkalmas $m > n$ dimenziós



(a) PCA

(b) Kernel-PCA .

Figure 2: A 300 leggyakoribb szó 2 dimenzióba történő vetítése.

térbe egy ϕ transzformációval már lineárisan szeparálható. Ez a tér akár egy végtelen dimenziós RKHS is lehet.

$$\mathcal{K}_{i,j} = \langle \phi(x_i), \phi(x_j) \rangle$$

A $\mathcal{K}$ mátrixot hívjuk kernelnek. Gyakori választás a Gauss vagy a polinomiális kernel. Ezek után a PCA lépéseit követjük. A K-PCA így nemcsak lineárisan szeparált adattal tud dolgozni, a mi adathalmazunkra ez jobbnak bizonyult mint a PCA. Az 2 ábrán látható a PCA és Kernel-PCA közötti különbség a leggyakoribb szavak esetében.

4.3 T-SNE

A T-SNE egy nemlineáris, az előző két módszertől jelentősen eltérő dimenziócsökkentő eljárás. Az adatpontpárokhoz valószínűségeket rendelünk normális eloszlás szerint, minél közelebb vannak egymáshoz, annál nagyobb ez a valószínűség. Ugyanígy a kétdimenziós tér pontpárjaihoz is rendelünk valószínűségeket student-t eloszlás szerint. Majd az eljárás a két eloszlás közötti KL divergenciát minimalizálja.
T-SNE algoritmust használtunk különböző paraméterezéseket kipróbálva a leggyakoribb szavak klaszterezéséhez, ez a 3 ábrán látható.

5 Klaszterezés: szavak vagy felhasználók?

A félév elején reménykedtünk abban, hogy a felhasználókat tudjuk klaszterekre osztani, de mivel túl sok felhasználó volt az adatban és nem estek szét nagyobb halmazokra még T-SNE-vel sem, erről letettünk. Ezek után a leggyakoribb szavakat klasztereztük, amihez több algoritmust is kipróbáltunk, de a legjobb a DBSCAN volt, ezzel ugyanis nemcsak konvex klaszterek jöhetnek létre. A klaszterezés során úgy optimalizáltuk a hiperparamétereket, hogy ne legyen se túl sok, se túl kevés klaszter, viszont ezek a klaszterek valóban különüljenek el egymástól. Az általánosságban elmondható, hogy valamennyi értelmes paraméterbeállítás mellett keletkezett egy nagy klaszter, mely eléggé általános szavakat tartalmazott. Érdekes volt megfigyelni, hogyan változnak a szavak ennek a klaszternek a szélén. Próbáltuk minimalizálni az outlierek számát is, de ezzel együtt nőttek az alig néhány pontból álló klaszterek, vagy ellenkezőleg, túl kevés klaszter keletkezett.

Figure 3: A 300 leggyakoribb szó két dimenzióba vetítése T-SNE-vel és a klaszterek.

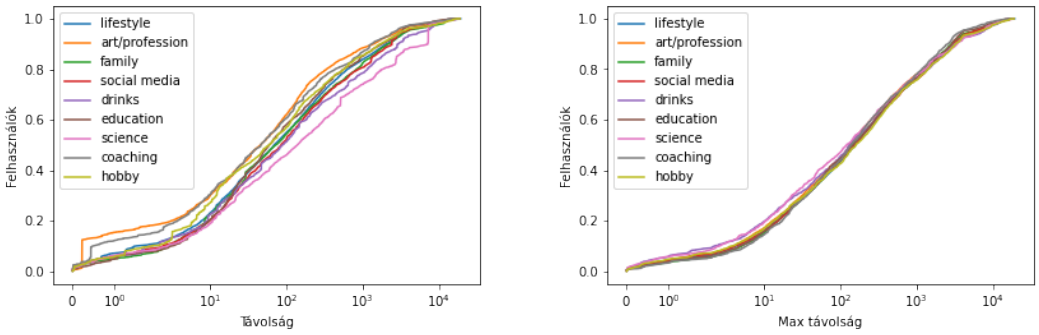
6 A felhasználók kategorizálása

Ugyan klaszterezni nem tudtuk a felhasználókat, ezért hogy kapjunk valamennyi információt a felhasználókról, megnéztük a klaszterközéppontoktól való távolságukat, valamint a klaszterektől való minimális távolságot. Ezek alapján azt lehetett megmondani, mennyire illik bele az adott felhasználó a különböző kategóriákba. Itt értelemszerűen az outlierek által alkotott klasztert figyelmen kívül hagytuk. Ismerve a távolságokat, a korrelációs mátrixot is ki lehet számolni, de az csak azt mutatja meg, hogy mely klaszterek vannak egymáshoz közel.

7 Összekapcsolás az előző félévvel

Végül a felhasználókat klasztereztük a minimális távolság alapján, és megnéztük, hogy van-e különbség a különböző klaszterekhez tartozó felhasználók kirándulási szokásai között abban a tekintetben, hogy milyen messze utaznak. Mivel eléggé általánosak a klaszterek, ezért nem vártunk nagy eltéréseket, és ez igaz is lett, amint az a 4 ábrán látható.

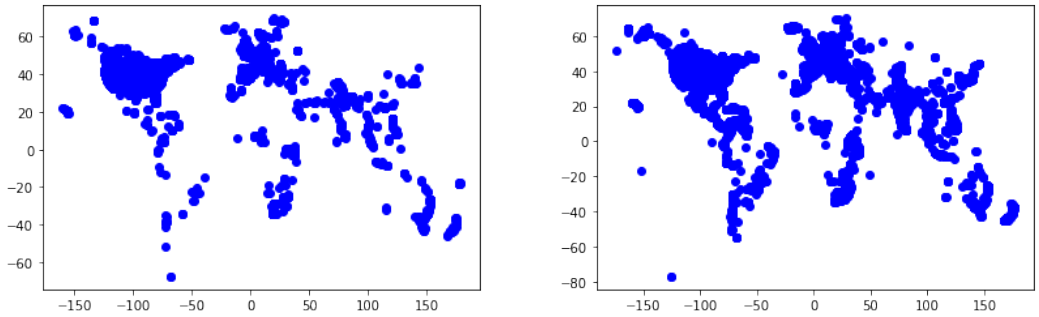
A felhasználók alapján a tweetek is klaszterekbe sorolhatók, így meg tudtuk nézni, hogy a különböző klaszterekbe tartozó felhasználók mely térségeket preferálják. Itt fontos kiemelni, hogy csak angol nyelvű felhasználók tweetjeit vettük figyelembe és az előző féléves elemzések alapján a külföldre utazás mennyisége eltörpül a belföldön tett kirándulásokétól. A 5 ábrán látható két klaszter és az ezekbe tartozó userek utazási célpontjai. Az education klaszterbe tartozó felhasználóktól nagyjából 30 000 tweetünk van, míg az art/profession klaszterbe tartozóktól 20 000. Így is látszik, hogy Dél-Amerikát, Afrikát, Ázsiát és Ausztráliát kevésbé preferálja az education kategória, ellenben az art/profession kategóriával.



(a) A távolságok eloszlása.

(b) A maximális távolságok eloszlása.

Figure 4: Távolságeloszlások klaszterenként



(a) Az education kategóriába sorolt felhasználók tweetjeinek térbeli eloszlása.

(b) Az art/profession kategóriába sorolt felhasználók tweetjeinek térbeli eloszlása.

Figure 5: A tweetek eloszlása különböző kategóriákban.

Irodalomjegyzék

[1] Arjan S. Gosal et al. "Using social media, machine learning and natural language processing to map multiple recreational beneficiaries". In: *Ecosystem Services* 38 (2019), p. 100958. ISSN: 2212-0416. DOI: https://doi.org/10.1016/j.ecoser.2019.100958. URL: https://www.sciencedirect.com/science/article/pii/S2212041618302985.

[2] Ricardo Moreno-Llorca et al. "Evaluating tourist profiles and nature-based experiences in Biosphere Reserves using Flickr: Matches and mismatches between online social surveys and photo content analysis". In: *Science of The Total Environment* 737 (2020), p. 140067. ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2020.140067. URL: https://www.sciencedirect.com/science/article/pii/S0048969720335877.

[3] Seunghyun Brian Park et al. "Visualizing theme park visitors' emotions using social media analytics and geospatial analytics". In: *Tourism Management* 80 (2020), p. 104127. ISSN: 0261-5177. DOI: https://doi.org/10.1016/j.tourman.2020.104127. URL: https://www.sciencedirect.com/science/article/pii/S0261517720300534.