Szóbeágyazás-modellek geometriai tulajdonságainak összehasonlító vizsgálata

Cserháti Réka
II. évf. matematika BSc

Témavezető: Dr. Berend Gábor

SZTE TTIK Számítógépes Algoritmusok és Mesterséges Intelligencia Tanszék

A napjainkban a tudományban és iparban is egyre népszerűbb természetesnyelv-feldolgozás területén a szóbeágyazások, vagyis olyan algoritmusok, melyek egy szövegkorpusz alapján annak szavait egy *n*-dimenziós vektortér pontjainak feleltetik meg, a korszerű technológiáknak elengedhetetlen részévé váltak. Az ezekkel létrehozott szóbeágyazás-modellek információt hordoznak az eltárolt szavak jelentéséről, ami sokat javít a gépi tanulásos modellek teljesítményén, legyen szó egyszerű klasszifikációs modellekről (dokumentumosztályozás, szentimentelemzés) vagy komplex feladatokhoz (mint gépi fordítás, kérdésmegválaszolás, információkeresés) szükséges bonyolultabb struktúrájú algoritmusokról.

Bár a szóbeágyazások területén is rendszeresen születnek új, fejlettebb megvalósítások, melyek előrelépést hoznak az alkalmazásokban, a régebbi módszerek működésében is sok megválaszolatlan kérdés maradt, és a szóbeágyazások a természetesnyelv-feldolgozás egy széles körben használt, de meg nem értett feketedobozává ('black box'-ává) váltak. Dolgozatomban a vektoros szóreprezentációk mélyebb megértésére tett erőfeszítéseket próbálom segíteni az ilyen modellek geometriai tulajdonságainak vizsgálatával.

Arra a kérdésre keresünk választ, hogy a különböző módszerekkel létrehozott modelleknek milyen közös tulajdonságai vannak, és mik azok, amiben különböznek. Ehhez kétféle korpuszon, 5 különböző algoritmussal, és ezeken belül is különböző hiperparaméterekkel létrehozott modelleket használunk.

A használt algoritmusok leírása és a kapcsolódó szakirodalom áttekintése után először a vektorok általános és egymáshoz viszonyított elhelyezkedését vizsgáljuk, a modelleket egymással és normáleloszlás szerint generált véletlenszerű vektorokkal is összehasonlítva. Itt a vektorok hossza érdekes tulajdonságnak bizonyul, így ezután az ezt befolyásoló tényezőkre derítünk több fényt, megkérdőjelezve és pontosítva azt az eddig a köztudatban élő nézetet, hogy mennél gyakoribb egy szó, annál hosszabb lesz a hozzá tartozó vektor.

Végül a szavak hasonlóságát és szomszédságát helyezzük a középpontba, feltételezve, hogy ha az algoritmusok ugyanazt az információt tanulják, egy adott szóhoz nagyjából ugyanazok lesznek hasonlóak modellválasztástól függetlenül – de kiderül, hogy ez sem ilyen egyszerűen igaz.