Szakdolgozat kivonat

Szóvektorok matematikai tulajdonságai

Szente Zsófia

Témavezető: Borbély Gábor

A számítástudomány és a mesterséges intelligencia egyik területe a természetes nyelvfeldolgozás. A természetes nyelvfeldolgozásnak az a célja, hogy az ember és a gép közötti kommunikációt segítse, melyet az emberi nyelv feldolgozásával éri el. Szakdolgozatomban a szavakat vektorként reprezentáló modellek egyikét a Word2Vec felépítését, működését mutatom be. A szavak olyan reprezentálását szeretnénk elérni, amellyel jól értelmezhető, számunkra hasznos információkat nyújtó ábrázolást kapjunk, ezáltal jósolni tudjuk a környező szavakat. Bemutatom a Word2Vec két modelljét a Skip-gram és a Continuous Bag of Words modelleket. A Skip-gram modell egy szó alapján próbálja megjósolni a környező szavakat, a Continuous Bag of Words viszont ellenkezőleg, egy szót jósol egy környezete alapján. Bevezetem a modellekhez használt legfontosabb definíciókat, állításokat és tételeket.

A modellek minőségének, számítási igényének javítása érdekében optimalizálási technikákat mutatok be (hierarchikus szoftmax, negatív mintavételezés, és a gyakori szavak alul-mintavételezése).

A Word2Vec modelljein kívül, a GloVe modellre is kitérek. A GloVe modell a korpuszban szereplő szavak előfordulásának a statisztikáját vizsgálja, ami az egyik legfontosabb információkat tartalmazza minden felügyelet nélküli folyamat számára a szavak reprezentálásához.

Ezután a Skip-gram modellre megvizsgálom a kompozíciót. A kompozíció additivitáshoz a szavak egyenletes eloszlása kell. Az átalakításomban az eredeti modellt próbáltam megváltoztatni oly módon, hogy a későbbiekben a szavak valószínűsége ne befolyásolják a kompozíció additivitás teljesülését. A levezetéshez felhasználtam a KL-divergencia és a negatív kereszt-entrópia közötti összefüggést, melynek felhasználásával elérjük, hogy a szavak parafrázisa felírható lesz a szavak összegeként.

A szakdolgozatom utolsó részében a szavak beágyazását felhasználó mondat beágyazást ismertetem, ahol a modell bemeneteként nem egy szót tekintünk, hanem egy egész mondatot.