Gensim word2vec amazon reviews

Ezek állítgatásával próbálgattam befolyásolni
- vektor dimenziók száma
- window
- epochs

szétszedve az inicializálást, vocab buildelést és trainelést
memóriában tárolás szükséges ha preprocesszelni szeretnénk a szöveget nem?

Conceptnet

Az ötlet az volt, hogy mivel minden sor egy élről tárol el információt és a kezdő- és végcsúcsot is, így megjelenésük számát számolnánk a csúcsoknak, ahányszor megjelenik annyi éle van (kimenő vagy bemenő), ahány sor beolvasva annyi él van összesen, ahány egyedi szó annyi csúcs összesen. Ezen információk alapján már ki lehetne rajzolni a grafikont, csak ki kellene nyerni az adatokat. < --- ezzel kezdtem el foglalkozni.

Top5 szó

Ezek állítgatásával próbálgattam befolyásolni
- szavak mennyisége
- dimenziók száma
- window

A leggyakoribb szó, “goes”, szinte mindig benne van ami gondolom jó dolog.

Eisenstein

14.1

ismeretlen szó jelentésének megtippelése kontextus alapján

14.2

szavak reprezentálása vektorokkal (word embedding)
valós számok helyett bitstringgel (brown cluster)
kontextus ha kicsi, általánosabb, ha nagyok specifikusabb

14.3

Latent Semantic Analysis (LSA)

korrelálatlan mátrixok, hogy minden dimenzió egyedi információt tároljon
pointvise mutual information (PMI), egyfajta transzformáció. Ha a mátrixot transzformáljuk ezzel mielőtt alkalmaznánk az SVD-t hatékonyabb lesz az LSA-nk

PMI

megnézi hogy egy szó egy kontextusba belepasszol-e (degree of assotiation)

a negatív és nulla értékek kiküszübölésére positive PMI. Ha PMI(i,j) <= 0 akkor return 0

14.4

Brown Clusters

Diszkrét reprezentáció, klaszterizálás által

Ha kevés klaszterünk van a benne lévő szavakban nem lesz túl sok közös

Megoldás: Hierarchikus klaszterezés

Minden bitstring egy útvonal egy fában

14.5

CBOW (Continous bag-of-words)

kontextusból jóslunk szót

helyi kontextus kiszámolása szomszédos szavak átlaga (a szavak vektorosan vannak)
sorrend nem számít, vektorokon kondícionálunk, folyamatos

Skipgram

szóból jóslunk kontextust

minden szó többször generálódik, minden alkalommal egy szón kondícionálódik

CBOW-t trainelni gyorsabb, de Skipgram jobban teljesít