Ανάκτηση Πληροφορίας 2022

Word Embeddings

Ομάδα

Φραγκαθούλας Χρήστος ΑΜ 4196

Λιάτσος Νικόλαος ΑΜ 4101

Το μοντέλο που χρησιμοποιήθηκε για την άσκηση είναι ένα μοντέλο της Google που είναι pre-trained

```
In [6]: import gensim.downloader as api
        wv = api.load('word2vec-google-news-300')

[==================================================] 100.0% 1662.8/1662.8MB downloaded
```

Δοκιμάσαμε κάποια παραδείγματα όπως

```
In [13]: vec1 = wv.most_similar(positive=['man', 'computer'], topn =8)
         vec1

Out[13]: [('woman', 0.6019380688667297),
          ('computers', 0.5453090667724609),
          ('teenager', 0.5446600317955017),
          ('laptop_computer', 0.5383511781692505),
          ('boy', 0.5313993692398071),
          ('teenage_girl', 0.5139628648757935),
          ('Cops_Burglar', 0.5133906602859497),
          ('laptop', 0.5061536431312561)]

In [14]: vec2 = wv.most_similar(positive=['woman', 'computer'], topn =8)
         vec2

Out[14]: [('girl', 0.5884929895401001),
          ('teenage_girl', 0.5812638998031616),
          ('laptop_computer', 0.5632834434509277),
          ('laptop', 0.5577383041381836),
          ('computers', 0.5565006136894226),
          ('man', 0.5512093901634216),
          ('Evi_Susilowati', 0.5224096775054932),
          ('teenager', 0.5131943225860596)]

In [17]: d1 = [item[1] for item in vec1]
         d2 = [item[1] for item in vec2]
         distance.cosine(d1,d2)

Out[17]: 0.00041727742253538924

In [19]: vec2 = wv.most_similar(positive=['woman', 'householder'], topn =8)
         vec2

Out[19]: [('man', 0.6526675224304199),
          ('pensioner', 0.6297963857650757),
          ('teenage_girl', 0.6045790314674377),
          ('girl', 0.5920445919036865),
          ('TEENAGE_girl', 0.5761831998825073),
          ('dementia_sufferer', 0.5746362209320068),
          ('householders', 0.5590190887451172),
          ('AN_ELDERLY_woman', 0.5589241981506348)]

In [20]: d1 = [item[1] for item in vec1]
         d2 = [item[1] for item in vec2]
         distance.cosine(d1,d2)

Out[20]: 0.00020304546170712978
```

Σε αυτό το παράδειγμα παρατηρούμε ότι το διάνυσμα (man+computer) έχει μεγάλη ομοιότητα με το διάνυσμα (woman+computer) με ποσοστό περίπου 0,00041. Με την αντικατάσταση του τελευταίου διανύσματος σε (woman+householder) παρατηρούμε ότι η συσχέτιση μεταξύ τους είναι μεγαλύτερη σε σχέση με πριν.

Ένα επιπλέον παράδειγμα

```
In [36]: vec1 = wv.most_similar(positive=['man', 'white','income'], topn = 8)
         vec1

Out[36]: [('Responded_Letterman_How', 0.6086456775665283),
          ('black', 0.5841068029403687),
          ('woman', 0.5747525691986084),
          ('dark_complected', 0.5744969844818115),
          ('wrote_Newitz', 0.5744592547416687),
          ('wearing_gray_hoody', 0.5666198134422302),
          ('gorgeously_photographed_black', 0.5484243631362915),
          ('5ft_##ins_slim', 0.5400959253311157)]

In [37]: vec2 = wv.most_similar(positive=['man', 'black', 'punch'], topn = 8)
         vec2

Out[37]: [('Responded_Letterman_How', 0.5782788991928101),
          ('white', 0.5762711763381958),
          ('woman', 0.5435894131660461),
          ('punching', 0.5211857557296753),
          ('scruffy_unshaven', 0.5159029960632324),
          ('6ft_1ins_tall', 0.5104131698608398),
          ('_White', 0.5098403096199036),
          ('horribly_horribly_deranged', 0.509628415107727)]

In [38]: d1 = [item[1] for item in vec1]
         d2 = [item[1] for item in vec2]
         distance.cosine(d1,d2)

Out[38]: 0.00046791627038555994
```

Αυτό το παράδειγμα δείχνει την προκατάληψη που υπάρχει, ότι δηλαδή οι λευκοί άντρες έχουν εισόδημα, ενώ οι μαύροι άντρες είναι ποιο εκμεταλλεύσιμοι.

Ακόμη το παρακάτω παράδειγμα

```
In [39]:
1 vec1 = wv.most_similar(positive=['greece', 'peace'])#, 'harmony'])
2 vec1

Out[39]: [('kosovo', 0.5778459906578064),
 ('sri_lanka', 0.5691853761672974),
 ('peacefull', 0.5572632551193237),
 ('stage_Yusafzai', 0.5539507865905762),
 ('sri_lankans', 0.550804078578949),
 ('syria', 0.5496231317520142),
 ('nepal', 0.5396113395690918),
 ('Mozilla_Safari_Konqueror', 0.539513885974884),
 ('lebanon', 0.5392201542854309),
 ('israelis', 0.5380474328994751)]

In [44]:
1 vec2 = wv.most_similar(positive=['russia', 'war'])#, 'conflict'])
2 vec2

Out[44]: [('wars', 0.6593061685562134),
 ('vietnam', 0.6408215165138245),
 ('iraq', 0.6100413203239441),
 ('ww2', 0.5873371958732605),
 ('afghanistan', 0.5872629880905151),
 ('russians', 0.577458381652832),
 ('afganistan', 0.574748158454895),
 ('iran', 0.5644713640213013),
 ('WW3', 0.5615954399108887),
 ('Afganistan', 0.5457037687301636)]

In [45]:
1 d1 = [item[1] for item in vec1]
2 d2 = [item[1] for item in vec2]
3 distance.cosine(d1,d2)

Out[45]: 0.0006170719152537307
```

Πολλά κείμενα της google βασιζόμενα στην Wikipedia, φαίνεται ότι υπάρχει η σύνδεση της Ελλάδος με την ειρήνη και της Ρωσίας με τον πόλεμο.

Παρακάτω θα δούμε μια διαφορά μεταξύ δύο μοντέλων

Μοντέλο της Google

```
In [46]:
vec1 = wv.most_similar(positive=['man', 'wheel'])#, 'harmony'])
vec1

Out[46]:
[('driver', 0.6070996522903442),
 ('woman', 0.5898844003677368),
 ('motorcyclist', 0.5674616098403931),
 ('motorbike_rider', 0.544006884098053),
 ('motorist', 0.539942741394043),
 ('steering_wheel', 0.5358079671859741),
 ('suspected_purse_snatcher', 0.5357204079627991),
 ('Cops_Drunken', 0.5299758911132812),
 ('teenager', 0.5289519429206848),
 ('motorcylist', 0.528925359249115)]
```

```
In [47]:
vec2 = wv.most_similar(positive=['woman', 'lipstick'])#, 'conflict'])
vec2

Out[47]:
[('lady', 0.586341381072998),
 ('red_lipstick', 0.5832858085632324),
 ('girl', 0.5680494904518127),
 ('lipliner', 0.5476288795471191),
 ('she', 0.5443190336227417),
 ('Lip_gloss', 0.5418884754180908),
 ('lipstick_mascara', 0.529711604183472),
 ('teenage_girl', 0.5287932753562927),
 ('pearl_earring', 0.525692880153656),
 ('thong_bikini', 0.5246971845626831)]
```

```
In [48]:
d1 = [item[1] for item in vec1]
d2 = [item[1] for item in vec2]
distance.cosine(d1,d2)

Out[48]:
8.67344923314306e-05
```

Άλλο μοντέλο

```
In [164]:
vec1 = model.wv.most_similar(positive=['man', 'wheel'])#, 'harmony'])
vec1

Out[164]:
[('rider', 0.6859514117240906),
 ('blade', 0.6823872327804565),
 ('sword', 0.6486030220985413),
 ('hammer', 0.6425479054450989),
 ('knife', 0.6404703855514526),
 ('finger', 0.6373097896575928),
 ('bow', 0.6263216137886047),
 ('fingers', 0.6163617968559265),
 ('candle', 0.6130890250205994),
 ('nose', 0.609165370464325)]
```

```
In [165]:
vec2 = model.wv.most_similar(positive=['woman', 'lipstick'])#, 'conflict'])
vec2

Out[165]:
[('senex', 0.7691648602485657),
 ('broomstick', 0.7670607566833496),
 ('prostitute', 0.7601841688156128),
 ('stepfather', 0.7555566430091858),
 ('sweetheart', 0.7533393502235413),
 ('stepmother', 0.749531626701355),
 ('erdrich', 0.7486035227775574),
 ('bruise', 0.7478093504905701),
 ('totoro', 0.7457727193832397),
 ('kitten', 0.7430563569068909)]
```

```
In [166]:
d1 = [item[1] for item in vec1]
d2 = [item[1] for item in vec2]
distance.cosine(d1,d2)

Out[166]:
0.0004104612019074638
```

Τα μοντέλα πρέπει να εκπαιδεύονται με κατάλληλα κείμενα, γιατί αυτά τα μοντέλα παίζουν σημαντικό ρόλο στην καθημερινότητά μας με παράδειγμα τα παραπάνω, όπου τα 2 μοντέλα συσχετίζουν διαφορετικά τον άνδρα και την γυναίκα στην καθημερινότητάς τους. Ακόμη παρατηρούμε το ένα μοντέλο εκφράζει και συνδέει την γυναίκα με προσβλητικούς χαρακτηρισμούς, ενώ το άλλο μοντέλο τους έχει αποφύγει.

Επιπρόσθετα το επόμενο παράδειγμα

```
In [69]: vec2 = wv.most_similar(positive=['greece', 'food'])
         vec2

Out[69]: [('malta', 0.5487498044967651),
          ('hungary', 0.5485798120498657),
          ('foodstuffs', 0.5457229018211365),
          ('tinned_tomatoes', 0.5281139016151428),
          ('Cloned_meat', 0.5273177623748779),
          ('döner', 0.5237293243408203),
          ('philippine', 0.5167913436889648),
          ('food_stuffs', 0.5088459253311157),
          ('gordon_brown', 0.5083537101745605),
          ('foods', 0.5020650625228882)]

In [70]: vec2 = wv.most_similar(positive=['portugal', 'football'] )
         vec2

Out[70]: [('footbal', 0.6720271110534668),
          ('fooball', 0.6572456359863281),
          ('juventus', 0.65140300989151),
          ('Real_madrid', 0.6427035331726074),
          ('fotball', 0.6399109363555908),
          ('bayern', 0.6386366486549377),
          ('real_madrid', 0.6306421756744385),
          ('soccer', 0.6281998157501221),
          ('totti', 0.620458722114563),
          ('La_liga', 0.6201213598251343)]

In [71]: vec2 = wv.most_similar(positive=['greece','portugal'])
         d1 = [item[1] for item in vec1]
         distance.cosine(d1,d2)

Out[71]: 0.0004293256459185768
```

Αυτό μας δείχνει ότι η Ελλάδα έχει συσχετιστεί με το φαγητό, λόγω του πολιτισμού και η Πορτογαλία με το ποδόσφαιρο λόγω τον πολλών γνωστών ομάδων που διακατέχει.

Τέλος το παράδειγμα που φαίνεται παρακάτω

```
In [49]:
vec1 = wv.most_similar(positive=['democracy', 'freedom'])
vec1

Out[49]:
[('freedoms', 0.7922826409339905),
 ('liberty', 0.7657341957092285),
 ('democratic', 0.7601733207702637),
 ('democratic_freedoms', 0.7149538993835449),
 ('pluralism', 0.700656533241272),
 ('democratic_ideals', 0.648664653301239),
 ('democratization', 0.6456856727600098),
 ('pluralist_democracy', 0.6272454261779785),
 ('fundamental_freedoms', 0.6250193119049072),
 ('multiparty_democracy', 0.6240731477737427)]

In [59]:
vec2 = wv.most_similar(positive=['king', 'slavery'])
vec2

Out[59]:
[('slave', 0.658708930015564),
 ('slaves', 0.6567034721374512),
 ('queens', 0.6397448778152466),
 ('abolitionist_Sojourner_Truth', 0.5947266817092896),
 ('very_pampered_McElhatton', 0.5828427076339722),
 ('princess', 0.581714391708374),
 ('chattel_slavery', 0.5796917676925659),
 ('enslaved', 0.5599312782287598),
 ('monarch', 0.5521868467330933),
 ('Abraham_Lincoln_emancipation', 0.5490185022354126)]

In [60]:
d1 = [item[1] for item in vec1]
d2 = [item[1] for item in vec2]
distance.cosine(d1,d2)

Out[60]:
0.00048542726066302944
```

Παρουσιάζει την συσχέτιση που έχει η δημοκρατία με την ελευθερία όσο και ο Βασιλιάς με την σκλαβιά.