Practical No. 1

Jan Dziuba
jd406140

March 2, 2023

1 Part 1

e)

What clusters together in 2-dimensional embedding space?
In unnormalized case there are 2 clusters: first containing "śpiewaczka", "literatura", "poeta", "obywatel", second containing "sztuka".

In normalized case there are 3 clusters: first containing "śpiewaczka", "poeta", "obywatel", second containing "sztuka", third containing "literatura".

What doesn't cluster together that you might think should have?
In both cases "literatura" and "sztuka" should cluster together, but don't.

Is normalization necessary?
Normalization separates "literatura" from words meaning people producing better clusters.

2 Part 2

a)

What clusters together in 2-dimensional embedding space?
Words are more spread apart. "sztuka" and "artystyczny" are closest.

What doesn't cluster together that you might think should have?
I'm suprised that words describing people are far apart.

How is the plot different from the one generated earlier from the co-occurrence matrix?
In general words are more separated. The dependencies between words seem more complex.

b)

Please state the polysemous word you discover and the multiple meanings that occur in the top 10.
Word - 'blok'
top 10 - 'barak', 'budynek', 'kolumna', 'sześcian', 'bunkier', 'sektor', 'platforma', 'prostopadłościan', 'segment', 'panel'

Why do you think many of the polysemous words you tried didn't work?
My guess is that algorithm tries to give a word a single meaning, because if word is close to 2 words that have different meanings then they have to be close to each other.

c)

Synonyms duży, tegi have cosine distance: 0.6592637896537781
Antonyms duży, mały have cosine distance: 0.25460952520370483

Once you have found your example, please give a possible explanation for why this counter-intuitive result may have happened.

Antonyms are used in same contexts, so it makes sense that their distance is small.
'tegi' is a rare word so its representation may be poor.
Additionaly it is probably used with other complicated words, whereas 'duży' is used in simple sentences.

d)

top 10 similar to 'mało', 'wyżej', dissimilar to 'niżej'
('dużo', 0.5422331690788269),
('wiele', 0.5120387673377991),
('naprawde', 0.5089786648750305),
('niewiele', 0.4938616454601288),
('zdrosny', 0.4818892478942871),
('chyba', 0.47972527146339417),
('mniemanie', 0.466522753238678),
('rzeczywiście', 0.45115557312965393),
('tyle', 0.4444846510887146),
('troskać', 0.4437904953956604)

e)

Incorrect analogy

ptak:latać :: ryba:pływać

top 10 similar to 'ryba', 'latać', dissimilar to 'ptak'
('miesiac', 0.7154368162155151),
('lato', 0.6962296366691589),
('tydzień', 0.6849808096885681),
('doba', 0.5967993140220642),
('dekada', 0.5900819897651672),
('dzień', 0.584226667881012),
('godzina', 0.5746790766716003),
('dziesieciolecie', 0.572984516620636),
('stulecie', 0.559540867805481),
('rok', 0.5404884219169617)

f)

Results are weird because code uses 'mezczyzna',
which is a form of 'meżczyzna' without polish letters.
With 'meżczyzna' instead top result is 'szefowa'.
I guess 'mezczyzna' word comes form dataset that rarely uses 'szefowa' word.

top 10 similar to 'kobieta', 'szef', dissimilar to 'mezczyzna'
('własika', 0.5678122639656067),
('agent', 0.5483713150024414),
('oficer', 0.5411549210548401),
('esperów', 0.5383270978927612),
('interpol', 0.5367037653923035),
('antyterrorystyczny', 0.5327680110931396),
('komisarz', 0.5326411128044128),
('europolu', 0.5274547338485718),
('bnd', 0.5271410346031189),
('pracownik', 0.5215375423431396)

top 10 similar to 'mezczyzna', 'prezes', dissimilar to 'kobieta'
('wiceprezes', 0.6396454572677612),
('czlonkiem', 0.5929950475692749),
('przewodniczacy', 0.5746127963066101),
('czlonek', 0.5648552179336548),
('przewodniczacym', 0.5586849451065063),
('wiceprzewodniczacy', 0.5560489892959595),

('obowiazków', 0.5549101233482361),
('obowiazani', 0.5544129610061646),
('dyrektor', 0.5513691306114197),
('obowiazany', 0.5471130609512329)

g)

top 10 similar to 'mężczyzna', dissimilar to 'kobieta'
('zaatakowac', 0.5070675611495972),
('dokonczyl', 0.5051069259643555),
('bagaz', 0.49883756041526794),
('niebezpieczenstwa', 0.49833419919013977),
('zmierzyc', 0.497328519821167),
('niedermayr', 0.4963065981864929),
('oszolomiony', 0.4955935776233673),
('wrzeszczec', 0.4900546371936798),
('przestancie', 0.4895210862159729),
('odwazny', 0.4894097149372101)

top 10 similar to 'kobieta', dissimilar to 'mężczyzna'
('dziewcze', 0.5510122776031494),
('dziewczyna', 0.5392330884933472),
('dziewczynka', 0.5020133852958679),
('mulatka', 0.4649435877799988),
('kobiecy', 0.4627985954284668),
('dziecko', 0.4600450098514557),
('niemowle', 0.44795161485671997),
('osoba', 0.4446863532066345),
('dzieciecy', 0.43624842166900635),
('młodzież', 0.43331262469291687)

Words similar to 'mężczyzna' are actions or character traits
Words similar to 'kobieta' are describing sex, age, or skin color,
so they are related to looks.
Interestingly results are significantly different when we change 'mężczyzna' to 'meżczyzna'
and moderately different when we change it to 'mezczyzna'.

h)

What might be the cause of these biases in the word vectors?
Vectors are biased because dataset is biased.
In my opinion bias may be a result of
- word being rare.
- dataset being old, or not representing real world.

- world being biased.

3 Part 3

dataset from here
https://www.kaggle.com/datasets/leadbest/googlenewsvectorsnegative300?resource=download

b)

Please state the polysemous word you discover and
the multiple meanings that occur in the top 10.
Word - 'block'
top 10 - 'blocks', 'Exume_tried', 'Kedvale_Avenue', 'Terroristic_threat', 'Hoyne_Avenue', 'Goguac_Street', 'Briarwood_Drive', 'Boxwood_Drive', 'Avers_Avenue', 'Woodale_Avenue'
Top words are related to one meaning - street.
there are also odd results like Exume_tried or Terroristic_threat.

c)

Synonyms big, colossal have cosine distance: 0.49641215801239014
Antonyms big, small have cosine distance: 0.5041321516036987
Here synonyms have smaller cosine distance.
Reasons can be either 'colossal' being more frequently used than 'tegi' or dataset being larger.

d)

top 10 -
[('bit', 0.7062703967094421),
('much', 0.5927826166152954),
('wee_bit', 0.5753095149993896),
('lot', 0.5397298336029053),
("everything'sa", 0.5307227373123169),
('teeny_bit', 0.5167810320854187),
('maybe', 0.5110480785369873),
('smidgen', 0.49733006954193115),
('litte', 0.49377959966659546),
('scant', 0.49189597368240356)]

Much is second.
English has more commonly used synonyms of 'little', than Poilsh has of 'mało'.

e)

bird:fly :: fish:swim
top 10 similar to 'fish', 'fly', dissimilar to 'bird'

[('longline_vessels', 0.48074957728385925),
('fishing', 0.4567126929759979),
('fished', 0.44685736298561096),
('fishes', 0.43080002069473267),
('tuna', 0.4262666702270508),
('shad_darts', 0.4228907823562622),
('overfish', 0.4222323000431061),
('mackerel', 0.41814762353897095),
('saithe', 0.4176707863807678),
('mangrove_snappers', 0.41631370782852173)]

No 'swim' in top 10.
My analogy was confusing enough.

f)

No bias observed.
top 10 similar to 'woman', 'boss', dissimilar to 'man'

[('bosses', 0.5522644519805908),
('manageress', 0.49151360988616943),
('exec', 0.45940810441970825),
('Manageress', 0.4559843838214874),
('receptionist', 0.4474116563796997),
('Jane_Danson', 0.44480547308921814),
('Fiz_Jennie_McAlpine', 0.4427576959133148),
('Coronation_Street_actress', 0.44275563955307007),
('supremo', 0.4409853219985962),
('coworker', 0.43986251950263977)]
top 10 similar to 'man', 'president', dissimilar to 'woman'

[('President', 0.6693714261054993),
('chairman', 0.6291338801383972),
('chief_executive', 0.5799599885940552),
('CEO', 0.5605502128601074),
('pesident', 0.5466867089271545),
('Chairman', 0.5464242696762085),
('vice_president', 0.537648737430572 5),
('prez', 0.5215684175491333),
('Presdient', 0.5107682347297668),

('presient', 0.498538076877594)]

g)

top 10 similar to 'man', dissimilar to 'woman'
[('Shaun_Maloney_Aiden_McGeady', 0.35027220845222473),
('tactically_adept', 0.3487197160720825),
('Matt_Bramald', 0.3400961458683014),
('strongside_LB', 0.337636798620224),
('newboy', 0.33329278230667114),
('Philip_Boampong', 0.33152341842651367),
('joker', 0.3312978446483612),
('superpest', 0.3302587866783142),
('TRENDING_UP', 0.3300756514072418),
('Felipe_Claybrooks', 0.3289523720741272)]
top 10 similar to 'woman', dissimilar to 'man'
[('she', 0.45412716269493103),
('her', 0.39712801575660706),
('Certified_Nurse_Midwife', 0.3824717402458191),
('Ms.', 0.37514764070510864),
('silicone_gel_implant', 0.3704040050506592),
('girlhood', 0.37001779675483704),
('nurse_midwife', 0.369699090719223),
('undergo_hysterectomy', 0.36893028020858765),
('silicone_breast_implants', 0.3683786392211914),
('breastfeeds', 0.36699435114860535)]

Words similar to 'men' are about achiving success.
Words similar to 'woman' are about motherhood.