# NLP course Assignment 1: Distributional Similarity

Daniel bazar 314708181

Lior krengel XXXXXXXXX

## Dense vectors (word2vec based similarities)

In this section of the assignment, we explored word-similarities induced by a word-embedding algorithm. We used the ‘*word2vec-google-news-300’* pre-trained vectors, from the *gensim* python package.

we exploreded word similarities in several manners as detailed below:

### Generating lists of the most similar words

We chose 5 words out of the vocabulary, and for each of them, we generated list of the 20 most similar words according to word2vec algorithm. Later on the second part of the assignment we will analyze these lists in comparison to ChatGPT output.

**Our 5 words are: espresso, game, spy, car, smartphone.**

Results:

|  |  |
| --- | --- |
| **word** | **20 most similar words (decreasing order)** |
| espresso | cappuccino, mocha, coffee, latte, caramel\_macchiato, ristretto, espressos, macchiato, chai\_latte, espresso\_cappuccino, caramel\_latte, vanilla\_latte, brewed\_coffee, iced\_coffee, nonfat\_latte, expresso, espresso\_latte, coffees, mocha\_latte, Lavazza\_coffee |
| game | games, play, match, matchup, agame, ballgame, thegame, opener, matches, tournament, playing, league, Game, scrimmages, fourgame, scrimmage, postseason, playoffs, gme, season |
| spy | spies, spying, espionage, spymaster, CIA, Spy, MI6, spymasters, intelligence, CIA\_operative, eavesdropping, covert, persecute\_dissidents, counterspy, counterintelligence, supersecret, counterspies, KGB, honeytrap, superspy |
| car | vehicle, cars, SUV, minivan, truck, Car, Ford\_Focus, Honda\_Civic, Jeep, pickup\_truck, Toyota\_Camry, scooter, Honda\_Accord, sedan, Toyota\_Corolla, motorcycle, Nissan\_Altima, Ford\_Explorer, Ford\_Escort, Ford\_Mustang |
| smartphone | smartphones, Smartphone, handset, Android\_smartphone, smart\_phones, Android\_phones, Android, Android\_smartphones, netbook, Android\_OS, touchscreen\_smartphone, iPhone, handsets, Smartphones, smartphone\_OS, Apple\_AAPL\_iPhone, Android\_handsets, Symbian\_smartphone, tablet\_PC, smarphone |

### Polysemous Words

Polysemous words are words that have several meanings.  
we asked to find three polysemous such that the top-10 neighbours of each word reflect both word meanings (Group 1), and three polysemous words such that the top-10 neighbours of each word reflect only a single meaning (Group 2).  
we actually found relatively a lot of polysemous words but most of them were belong to group 2. We tried to think about attributes that associate word to group 1 like the sense frequency (maybe if one sense is significantly more frequent, the word vectors might be biased towards that dominant sense) or sense context (for example ‘bank’ represent two meanings but these are both places, in contrast to bass that reflects completely different meanings) but still, most of the words were belong to group 2. So, at the end it was somewhat trial and error.  
in our opinion, the sense frequency (mention above) is reasonable possible explanation for why the second group words neighbours reflect only one sense.   
~~In addition (and from here on these are just assumptions because we don’t know the training settings of the model), both training data and training settings have an influence on the embedding. For example, some settings would prefer morphologically similar words (like plurals, cat->cats) and other settings would consider semantically relevant words (like cat->dog). Different settings\algorithms can also differ in sensitivity to rare and frequent words and that connect to our first explanation.~~

**Group 1**

|  |  |  |
| --- | --- | --- |
| **word** | **possible senses** | **top-10 neighbours** |
| mole | small burrowing mammal, a common type of skin growth, spy, unit in chemistry | moles, pollo\_en, freckle, cancerous\_mole, birthmark, unibrow, spies, codenamed\_Stakeknife, nodule, pube |
| bass | type of fish, low-pitched musical instrument, low voice | crappie, largemouth, largemouths, largemouth\_bass, striper, stripers, smallmouth, Spotted\_bass, acoustic\_bass, upright\_bass |
| fall | season (also known as autumn), to drop or descend | falling, falls, drop, tumble, rise, plummet, fell, spring, Fall, sag |

Reflected sense are represented by colors

**Group 2**

|  |  |  |
| --- | --- | --- |
| **word** | **possible senses** | **top-10 neighbours** |
| rock | A relatively hard naturally occurring mineral material, music genre | rock\_n\_roll, rockers, punk\_emo, punk\_rock, alt\_rock, station\_WHJY\_FM, rocks, indie\_rock, star\_Gustavo\_Cerati, rock'n'roll |
| bar | drinking establishment, a straight piece (as of wood or metal) that is longer than it is wide (beam\rod) | Bar, bars, tavern, pub, nightspot, nightclub, Pub, bartender, restaurant, Lounge |
| plant | botanical organism (seed), factory | plants, Plant, factory, paperboard\_mill, containerboard\_mill, factories, megawatt\_MW\_biomass, refinery, Plants, mill |

Reflected sense are represented by colors

### Synonyms and Antonyms

We asked to find a triplet of words (w1, w2, w3) such that all of the following conditions hold:

1. w1 and w2 are synonyms or almost synonyms.
2. w1 and w3 are antonyms.
3. sim(w1, w2) < sim(w1, w3)

**our words that meet all the conditions are: love, like, hate (w1, w2, w3 respectively)**

This behavior in which the antonyms are more similar than the synonyms can be explained by that in word embeddings, words that are often used in similar contexts are closer. In our example, "love" and "hate" only appear in the context of "feelings" while "like" can appear in context of "similarity" as well.

### The Effect of Different Corpora

In this part we compared two models based on two sources Wikipedia and twitter.

We asked to find 5 words whose top 10 neighbors based on the Wikipedia corpus are very similar to their top 10 neighbors based on the twitter corpus. And vice versa, 5 words whose top 10 neighbors based on the news corpus are substantially different from the top 10 neighbors based on the twitter corpus.our strategy to find those words was thinking about words that in the “twitter world” have different meanings. As we mentioned before, the similarities between words depend on if they appear in the same context so we search for words that in twitter will be in different context.

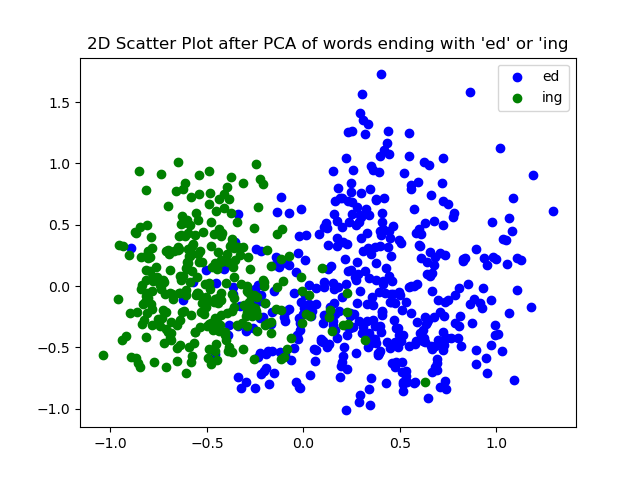
**Cross corpus similarity**

|  |  |  |
| --- | --- | --- |
| **Word** | **Wikipedia top-10 neighbours** | **Twitter top-10 neighbours** |
| yellow | red, pink, purple, blue, green, bright, colored, orange, black, colour | blue, purple, red, green, pink, white, orange, black, colored, bright |
| morning | afternoon, evening, friday, monday, thursday, night, tuesday, sunday, saturday, wednesday | afternoon, night, good, sunday, day, evening, monday, today, early, mornin |
| dog | dogs, cat, pet, puppy, horse, animal, cats, wolf, hound, pets | dogs, cat, puppy, pet, cats, horse, animal, kitten, little, kid |
| car | cars, vehicle, driver, truck, driving, vehicles, motorcycle, parked, drivers, bus | truck, cars, driving, drive, front, vehicle, bus, bike, cause, house |
| coffee | tea, cocoa, beans, espresso, drinks, drink, beer, wine, starbucks, sugar | tea, starbucks, coffe, beer, drink, iced, breakfast, milk, wine, latte |

**Cross corpus difference**

|  |  |  |  |
| --- | --- | --- | --- |
| **Word** | **Wikipedia top-10 neighbours** | **Twitter top-10 neighbours** | **difference description** |
| umbrella | organisation, organization, organisations, organizations, groups, alliance, grouping, group, non-governmental, non-profit | umbrellas, coat, rain, jacket, raining, parasol, bag, rains, outside, backpack | Wikipedia meaning: association of institutions who work together  Twitter meaning: protection against the rain |
| troll | trolls, monster, witch, doll, ape, mermaid, tentacled, sorcerer, monsters, medusa | trolls, trolling, noob, hacker, fail, bully, pervert, nerd, hack, trolled | Wikipedia meaning: a mythical creature  Twitter meaning: to antagonize others online |
| profile | recent, prominent, profiles, high-profile, attention, publicized, celebrity, political, similar, focus | page, picture, check, website, view, account, pic, click, add, visit | Wikipedia meaning: level of public exposure twitter meaning: user profile in social media |
| mute | deaf, blind, helpless, silent, storks, bedridden, ignorant, swans, cornetts, motionless | refresh, button, disable, delete, deaf, screen, remote, pause, block, muting | Wikipedia meaning: person who have speech disorder  Twitter meaning: function to eliminate sound in apps |
| gaming | gambling, casino, casinos, entertainment, poker, interactive, gamers, bingo, online, multiplayer | xbox, console, playstation, gamer, gamers, videogames, tech, desktop, controller, nintendo | Wikipedia meaning: playing gambling  Twitter meaning:  playing video games |

### Plotting words in 2D - Dimensionality Reduction

In this part we took the first 5000 words in the google model vocabulary, keeping just the words ending with “ed” and “ing” (verbs) and implemented a dimensionality reduction using PCA to transform the 300-dim matrix to 2-d so we can plot it:

points that correspond to words that end with “ed” are colored in blue and points that correspond to words that end with “ing” in green.

As we can see the two classes are slightly separate, indicates that the model might distinguished tense in words. However, there are quite a few overlapping points. We tried to analyze the controversial points using KNN by looking at points that all their closest neighbors are from different class but it ain’t provided good enough distinction, seems like the words lost most of their meaning. We looked at the percentage of variance explained by each component and we observed that they actually explained very few of it. The variance explained by the first component is around 4.2% and the variance explained by the second component is around 3.4% so the overall percentage of variance explained is just around 7.6%! so by considering only those two components we barely preserve the essential characteristics of the original data. Nevertheless, it could be useful (although, not successful) in terms of classification which suffix the word have.