

NLP course Assignment 1: Distributional Similarity

Daniel Bazar 314708181

Lior Krenkel 315850594

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 explored 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 a 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, Computer, run.

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, espresso, 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
Computer	Computers, computer, Computing, Enterprises_GoGrid_Gomez, computers, Electronics, Genuitec_GizmoX_Glassbox_Global, COMPUTER, Modules_COMs, Information_Technology, Microcomputer, Applied_Computing, Lab_International_Concensus, Marvell_Plug, Microcomputers, Digital_Forensics, PC, Malicious_Attacks, Software, Laptops
run	runs, running, drive, ran, scamper, tworun_double, go, twoout, walk, Mark_Grudzielanek_singled, Batterymate_Miguel_Olivo, homerun, threerun, Collin_Salzenstein, basesloaded, fielder's_choice_grounder, Peter_Bourjos_tripled, Casey_Kalenkosky, Scutaro_singled, clubbed_solo_homer

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 (w_1, w_2, w_3) such that all the following conditions hold:

- a) w_1 and w_2 are synonyms or almost synonyms.
- b) w_1 and w_3 are antonyms.
- c) $\text{sim}(w_1, w_2) < \text{sim}(w_1, w_3)$

our words that meet all the conditions are: love, like, hate (w_1, w_2, w_3 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

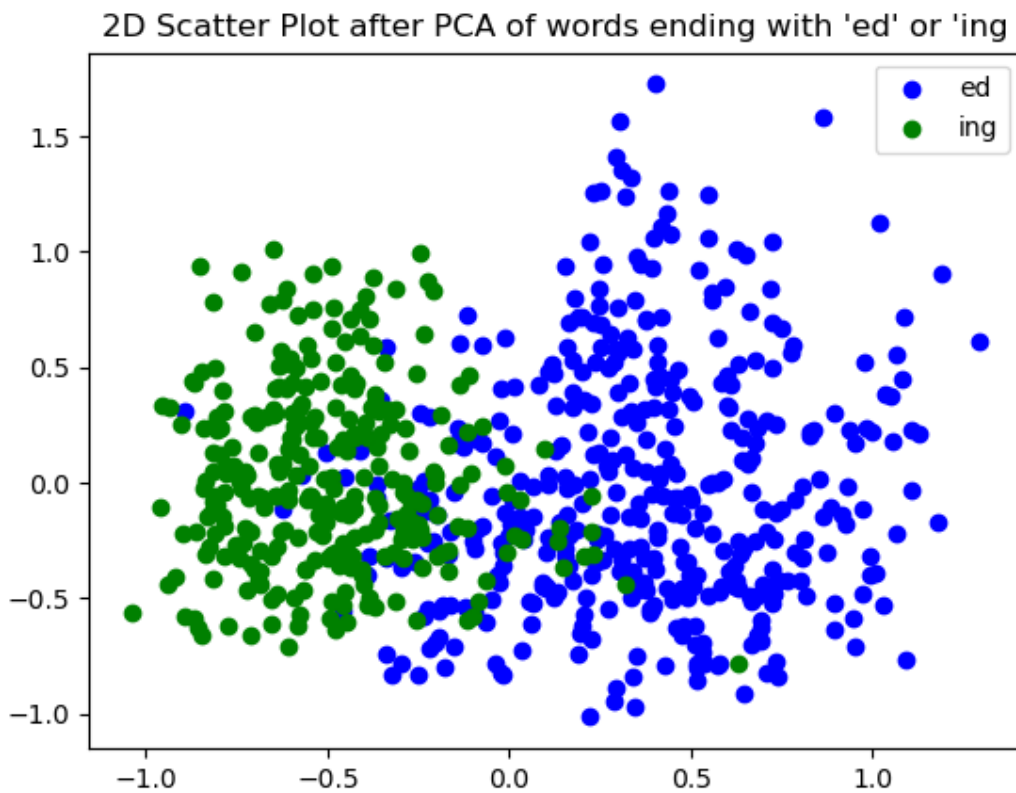
Word	Wikipedia top-10 neighbours	Twitter top-10 neighbours
yellow	red, pink, purple, blue, green, bright, colored, orange, black, colour	blue, purple, red, green, pink, white, orange, black, colored, bright
morning	afternoon, evening, friday, monday, thursday, night, tuesday, sunday, saturday, wednesday	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, tentacted, 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 very successful) in terms of classification which suffix the word have.

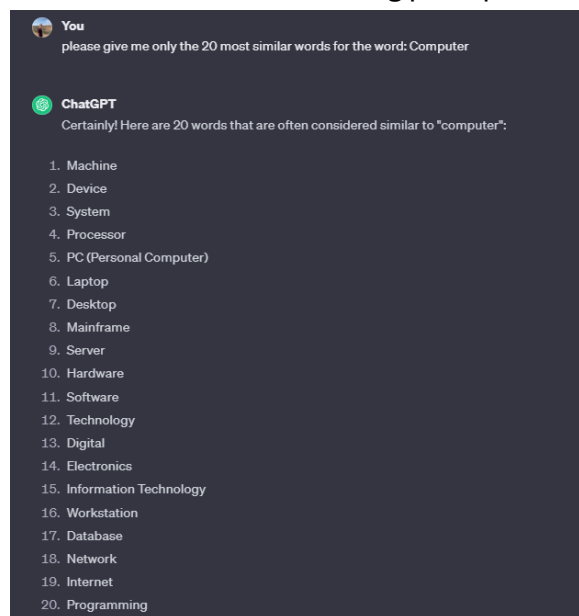
Word-similarities in Large Language Model

Related words

In this part we compared word2vec and ChatGPT word similarity result.

For each of the 5 words for which we generated the similarity lists in the word2vec part above (Computer, espresso, spy, run, game), we asked ChatGPT to produce a list of the 20 most similar words by itself. We did it word by word instead of asking for all words in the same prompt.

we started with the following prompt:



As we can see, on the first try, for the word Computer, we got exactly 20 similar words, but we also got some phrases. So, we asked him again to give us similar words but this time without phrases and he succeeded.



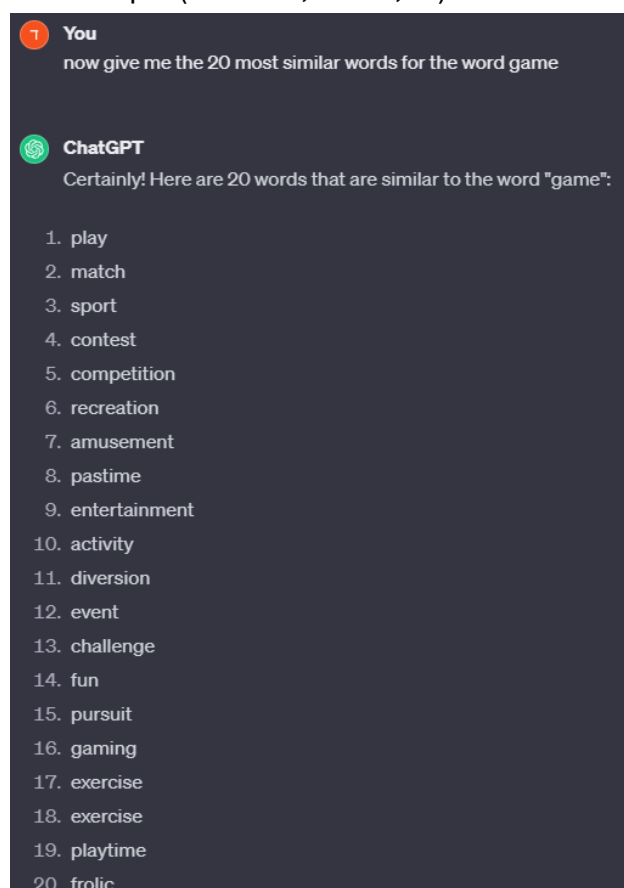
We can see that for acronym like PC (no.5), he just removed the interpretation when phrases like information technology (no.15) he just threw it away. Apart from those changes, the list is pretty the same.

We noticed that after asking from ChatGPT for ‘no phrases’ for the first time, for the rest of the words we didn’t receive any phrases. For example:

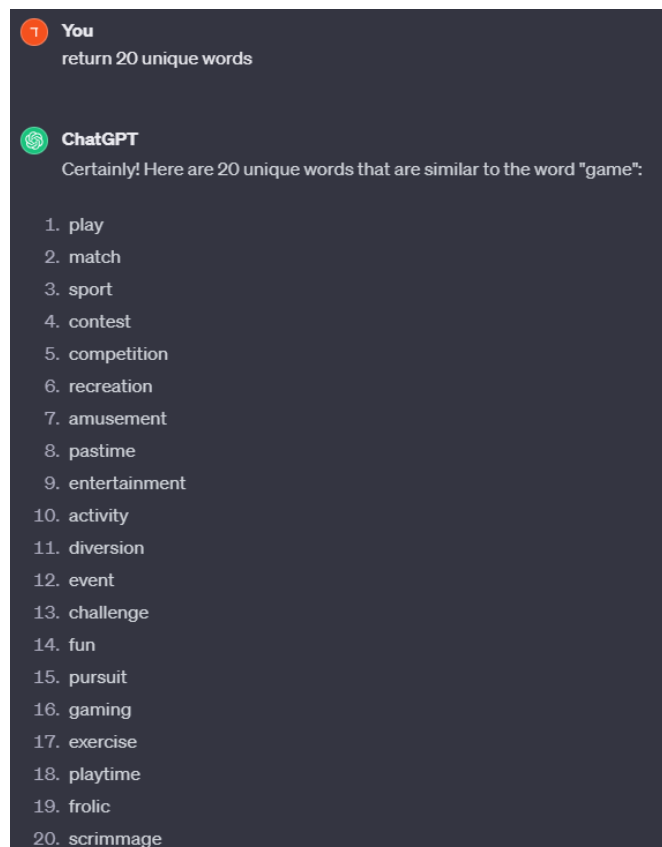


From this word we kept asking the same question and got no phrases.

One more thing we noticed that ChatGPT sometimes produced duplicates words. for example (exercise, no.17,18):



So to handle this we asked him to return unique words and it worked:



Both models give pretty accurate results for the similar words but have different approaches for the similarity. It seems like word2vec similar words are more morphologically related words (the first words are always inflections of the query like tense, plurals, misspelling, capitals, etc.) while ChatGPT provides more semantically related words like synonyms.

In terms of diversity, on the one hand word2vec is lacking because it returns inflections of the word and because it's based on co-occurrences of words and as a result he can return different words but in the same "type" (like a lot of runners names when query with run). On the other hand, ChatGPT sometimes produces duplicates words while with word2vec we always get words once (but it could be handle easily with another prompt).

Another difference between the models is when asking for a word with initial cap, ChatGPT isn't sensitive to it and refers to this word as the same word as lower caps, while for word2vec it gives different similar words for each one.

(produced lists below)

Word	word2vec 20 most similar words (decreasing order)	ChatGPT 20 most similar words (decreasing order)	ChatGPT 20 most similar words without phrases (decreasing order)
Computer	Computers, computer, Computing, Enterprises_GoGrid_Gomez, computers, Electronics, Genuitec_GizmoX_Glassbox_Global, COMPUTER, Modules_COMs, Information_Technology, Microcomputer, Applied_Computing, Lab_International_Concensus, Marvell_Plug, Microcomputers, Digital_Forensics, PC, Malicious_Attacks, Software, Laptops	Machine, Device, System, Processor, PC (Personal Computer), Laptop, Desktop, Mainframe, Server, Hardware, Software, Technology, Digital, Electronics, Information Technology, Workstation, Database, Network, Internet, Programming	Machine, Device, System, Processor, PC, Laptop, Desktop, Mainframe, Server, Hardware, Software, Technology, Digital, Electronics, Workstation, Database, Network, Internet, Programming, Coding
espresso	cappuccino, mocha, coffee, latte, caramel_macchiato, ristretto, espressos, macchiato, chai_latte, espresso_cappuccino, caramel_latte, vanilla_latte, brewed_coffee, iced_coffee, nonfat_latte, espresso, espresso_latte, coffees, mocha_latte, Lavazza_coffee	Coffee, Brew, Shot, Beverage, Caffeine, Roast, Italian, Java, Cup, Barista, Crema, Aroma, Beans, Mocha, Latte, Americano, Macchiato, Cappuccino, Bar, Sip	<no phrases>
spy	spies, spying, espionage, spymaster, CIA, Spy, MI6, spymasters, intelligence, CIA_operative, eavesdropping, covert, persecute_dissidents, counterspy, counterintelligence, supersecret, counterspies, KGB, honeytrap, superspy	Agent, Operative, Infiltrator, Mole, Sleuth, Spook, Informant, Saboteur, Covert, Espionage, Surveillance, Undercover, Intelligence, Detective, Reconnaissance, Secret, Cryptanalyst, Infiltration, Stealth, Clandestine	<no phrases>
run	runs, running, drive, ran, scamper, tworun_double, go, twoout, walk, Mark_Grudzielanek_singled, Batterymate_Miguel_Olivo, homerun, threerun, Collin_Salzenstein, basesloaded, fielder's_choice_grounder, Peter_Bourjos_tripled, Casey_Kalenkosky, Scutaro_singled, clubbed_solo_homer	jog, sprint, dash, race, trot, gallop, stride, hurry, bolt, scamper, dart, canter, amble, lope, speed, pace, hasten, charge, skip, jogtrot	<no phrases>
game	games, play, match, matchup, agame, ballgame, thegame, opener, matches, tournament, playing, league, Game, scrimmages, fourgame, scrimmage, postseason, playoffs, gme, season	play, match, sport, contest, competition, recreation, amusement, pastime, entertainment, activity, diversion, event, challenge, fun, pursuit, gaming, exercise, playtime, frolic, scrimmage	<no phrases>

Now we selected two of the words (run and espresso) and increased the number of neighbors from 20 to 100:

Word	word2vec 100 most similar words (decreasing order)	ChatGPT 100 most similar words (decreasing order)
run	<p>runs, running, drive, ran, scamper, tworun_double, go, twoout, walk, Mark_Grudzielanek_singled, Batterymate_Miguel_Olivo, homerun, threerun, Collin_Salzenstein, basesloaded, fielder's_choice_grounder, Peter_Bourjos_tripled, Casey_Kalenkosky, Scutaro_singled, clubbed_solo_homer, Reliever_Macay_McBride, solo_round_tripper, Juan_Uribe_sacrifice_fly, Melisa_Koutz, Amonite, Geoff_Blum_sacrifice_fly, roundtripper, ribbie, blooping_single, Gelalich, Jonathan_Lucroy_singled, Alec_Lowrey, DeSico, Earnest_Rhone, Bucky_Aona, Felix_Fanaselle, Jed_Lowrie_sacrifice_fly, blooped_leadoff, Miguel_Cabrera_belted, Andy_Schutzenhofer, Joey_Swatfager, threerun_homer, Edwin_Encarnacion_sacrifice_fly, suicide_squeeze_bunts, Stack_Babich, Sean_Gusrang, Daniel_Nottebart, Baron_Batch_tacked, Bo_Cogbill, leftcenter, Omar_Infante_sacrifice_fly, Nate_Rolison, sacrifice_fly, leadoff_triple, Jose_Tabata_singled, Christina_Capobianco, reliever_Manny_Parra, Cameron_Seitzer, Ashley_Razey, Micheal_Dabbs, sac_fly, Ronnie_Belliard_singled, Austin_Kearns_grounded, Gemoll, LaNinfa, Rich_Aurilia_singled, clubbed_grand_slam, Kenton_Gedwed, Ronny_Paulino_singled, Grudzielanek_singled, Russell_Branyan_belted, Brian_Shouse_relieved, Jorge_Cantu_sacrifice_fly, Ryan_Uphouse, Wigginton_flied, trot, Span_flied, Markakis_homered, Matt_Untiet, Ted_Zentek, Deivi_Cruz_homered, solo_homer, bullpen_imploded, Yadier_Molina_grounded, sacrifice_fly_plated, Ellsbury_looped, Matt_Wieters_sacrifice_fly, Melvin_Mora_bunted, Jamie_Keltz, Blake_Hurlbutt, Batkoski, sacrifice_fly, Matt_Repec, Slumping_slugger_Carlos_Pena, Hawpe_homered, Hatteberg_singled,</p>	<p>sprint, jog, dash, race, trot, gallop, stride, hurry, bolt, scamper, dart, canter, amble, lope, speed, pace, hasten, charge, skip, jogtrot, tear, flee, escape, rush, galumph, lunge, scurry, trudge, march, saunter, meander, hike, plod, stampede, fly, bound, vault, hop, jump, propel, shoot, scuttle, accelerate, expedite, navigate, zip, cruise, whisk, galvanize, surge, slalom, glide, whisk, bomb, streak, hurtle, whiz, breeze, whirl, circulate, flow, stream, cascade, course, sail, zoom, scoot, swoop, scud, skim, soar, skitter, slide, wander, travel, navigate, slip, bolt, hasten, rush, fly, flee, skim, sail, amble, traverse, roll, breeze, sprint, lope, scamper, dash, jogtrot, trek, plod, speed, whisk, race, trip, journey</p>

	Tyler_Doke, Eugenio_Velez_infield, Casey_McGehee_singled, RBI_groundout	
espresso	Coffee, Cappuccino, Latte, Americano, Macchiato, Mocha, Cortado, Ristretto, Affogato, Breve, Doppio, Lungo, Flat White, Turkish Coffee, Café au Lait, Irish Coffee, Red Eye, Vienna Coffee, Frappuccino, Turkish Coffee, Barista, Crema, Decaf, Froth, Baristress, Brew, Espresso Machine, Shot, Tiramisu, Crema, Mocha, Moka, Italian Roast, Arabica, Robusta, Filter Coffee, French Press, Cold Brew, Pour-over, Chemex, Coffee Bean, Espresso Cup, Coffeehouse, Bar, Blend, Grind, Roast, Barista, Cupping, Siphon Coffee, French Roast, Frother, Brewmaster, Coffee Roaster, Single Origin, Espresso Shot, Coffee Press, Coffee Pot, Demitasse, Coffee Connoisseur, Crema, Espresso Bar, Coffee Filter, Coffee Scoop, Coffee Grinder, Espresso Beans, Coffee Mill, Coffee Creamer, Coffee Syrup, Coffee Spoon, Coffee Napkin, Coffee Tumbler, Coffee Stirrer, Coffee Canister, Coffee Sack, Coffee Subscription, Coffee Sampler, Coffee Tasting, Coffee Fest, Coffee Roasting, Coffee Brewing, Coffee Culture, Coffee Ritual, Coffee Trade, Coffee Origin, Coffee Plantation, Coffee Harvest, Coffee Blend, Coffee Aroma, Coffee Flavor, Coffee Terroir, Coffee Bar, Coffee Break, Coffee Date, Coffee Hour, Coffee Snack, Coffee Dessert, Coffee Pairing, Coffee Moment, Coffee Ritual	cappuccino, mocha, coffee, latte, caramel_macchiato, ristretto, espressos, macchiato, chai_latte, espresso_cappuccino, caramel_latte, vanilla_latte, brewed_coffee, iced_coffee, nonfat_latte, espresso, espresso_latte, coffees, mocha_latte, Lavazza_coffee, Knotty_Bodies, cappuccinos, espresso_machine, Caramel_Macchiato, Espresso, chai_tea_latte, Starbucks_coffee, frappuccino, latté, lattes, freshly_brewed_espresso, soy_latte, barista, Illy_coffee, smoothie, Frappuccino, decaf, mochaccino, espresso_drinks, fresh_brewed_coffee, caramel_mocha, iced_mocha, grande_mocha, skim_latte, Lavazza_espresso, cappuccino_latte, latte_mocha, java, Raspberry_Mocha, liqueur, drip_coffee, White_Chocolate_Mocha, café_latte, espresso_beverages, Iced_Mocha, espressos_lattes, capuccino, café_mocha, squeezed_juices, lattés, java_junkies, cappucino, coffees_lattes, gourmet_coffee, flavored_syrups, Espressos, brewed_coffees, venti, caffe_latte, freshly_brewed_coffee, o_joe, frappuccinos, au_lait, iced_coffees, espresso_beans, frothed_milk, iced_latte, gelato, raspberry_mocha, cappuccinos_lattes, squeezed_orange_juice, macchiatos, mocha_frappuccino, mango_smoothie, flavored_lattes, barristas, Iced_coffee, vanilla_lattes, mocha_lattes, frappucinos, lattes_mochas, Caffe_Mocha, Chai_latte, venti_mocha, crema, chocolate_martini, caramel_macchiatos, cappucinos, caffe_mocha, java_jolt

We can see some trends in the data:

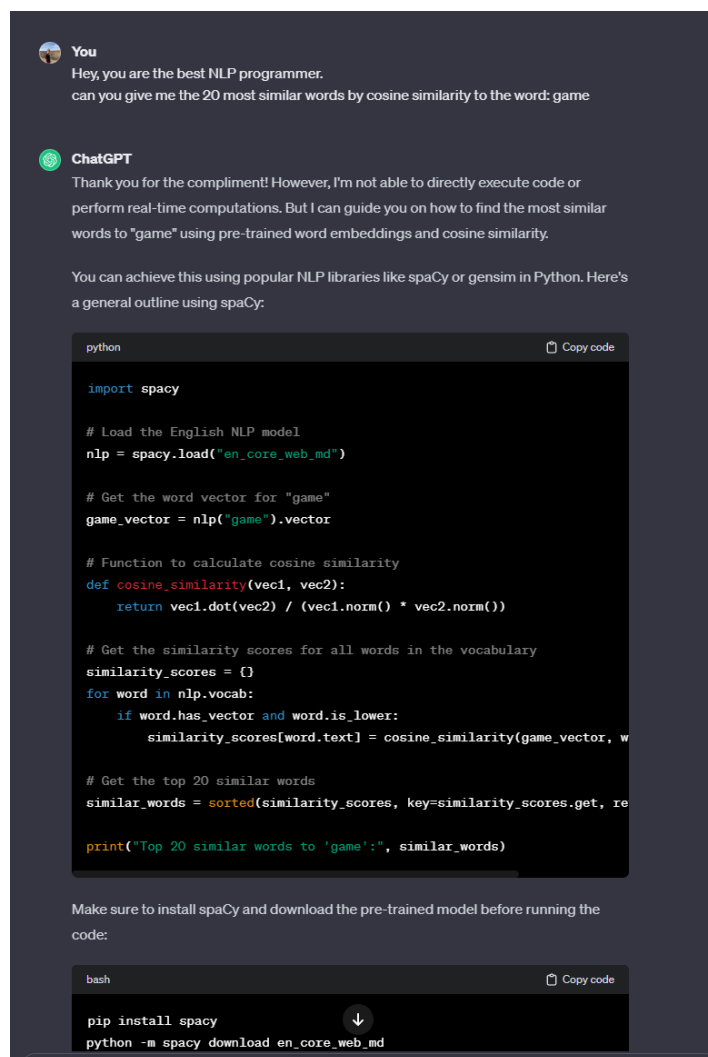
For the word “run” we can say that ChatGPT after 30 words starts repeating the same word multiple times while word2vec starts examples of runners (with 100 neighbors, even after asking multiple times in different ways to return only unique words ChatGPT still return duplicates).

For the word “espresso” we do see a similar trend that shows multiple types of coffee for both models.

Important to note that even after 100 words both models still produced words in the same topic.

While working with ChatGPT we had to do make some adjustments with our prompts.

for example, we tried to encourage him to give us good results, so we told him that he is the best programmer and asked for the most similar words by cosine similarity. The result was a code to do so instead of words. We also saw earlier how we had to ask it explicitly to remove phrases and not to duplicates words and so on.



You
Hey, you are the best NLP programmer.
can you give me the 20 most similar words by cosine similarity to the word: game

ChatGPT
Thank you for the compliment! However, I'm not able to directly execute code or perform real-time computations. But I can guide you on how to find the most similar words to "game" using pre-trained word embeddings and cosine similarity.

You can achieve this using popular NLP libraries like spaCy or gensim in Python. Here's a general outline using spaCy:

```
python
import spacy

# Load the English NLP model
nlp = spacy.load("en_core_web_md")

# Get the word vector for "game"
game_vector = nlp("game").vector

# Function to calculate cosine similarity
def cosine_similarity(vec1, vec2):
    return vec1.dot(vec2) / (vec1.norm() * vec2.norm())

# Get the similarity scores for all words in the vocabulary
similarity_scores = {}
for word in nlp.vocab:
    if word.has_vector and word.is_lower:
        similarity_scores[word.text] = cosine_similarity(game_vector, word.vector)

# Get the top 20 similar words
similar_words = sorted(similarity_scores, key=similarity_scores.get, reverse=True)

print("Top 20 similar words to 'game':", similar_words)
```

Make sure to install spaCy and download the pre-trained model before running the code:

```
bash
pip install spacy
python -m spacy download en_core_web_md
```

Synonyms and Antonyms

ChatGPT easily produced us multiple examples of Synonyms and Antonyms words with simple instructions of “can you give me 5 synonyms and 5 antonyms words for the words: happy, big, smart, beautiful, fast”.

Word	Synonyms	Antonyms
Happy	Joyful, Content, Pleased, Delighted, Merry	Sad, Unhappy, Miserable, Discontent, Gloomy
Big	Large, Huge, Enormous, Massive, Gigantic	Small, Tiny, Little, Miniature, Diminutive
Smart	Intelligent, Clever, Bright, Sharp, Wise	Stupid, Dull, Unintelligent, Dim-witted, Slow
Beautiful	Attractive, Gorgeous, Lovely, Stunning, Charming	Ugly, Plain, Unattractive, Homely, Repulsive
Fast	Quick, Speedy, Rapid, Swift, Fleet	Slow, Sluggish, Gradual, Leisurely, Languid



You

can you give me 5 synonyms and 5 antonyms words for the words: happy, big, smart, beautiful, fast



ChatGPT

Certainly! Here are 5 synonyms and 5 antonyms for each of the given words:

1. Happy:

- Synonyms: Joyful, Content, Pleased, Delighted, Merry
- Antonyms: Sad, Unhappy, Miserable, Discontent, Gloomy

2. Big:

- Synonyms: Large, Huge, Enormous, Massive, Gigantic
- Antonyms: Small, Tiny, Little, Miniature, Diminutive

3. Smart:

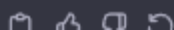
- Synonyms: Intelligent, Clever, Bright, Sharp, Wise
- Antonyms: Stupid, Dull, Unintelligent, Dim-witted, Slow

4. Beautiful:

- Synonyms: Attractive, Gorgeous, Lovely, Stunning, Charming
- Antonyms: Ugly, Plain, Unattractive, Homely, Repulsive

5. Fast:

- Synonyms: Quick, Speedy, Rapid, Swift, Fleet
- Antonyms: Slow, Sluggish, Gradual, Leisurely, Languid



Polysemy

When asked ChatGPT for similarities for the polysemous words, like word2vec, for group 1 (where each word reflects both word meanings), we got multiple meanings, and for group 2 (words such that the top-10 neighbours of each word reflect only a single meaning), we see that it indeed gave only one meaning.

Although the similar behavior, there are some differences.

In our experiment, ChatGPT is more diverse and return in one case more meanings than word2vec. In addition, for 2 out of 3 words in group 2 ChatGPT return the opposite meaning from word2vec.

Word	Word Group	possible senses	ChatGPT 10 most similar words (decreasing order)
mole	Group 1	small burrowing mammal, a common type of skin growth, spy, unit in chemistry	Burrow, Rodent, Tunneler, Dermatologist, Spy, Spot, Blemish, Nevus (a type of skin mole), Underground, Molehill
bass	Group 1	type of fish, low-pitched musical instrument, low voice	Low, Deep, Resonant, Fish, Sound, Musical, Pitch, Notes, Frequency, Double bass
fall	Group 1	season (also known as autumn), to drop or descend	Autumn, Descend, Drop, Plummet, Tumble, Decline, Collapse, Autumnal, Precipitate, Topple
rock	Group 2	A relatively hard naturally occurring mineral material, music genre	Stone, Boulder, Pebble, Cliff, Granite, Mineral, Formation, Geology, Mountain, Limestone
bar	Group 2	drinking establishment, a straight piece (as of wood or metal) that is longer than it is wide (beam\rod)	Pub, Tavern, Saloon, Lounge, Counter, Establishment, Inn, Club, Public house, Watering hole
plant	Group 2	botanical organism (seed), factory	Flora, Greenery, Botanical, Vegetation, Foliage, Herb, Shrub, Flower, Crop, Garden

Reflected sense are represented by colors

Mean Average Precision (MAP) evaluation

In this part we chose 2 of the 5 words above, spy and game. Then we generated lists of the 20 most similar words for each one by each method (word2vec and ChatGPT).

we judged the correctness of their similarities manually, by two approaches: topically related and same semantic class.

spy:

word2vec	Topically related	Same semantic class	ChatGPT	Topically related	Same semantic class
spies	1 (1)	1 (1)	Agent	1 (1)	1 (1)
spying	1 (1)	1 (1)	Operative	1 (1)	1 (1)
espionage	1 (1)	1 (1)	Infiltrator	1 (1)	1 (1)
spymaster	1 (1)	1 (1)	Mole	1 (1)	1 (1)
CIA	1 (1)	0 (0)	Sleuth	1 (1)	1 (1)
Spy	1 (1)	1 (0.833)	Spook	1 (1)	1 (1)
MI6	1 (1)	0 (0)	Informant	1 (1)	0 (0)
spymasters	1 (1)	1 (0.75)	Saboteur	1 (1)	0 (0)
intelligence	1 (1)	0 (0)	Covert	1 (1)	0 (0)
CIA_operative	1 (1)	0 (0)	Espionage	1 (1)	1 (0.7)
eavesdropping	1 (1)	0 (0)	Surveillance	1 (1)	0 (0)
covert	1 (1)	1 (0.583)	Undercover	1 (1)	1 (0.667)
persecute_dissidents	0 (0)	0 (0)	Intelligence	1 (1)	0 (0)
counterspy	1 (0.929)	1 (0.571)	Detective	1 (1)	1 (0.643)
counterintelligence	1 (0.933)	0 (0)	Reconnaissance	1 (1)	0 (0)
supersecret	1 (0.938)	0 (0)	Secret	1 (1)	0 (0)
counterspies	1 (0.941)	1 (0.529)	Cryptanalyst	1 (1)	0 (0)
KGB	1 (0.944)	0 (0)	Infiltration	1 (1)	0 (0)
honeypot	1 (0.947)	0 (0)	Stealth	1 (1)	0 (0)
superspy	1 (0.95)	1 (0.5)	Clandestine	1 (1)	0 (0)
AP (average precision)	0.98	0.78		1.00	0.89

game:

word2vec	Topically related	Same semantic class	ChatGPT	Topically related	Same semantic class
games	1 (1)	1 (1)	play	1 (1)	1 (1)
play	1 (1)	1 (1)	match	1 (1)	1 (1)
match	1 (1)	1 (1)	sport	1 (1)	0 (0)
matchup	1 (1)	0 (0)	contest	1 (1)	0 (0)
agame	1 (1)	1 (0.8)	competition	1 (1)	0 (0)
ballgame	1 (1)	1 (0.833)	recreation	1 (1)	0 (0)
thegame	1 (1)	1 (0.857)	amusement	1 (1)	0 (0)
opener	1 (1)	0 (0)	pastime	1 (1)	0 (0)
matches	1 (1)	1 (0.778)	entertainment	1 (1)	0 (0)
tournament	1 (1)	0 (0)	activity	1 (1)	0 (0)
playing	1 (1)	0 (0)	diversion	1 (1)	0 (0)
league	1 (1)	0 (0)	event	0 (0)	0 (0)
Game	1 (1)	1 (0.615)	challenge	1 (0.923)	0 (0)
scrimmages	0 (0)	0 (0)	fun	1 (0.929)	0 (0)

fourgame	1 (0.933)	1 (0.6)	pursuit	0 (0)	0 (0)
scrimmage	0 (0)	0 (0)	gaming	1 (0.875)	1 (0.188)
postseason	1 (0.882)	0 (0)	exercise	1 (0.882)	0 (0)
playoffs	1 (0.889)	0 (0)	playtime	1 (0.889)	1 (0.222)
gme	1 (0.895)	1 (0.526)	frolic	1 (0.895)	0 (0)
season	1 (0.9)	0 (0)	scrimmage	0 (0)	0 (0)
AP (average precision)	0.97	0.80		0.96	0.60

Mean Average Precision (MAP):

	Word2vec	ChatGPT
Topically related	0.98	0.98
Same semantic class	0.79	0.75

So, in our experiment, for “topically related judgment”, the two models are the same and for “same semantic class judgment” word2vec is slightly better. But it’s important to note that 2 words are not representative, and we can’t really conclude that one method is better than the other without sufficient data.