# CSE40201 - Natural Language Processing
# Assignment 1
# Distributional Word Representations

Student name: Nguyen Minh Duc
Student ID: 20202026

## 1 Experimentation of Counting and PMI

In this section, we will discuss the changes in Spearmanr's correlation coefficient for three different values of context windows size: $1, 3$, and 6. Please refer to Table 1, 2, and Figure 1 for more details.

| **w** | **men.txt** | **simlex-999.txt** |
|---|---|---|
| 1 | 0.2099 | 0.0764 |
| 3 | 0.2243 | 0.0602 |
| 6 | 0.2379 | 0.0394 |

Table 1: Distributional Counting

| **w** | **men.txt** | **simlex-999.txt** |
|---|---|---|
| 1 | 0.4653 | 0.2687 |
| 3 | 0.5340 | 0.2245 |
| 6 | 0.5423 | 0.1826 |

Table 2: Distributional Counting with Pointwise Mutual Information

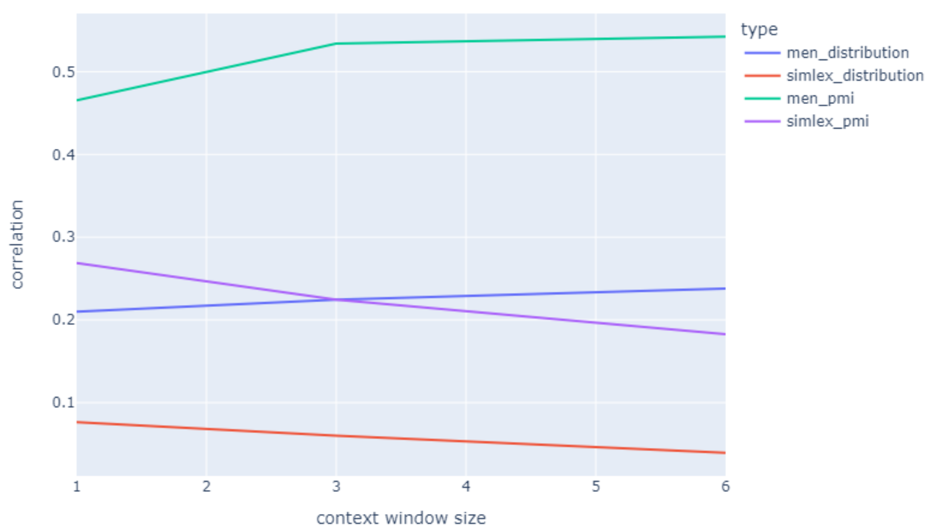


Figure 1: Correlation trend

One could immediately observe that the correlation between our score and the human-annotated score in `MEN` data set is trending upwards, while the correlations between `SimLex-999`'s score and ours are decreasing as $w$ increases. This is a very interesting behavior since the algorithm should be able to learn better as the context window increases, but it performs worse on the `SimLex-999` data set. Recall from the lecture that the `MEN` data set represents the "relatedness" between two words while the `SimLex-999` data set represents the "closeness in meaning" between two terms [1]. This suggests that Distributional Counting can learn how related the words are better than the meaning of the words. This makes sense since we are counting the appearance of surrounding words, which tend to be related to the center word. The more related two words are, the more similar their surrounding words are. For example, "apple" and "orange" are related because they are fruits (suggesting that the word "fruit" appears many times around them). However, when it comes to meaning, they are different concepts with some overlapping in their definitions. This is why "apple" and "lemon" are given a score of 4.05, an average score in `SimLex-999` where the maximum is about 9.8, while "apple" and "orange" are given of 43, a very high score in `MEN` where the maximum is 50. Another example that appears in both data sets could be "meat" and "bacon". In `MEN`, they receive a score of 44, almost the maximum, while they receive 5.8 out of 9.8 in `SimLex-999`. As $w$ increases, the center word can "see" more context words, but they are further away, hence, might not have a similar meaning to the center words anymore. Therefore, it will be harder for our algorithm to capture the similarity in meanings if the vector representation is not meaningfully stable. On the other hand, the wider the center word's horizon is, the more context it can perceive, contributing to the relatedness between it and the context since they are still related to each other in a particular sentence.

## 2 Nearest neighbors of monster

In this section, we are asked to output the 10 nearest neighbors of the word "monster" for two different context window sizes 1 and 6. The results are as follows

| **Rank** | **The word** |
|---|---|
| 1 | dragon |
| 2 | tyrant |
| 3 | creatures |
| 4 | monsters |
| 5 | jar |
| 6 | hornet |
| 7 | gangster |
| 8 | invaders |
| 9 | rhinoceros |
| 10 | robot |

Table 3: Ten nearest neighbors of monster when $w = 1$

| **Rank** | **The word** |
|---|---|
| 1 | evil |
| 2 | giant |
| 3 | creatures |
| 4 | monsters |
| 5 | godzilla |
| 6 | dragon |
| 7 | dog |
| 8 | ghost |
| 9 | horror |
| 10 | girl |

Table 4: Ten nearest neighbors of monster when $w = 6$

As we can see from Table 3 and 4, the top 1 nearest neighbors of "monster" are consistent with our desired result, which are dragon and evil for $w = 1$ and $w = 6$ respectively. We can also see that the majority of words are very similar to "monster", with some strange outliers such as "girl" and "jar".

## 3 Part-of-speech tag analysis

In this section, we will investigate whether the set of nearest neighbors produced by our algorithm maintains the word's part-of-speech tag of the query word. Doing this requires preparing a set of query words with different part-of-speech tags, inflected forms, and different context window sizes. For this exercise, I experimented with two different sizes the same as in the assignment handout, and the following set of words:

```
PART_OF_SPEECH_VERBS = ["transport", "transports", "transporting", "transported",
                        'eat', 'eats', 'ate', 'eaten', 'eating',
                        'fly', 'flies', 'flew', 'flown', 'flying']

PART_OF_SPEECH_NOUNS = ['dog', 'dogs',
                        'city', 'cities',
                        'person', 'people',
                        'leaf', 'leaves']

PART_OF_SPEECH_ADJECTIVES = ['big', 'bigger', 'biggest',
                             'good', 'better', 'best',
                             'happy', 'exceptional', 'exceptionally',
                             'far', 'farther', 'further']

PART_OF_SPEECH_PREPOSITIONS = ['as', 'in', 'on', 'at', 'of', 'to', 'with', 'up']
```

Firstly, let's analyze the behavior when $w = 1$. Please refer to Table 5 for sets of nearest neighbors of verbs, Table 6 for nouns, Table 7 for adjectives, and Table 8 for prepositions.

Let's investigate table by table. In the verb category, the word "transport" and "transports" does not keep their part-of-speech tags as their nearest neighbors are mostly nouns. However, the remaining words behave as expected since the majority of their nearest neighbors are verbs with different inflected forms. The same thing applies to nouns and adjectives as well, where nearest neighbors still keep their tags the same as the query's with a few exceptions. As for the prepositions, the results are a bit weird as they include punctuations as one of their close neighbors, nevertheless, most of the neighbors are of the same category. Thus, nearest neighbors tend to preserve the query's tag.

Secondly, let's investigate the algorithm's behavior when $w = 6$. Please refer to Table 9 for sets of nearest neighbors of verbs, Table 10 for nouns, Table 11 for adjectives, and Table 12 for prepositions.

Again, let's go through these data table by table. In the verb category, the part-of-speech tags change completely, it does not seem to preserve the query's tag anymore when $w = 6$. The nearest neighbors now consist of both verbs and nouns. The "relatedness" among them remains somewhat high since the output does indeed go with the queried word in a normal sentence. However, the difference in meaning between the query words and their nearest neighbors decreases dramatically. For example, the top 5 nearest neighbors of the word "eats" are "crustaceans, snails, eat, eating, feeds", among which three out of five words has totally different meaning from the original query words. The same is also true for the rest. In the nouns category, the nearest neighbors can still keep their query words' part-of-speech tags to some extent. More related words are added, but they tend to lose their original meanings. As for the adjectives, they could not maintain the part-of-speech tags as well as they do when $w = 1$. They still maintain the tags when the query is "bigger", and "exceptionally", but for all other adjectives, they have totally different part-of-speech tags. The prepositions even suffer from a greater loss of part-of-speech tags. On average, only about 2 words in the top 5 nearest neighbors still keep the same tag as their query word.

| query | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
|---|---|---|---|---|---|
| transport | transportation | rail | reconnaissance | services | service |
| transports | battleships | citrouillestally-ho | 1697 | aruba | histoire |
| transporting | embarking | transported | hauling | hauled | dispose |
| transported | ejected | marched | deported | shipped | transporting |
| eat | eating | infect | buy | forget | disable |
| eats | pretended | drowned | gastropod | flies | plight |
| ate | eat | harass | greet | hungry | misunderstood |
| eaten | overlooked | worshipped | consumed | worn | regarded |
| eating | eat | fried | cooked | drinking | eaten |
| fly | migrate | go | flown | wander | come |
| flies | eats | wander | bounced | fishes | pulled |
| flew | flown | interceptor | crashed | defected | combat |
| flown | pumped | shipped | refloated | interned | overthrown |
| flying | flight | flights | raf | patrol | aircraft |

Table 5: Nearest neighbors for verbs when $w = 1$

| query | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
|---|---|---|---|---|---|
| dog | dogs | cat | horse | rabbit | goat |
| dogs | dog | cats | humans | cat | animals |
| city | town | county | university | area | district |
| cities | towns | villages | settlements | provinces | countries |
| person | man | woman | persons | people | individuals |
| people | students | men | households | families | approximately |
| leaf | tree | trees | grove | yellow | plantations |
| leaves | trees | tells | wife | gets | married |

Table 6: Nearest neighbors for nouns when $w = 1$

In conclusion, the algorithm produces nearest neighbors that tend to have the same part-of-speech tag as the query word when $w$ is small, i.e., $w = 1$. However, as $w$ increases, i.e., $w = 6$, the algorithm obtains a worse understanding of the meaning of each word, but its understanding of relationships among the words is better, which makes the algorithm produce illogical sense neighbors, hence, cannot preserve part-of-speech tags, but they are still related to each other to some extent.

# 4 Words with multiple senses analysis

In this section, we are going to analyze the behavior of multiple senses words as $w$ is changing from $w = 1$ to $w = 6$. Please refer to Table 13 for $w = 1$, and Table 14 for $w = 6$ for the raw data obtained from the algorithm. In this section, I used the following set of words for analysis:

```
MULTIPLE_SENSES_WORDS=["bank", "cell", "apple", "apples", "axes", "frame", "light",
                        "well", "bat", "book", "break", "change", "date", "watch"]
```

Let's first investigate the case where $w = 1$. The nearest neighbors of our list of words seem to produce the words with the commonly known meanings of each word. For example, we usually associate the bank as a financial institution where people deposit and withdraw money, which has a similar meaning/related to the produced nearest neighbors by our algorithm. The two "apple" words in this case produce words that are all related to fruits; The word "cell" provides biological meaning; The word "book" is a literature-related concept; And the word "change" means to make something different or to move to a different place according to our vector representations. However, there are words that our algorithm successfully captures the duality in their semantics.

| query | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
|---|---|---|---|---|---|
| big | little | huge | biggest | large | great |
| bigger | larger | worse | smaller | closer | broader |
| biggest | major | largest | greatest | main | big |
| good | bad | poor | better | excellent | useful |
| better | good | worse | poorly | poor | best |
| best | top | good | american | better | canadian |
| happy | pleased | afraid | hesitant | glad | worried |
| exceptional | extraordinary | outstanding | exemplary | remarkable | impressive |
| exceptionally | extraordinarily | extremely | unusually | very | sufficiently |
| far | much | considerably | slightly | significantly | lot |
| farther | curving | kilometers | deeper | kilometres | worse |
| further | some | any | these | into | their |

Table 7: Nearest neighbors for adjectives when $w = 1$

| query | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
|---|---|---|---|---|---|
| as | is | was | by | for | with |
| in | . | at | from | february | january |
| on | at | for | as | with | in |
| at | from | on | in | ( | , |
| of | . | for | in | from | ’s |
| to | would | not | ’t | will | can |
| with | and | by | are | or | for |
| up | out | down | off | back | away |

Table 8: Nearest neighbors for prepositions when $w = 1$

For instance, the nearest neighbors of “light” contains both heavy and dark, which are related to two meanings of light. Last but not least, the representation of “well” produces confusing results, but it could still capture one meaning of it, the other five are prepositions. Therefore, with $w = 1$, our algorithm could correctly capture one meaning of such multiple senses words most of the time.

Secondly, let’s start analyzing the case where $w = 6$. The nearest neighbors of our list of words seem to be able to correctly capture the ambiguity of more words than in the previous case. For instance, it could capture both the financial and river meanings of “bank”; mass, and luminance for “light”; the tool for chopping, and mathematical meaning for “axes”. Surprisingly, there are some words that our algorithm switches to another meaning for the entire nearest neighbors. For example, “apple” is no longer a fruit but a tech company; “bat” is no longer an animal, but a tool for playing baseball. Unfortunately, our algorithm still cannot capture the meaning of “well”, maybe because it can be used as a preposition sometimes. As for the remaining words, the vector representation keeps the same meanings as in the previous case. In conclusion, with $w = 6$, our nearest neighbor can successfully explore other meanings of multiple senses words.

# References

[1] Professor Kim, Lecture 9, CSE40201 - Natural Language Processing, 2023.

| query | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
|---|---|---|---|---|---|
| transport | transportation | aircraft | rail | services | passenger |
| transports | convoy | leyte | boats | transporting | ships |
| transporting | carrying | transported | transport | supplies | cargo |
| transported | transporting | transport | prisoners | camps | supply |
| eat | eaten | eating | fish | food | insects |
| eats | crustaceans | snails | eat | eating | feeds |
| ate | eat | vegetables | mushrooms | eating | meat |
| eaten | vegetables | boiled | eat | meat | salad |
| eating | eat | meat | food | fruit | eaten |
| fly | flying | flight | aircraft | flies | plane |
| flies | beetles | larvae | fly | insects | geometridae |
| flew | flying | missions | squadron | aircraft | sorties |
| flown | aircraft | flew | bomber | squadron | flying |
| flying | aircraft | squadron | flight | air | fighter |

Table 9: Nearest neighbors for verbs when $w = 6$

| query | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
|---|---|---|---|---|---|
| dog | dogs | cat | boy | man | breed |
| dogs | cats | animals | dog | pigs | sheep |
| city | town | north | located | south | county |
| cities | towns | region | city | capital | areas |
| person | me | someone | your | subject | any |
| people | them | about | so | we | you |
| leaf | leaves | larvae | yellow | flowers | clusters |
| leaves | flowers | dark | tree | leaf | usually |

Table 10: Nearest neighbors for nouns when $w = 6$

| query | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
|---|---|---|---|---|---|
| big | show | rock | featured | black | tv |
| bigger | larger | faster | expensive | smaller | decent |
| biggest | largest | greatest | hosted | selling | hit |
| good | we | you | my | think | ’t |
| better | we | need | think | good | your |
| best | award | film | awards | music | won |
| happy | ’m | wants | someone | everyone | ’re |
| exceptional | extraordinary | outstanding | talent | skills | skill |
| exceptionally | extremely | relatively | brittle | extraordinarily | clever |
| far | much | too | even | very | less |
| farther | inland | coastline | kilometers | orbit | southwestern |
| further | should | discussion | page | talk | do |

Table 11: Nearest neighbors for adjectives when $w = 6$

| query | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
|---|---|---|---|---|---|
| as | well | such | is | ” | be |
| in | at | was | the | of | university |
| on | at | it | was | this | in |
| at | in | university | school | he | was |
| of | the | in | region | province | its |
| to | that | be | not | would | it |
| with | and | two | often | along | each |
| up | out | them | him | into | then |

Table 12: Nearest neighbors for prepositions when $w = 6$

| query | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
|---|---|---|---|---|---|
| bank | banks | insurance | company | corporation | banking |
| cell | cells | cellular | tissue | neuronal | protein |
| apple | chili | cherry | olive | plum | almond |
| apples | bananas | brains | kinds | israelis | olives |
| axes | facets | paths | phases | tributaries | concurrency |
| frame | frames | brick | two-story | tubing | rear |
| light | heavy | lights | bright | dark | radiation |
| well | poorly | be | been | however | there |
| bat | bats | crabs | rodents | jharkhand | equator |
| book | books | novel | album | film | story |
| break | breaks | hiatus | breaking | stay | come |
| change | changes | difference | changing | decrease | shift |
| date | dates | location | value | timing | name |
| watch | watches | want | deny | ignore | remember |

Table 13: Nearest neighbors of multiple senses words when $w = 1$

| query | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
|---|---|---|---|---|---|
| bank | banks | river | corporation | company | west |
| cell | cells | protein | proteins | membrane | cellular |
| apple | os | macintosh | microsoft | ios | mac |
| apples | oranges | sugarcane | fruits | fruit | citrus |
| axes | grind | angles | vectors | axe | flint |
| frame | roof | steel | rear | brick | frames |
| light | using | surface | water | dark | red |
| well | such | many | other | including | most |
| bat | bats | innings | batting | bowler | ball |
| book | published | books | written | wrote | story |
| break | off | trying | get | down | breaking |
| change | changes | process | your | need | different |
| date | dates | release | period | dating | days |
| watch | watching | someone | wait | everyone | online |

Table 14: Nearest neighbors of multiple senses words when $w = 6$