NLP - Assignment 1: Distributional Similarity

Shani Finkelstein, Devora Siminovsky

January 21, 2024

1 Generate most similar words

We chose 5 words in the vocabulary, and for each of them we generated list of 20 most similar words according to word2vec:
We choose the following words: ['game', 'back', 'now', 'some', 'million'] and for each of them got the following 20 most similar words.

```
'game' most similar:
[('games', 0.8204563856124878), ('play', 0.7528691291809082), ('players', 0.6834223866462708), ('player', 0.6758631467819214), ('playing',
0.6501080989837646), ('match', 0.6471269130706787), ('scoring', 0.6354659795761108), ('playoffs', 0.619382917881012), ('score', 0.6164776682853699),
('played', 0.6080259680747986), ('season', 0.6049267053604126), ('postseason', 0.598791241645813), ('playoff', 0.5972214341163635), ('team',
0.5896912813186646), ('baseball', 0.587356686592102), ('matches', 0.5809256434440613), ('teams', 0.5808079838752747), ('football',
0.5802282691001892), ('first', 0.5791586637496948), ('final', 0.5779183506965637)]

'back' most similar:
[('out', 0.7960792779922485), ('then', 0.7879571914672852), ('away', 0.7843512296676636), ('again', 0.7833052277565002), ('up', 0.7675327062606812),
('when', 0.7656893134117126), ('come', 0.7635478973388672), ('go', 0.7526348233222961), ('before', 0.7505038976669312), ('coming',
0.7458189129829407), ('but', 0.743777871131897), ('put', 0.7436739206314087), ('down', 0.7401796579360962), ('came', 0.7386630773544312), ('get',
0.7385607361793518), ('just', 0.7360021471977234), ('off', 0.7346910238265991), ('into', 0.7305406928062439), ('return', 0.7279360294342041), ('way',
0.7277121543884277)]

'now' most similar:
[('still', 0.8729714751243591), ('it', 0.8166109919548035), ('but', 0.7990895509719849), ('already', 0.7946615815162659), ('so', 0.7943700551986694),
('just', 0.7919618487358093), ('what', 0.782635509967804), ('only', 0.7825332880020142), ('even', 0.7819492220878601), ('is', 0.7771465182304382),
('though', 0.775760293006897), ('because', 0.7727454900741577), ('once', 0.766930878162384), ('if', 0.765724241733551), ('has', 0.7576674222946167),
('come', 0.7558827996253967), ('going', 0.7558476537323), ('that', 0.7540110349655151), ('well', 0.7533955574035645), ('much', 0.7505015730857849)]

'some' most similar:
[('many', 0.889228880405426), ('few', 0.8672340512275696), ('have', 0.8217975497245789), ('those', 0.8183451890945435), ('more', 0.8131898045539856),
('other', 0.8075711727142334), ('several', 0.7889262437820435), ('these', 0.7845801711082458), ('others', 0.7842263579368591), ('even',
0.7840512990951538), ('there', 0.7832342982292175), ('than', 0.7732689380645752), ('least', 0.7643056511878967), ('most', 0.7634771466255188), ('but'
0.7624861598014832), ('they', 0.7607685327529907), ('well', 0.754676103591919), ('about', 0.7521328926086426), ('only', 0.751836895942688), ('people'
0.7515760660171509)]

'million' most similar:
[('billion', 0.8441469073295593), ('$', 0.7920072078704834), ('dlrs', 0.7894065380096436), ('dollars', 0.7873415350914001), ('1.5',
0.7485544085502625), ('worth', 0.7404153943061829), ('estimated', 0.7357392311096191), ('1.2', 0.7296980023384094), ('euros', 0.7118194699287415),
('100,000', 0.7116228342056274), ('500,000', 0.7112492918968201), ('2.5', 0.7097642421722412), ('1.3', 0.7062451839447021), ('300,000',
0.6951659321784973), ('1.4', 0.6907528042793274), ('nearly', 0.6790913939476013), ('millions', 0.6785868406295776), ('1.6', 0.675882875919342),
('1.7', 0.6758379340171814), ('400,000', 0.6747632622718811)]
```

Figure 1: Most 20 similar words

2 Polysemous words

Polysemous words are words that have several meanings. We found three polysemous words (words with at least two different meanings) such that the top-10 neighbors of each word reflect both word meanings and three polysemous words such that the top-10 neighbors of each word reflect only a single meaning.

2.1 Two meaning in top 10 neighbors

The words we choose are ['draft', 'bat','tie']. The top 10-neighbors for each of them are in the picture below. 2

2.2 One meaning in top 10 neighbors

The possible senses for 'bank' are a financial institution licensed to receive deposits and make loans or a slope. Only the first sense was reflected in the top-10 neighbors.
The possible senses for 'book' are like reading a book, writing a book or to book a fly. Only the first sense was reflected in the top-10 neighbors.
The possible senses for 'fan' are to admire someone or a cool by waving something to create a current of air. Only the first sense was reflected in the top-10 neighbors.

```
Here are the top-10 neighbors of the word 'draft' that reflects both word meanings
[('nfl', 0.6856388449668884), ('qb', 0.6538099505021851), ('drafted', 0.640170693397522), ('pick', 0.630166232585907), ('rookie', 0.605609655380249),
('picks', 0.5945783853530884), ('roster', 0.5893217921257019), ('nba', 0.5856550335884094), ('playoff', 0.5816650986671448), ('playoffs',
0.575737714767456)]

Here are the top-10 neighbors of the word 'bat' that reflects both word meanings
[('bats', 0.6283668875694275), ('ko', 0.5676717162132263), ('bet', 0.5499674081802368), ('ba', 0.5475848913192749), ('wala', 0.5306985974311829),
('ka', 0.527642011642456), ('din', 0.5267859697341919), ('sa', 0.5260174870491028), ('bakit', 0.5182580351829529), ('nga', 0.5024552941322327)]

Here are the top-10 neighbors of the word 'tie' that reflects both word meanings
[('ties', 0.6625783443450928), ('shirt', 0.6500120162963867), ('tied', 0.6126130223274231), ('bow', 0.6059316396713257), ('dress', 0.5931714177131653),
('wear', 0.5844056010246277), ('pants', 0.5790060758590698), ('knot', 0.5751271843910217), ('wearing', 0.5740670561790466), ('suit',
0.5729429721832275)]

Here is the top-10 neighbors of the word 'bank' reflect only a single meaning
[('banking', 0.6421496868133545), ('credit', 0.6313413977622986), ('atm', 0.5995813608169556), ('banks', 0.587649941444397), ('cash',
0.5765193104743958), ('finance', 0.5761327147483826), ('financial', 0.5414320230484009), ('money', 0.5411860942840576), ('bri', 0.5336347222328186),
('card', 0.5313737392425537)]


Here is the top-10 neighbors of the word 'book' reflect only a single meaning
[('books', 0.8519430160522461), ('read', 0.767386794090271), ('reading', 0.7237954139709473), ('review', 0.7215535640716553), ('writing',
0.710425853729248), ('write', 0.709968626499176), ('novel', 0.7027835845947266), ('author', 0.6985722780227661), ('story', 0.6920100450515747),
('ebook', 0.6802876591682434)]


Here is the top-10 neighbors of the word 'fan' reflect only a single meaning
[('fans', 0.7911291718482971), ('belieber', 0.6327586770057678), ('one', 0.6157400012016296), ('directioner', 0.6107414364814758), ('hater',
0.5888530611991882), ('star', 0.5838051438331604), ('just', 0.5834642648696899), ('i', 0.5791181921958923), ('proud', 0.5778546929359436), ('justin',
0.5716624855995178)]
```

Figure 2: Polysemous words

The second group words neighbors reflect only one sense, because probably the frequency of the first sense of the word is higher than the second sense of the word.

3 Synonyms and Antonyms

The triplet of words such that the given conditions are hold are:
w1 = 'happy', w2 = 'joyful', w3 = 'sad'.
This behavior in which the antonyms are more similar than the synonyms can be explained by the fact that it makes more sense that the probability to see the same neighbors next to the words happy and sad (they can match in the same sentence, in the same context) is higher then happy and joyful, because happy and sad tend to be in the same linguistic phase, while joyful is in a higher linguistic phase.
Note: we tried couple of different antonyms that for them the conditions didn't hold- we think it depends on the probability of the words itself to show up in a sentence.

4 The Effect of Different Corpora

In this section, we would like to compare models based on two sources. The first model is based on wikipedia and news text, and the second based on twitter data. For the wikipedia and news model, use the gensim model glove-wiki-gigaword-200 . For the twitter data, use the gensim model glove-twitter-200("glove" is a different algorithm than word2vec, but its essence is similar).

The 5 words that we found whose top 10 neighbors based on the news corpus are very similar to their top 10 neighbors based on the twitter corpus are 'white', 'green', 'blue', 'gray' and 'red'.
Similarity score for the word blue: 8 out of 10.
Similarity score for the word green: 9 out of 10.
Similarity score for the word red: 8 out of 10.
Similarity score for the word white: 8 out of 10.
Similarity score for the word gray: 8 out of 10 .

The 5 words that we found whose top 10 neighbors based on the news corpus are substantially different from the top 10 neighbors based on the twitter corpus are 'towel', 'current', 'bark', 'rose' and 'bat'.
Note: list1 refers to model glove-wiki-gigaword-200 and list2 refers to model glove-twitter-200.

Similarity score for the word towel: 1 out of 10.
list1: ['towels', 'cloth', 'linen', 'napkin', 'scarf', 'wrap', 'toilet', 'cloths', 'washcloth', 'napkins']
list2: ['towels', 'shower', 'blanket', 'bath', 'bathroom', 'wash', 'dryer', 'laundry', 'washing', 'tub']

Similarity score for the word bat: 1 out of 10.
list1: ['bats', 'batting', 'balls', 'batted', 'toss', 'wicket', 'pitch', 'bowled', 'hitter', 'batsman']
list2: ['bats', 'ko', 'bet', 'ba', 'wala', 'ka', 'din', 'sa', 'bakit', 'nga']

Similarity score for the word bark: 0 out of 10. list1: ['twigs', 'mottled', 'birch', 'mulch', 'sawdust', 'trees', 'scaly', 'tree', 'cinchona', 'twig']
list2: ['barking', 'woof', 'bite', 'dog', 'barks', 'meow', 'dogs', 'tick', 'snore', 'chew']

Similarity score for the word rose: 0 out of 10.
list1: ['fell', 'climbed', 'surged', 'jumped', 'shares', 'percent', 'dropped', 'slipped', 'soared', 'dipped']
list2: ['derrick', 'flower', 'blue', 'green', 'roses', 'diamond', 'violet', 'white', 'pink', 'brown']

Similarity score for the word current: 1 out of 10.
list1: ['present', 'future', 'its', 'the', 'change', 'term', 'same','this', 'of', 'year']
list2: ['previous', 'recent', 'future', 'currently', 'despite', 'conditions', 'latest', 'according', 'although', 'potential']

For the second case, we can see a differences in the neighbors words we observed.
The reason is kind of simple, we compared Wikipedia and news model to a twitter model- the writing style in each model is very different.
In Wikipedia and in the news the words are more formal and on Twitter more slang.
In addition, the context of the word in each model is different, in Wikipedia and in the news it will be more based on facts and events that happened. Compared to Twitter, which receives information from tweets posted by private individuals.

Our strategy for finding 5 words whose top 10 neighbors based on the news corpus are very similar to their top 10, was based on thinking about words that don't have much different senses over platforms, models so we would get kind of similar neighbors in each model for the same word.
In order to find 5 words whose top 10 neighbors based on the news corpus are substantially different from the top 10 neighbors based on the twitter corpus was choosing words with several meanings so the output of the most similar words on each model will be different.

5 Plotting words in 2D

We took the first 5000 words in the vocabulary, and kept only words that end either with "ed" or with "ing".
We left with 708 words, we created a matrix with 708 rows where each row is a 300-dim vector for one word, and reduced the dimensionality of the matrix to 2-d using PCA algorithm, the output is a matrix with 708 rows and 2 columns.
In the plot we can see that the model separated the colors pretty well.
About separating, we are talking about words that end with 'ing' and words that end with 'ed'- each group refers to a different time expression, so it's kind of obvious that the colors are separated because the vectors are different.
About the "mutual" section, words that ends with 'ing' and 'ed' can be used in the same context in the same sentence. For example: "While I was runn**ing**, the dog bark**ed**".
From the model, we can get the information about the time expressions, that 'ing' and 'ed' ending can be used in the same context, sentence. Or as usual, the word ending can be individual in the sentence and this is why we see that the model separated the colors(time expressions) really good.

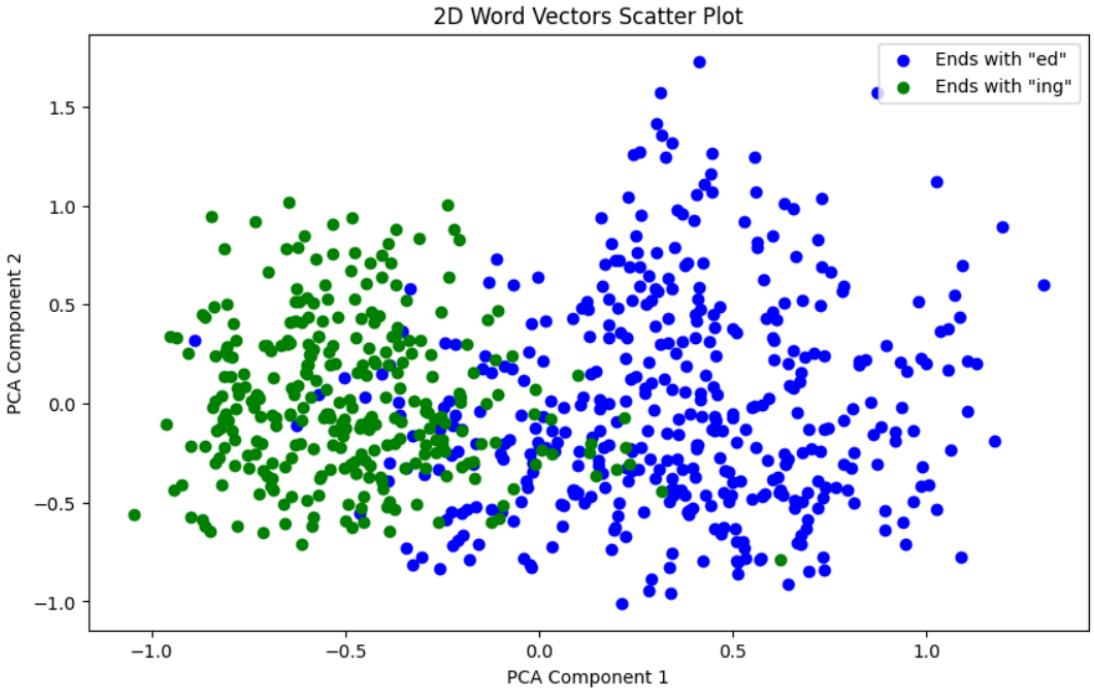

Figure 3: Plotting words in 2D

6 Word-similarities in Large Language Model

6.1 Related words

For each of the 5 words for which we generated the similarity lists in the word2vec part ('game', 'back', 'now', 'some' and 'million'),we will ask ChatGPT to produce a list of 20 similar words for each word. The prompt we used - what are 20 most similar words to the word "–"?

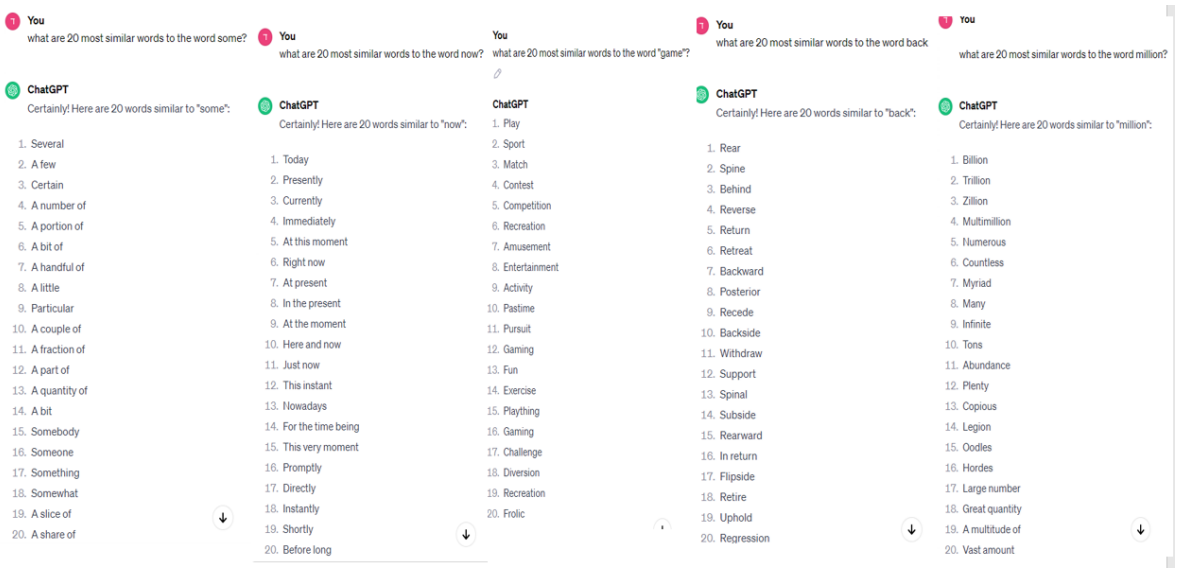

Figure 4: Most similar words in ChatGPT

As we can see, in ChatGPT the words are more general, and close to words that a human will give as similar words to the given word, contrast to result1s of word2vec which some of them are more specific, such as numbers and inflections of the word.
Note: The similar words that ChatGPT produced are all written in capital first letter compared to word2vec which is sensitive to a capitalized first letter in a word and produces a different vector accordingly. For the comparison we will refer to the words in ChatGPT that all of them in lowercase letters.
We got different words in the two methods:
For the word 'game' we found 2/20 matching words: 'play' and 'match'.
For the word 'back' we found 1/20 matching word: 'return'.
For the word 'now' we found 2/20 matching words: 'presently' and 'currently'.
For the word 'some' we found 2/20 matching words: 'certain' and 'several'. (there is also 'A few', 'A handful of' and 'A couple of' that come as a multi-word phrases in ChatGPT compared to word2vec that produced those as a specific words: 'few', 'handful' and 'couple', we didn't included those words on counting matching words)
For the word 'million' we found 1/20 matching word: 'billion'.

In addition, in ChatGPT we got larger diversity, probably because word2vec is based on a narrower corpus.
In ChatGPT the output includes also some multi-word phrases such as 'Large number' and more...
We can make it produce only words by using the prompt - "what are 20 most similar words to the word 'some'? just words without multi-word phrases"

Different types of similarity that are reflected in the lists, are:
1. Inflections of a word - for example: "game" - "games", "million" - "millions". 2. Inclusion of a word - for example: "game" - "activity", "entertainment", "sport"..., "million" - "trillion" 3. Detail of a word - for example: "game" - "football", "game" - "baseball", "now" - "at the moment".
Type 1 appears in word2vec, type 2 appears in word2vec and ChatGPT, type 3 appears in word2vec and ChatGPT.
Now we will select the words 'game' and 'back' and increase the number of neighbors from 20 to 100, for both ChatGPT and word2vec.

The trends we observe are:
In ChatGPT:
1. The first 20 words that ChatGPT will produce will be the same words as in the list of 20 words.
2. As we go down in the list we get more phrases, at least 30 last words are phrases.

In word2vec:
1. The first 20 words that 'word2vec' will produce will be the same words as in the list of 20 words.
2. As we go down in the list we will get the same words just with different inflections, different endings.
3. We will get more phrases, maybe even the same ones just with different inflections.

We used the following prompt to produce 100 neighbors for words above in ChatGPT: 'by giving the words: 'game', 'back'. please produce a list of 100 similar words for each word'.
We used the following command to produce 100 neighbors for the words above in word2vec:
$list = ['game','back']$
$for\ i \in list:$
$print(f"\ '\{i\}'\ most\ similar: model.most\_similar(i,\ topn = 100)")$

We got what we expected.

6.2 Synonyms and Antonyms

Q: Can you get ChatGPT to reliably produce synonyms? Which prompt did you use?
A: Yes, by using the prompt 'can you give me 2 synonyms words: 2 different words with the same meaning?'
Then,we got 2 synonyms words, now after executing the prompt a few times we got 2 synonyms words each time.
The output using the prompt- "can you give me 2 synonyms words?", here it gave us 2 very similar words but not exactly synonyms in some cases so we added to the prompt the following sentence: "(can you give me 2 synonyms words:) 2 different words with the same meaning" as explained above.

Q: Can you get ChatGPT to reliably produce antonyms? Which prompt did you use?
A: Yes, by using the prompt - 'can you give me 2 antonyms words?'.
Then,we got 2 antonyms words, now ran the prompt a few times and got 2 antonyms words each time.

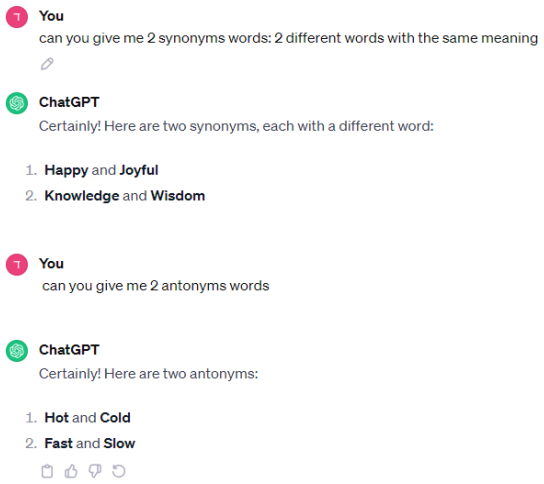

Figure 5: Synonyms and Antonyms words in ChatGPT

6.3 Polysemy

We checked how does ChatGPT behave when we ask him for similarities for the polysemous words we found in the word2vec part: 'draft', 'bat','tie', 'bank', 'book' and 'fan'.
Q: In what senses is it similar, and in what senses is it different from word2vec?
A: When we asked ChatGPT for similarities for the polysemous words we found in the word2vec part, in some cases we got words connected to both meanings as in word2vec and in some words differently, as shown in the table below.
The words we got in ChatGPT are more general than in word2vec (as explained above).

	word2vec	ChatGPT
one meaning	bank book fan	draft bat bank book
two meanings	draft bat tie	tie fan

7 Mean Average Precision (MAP) evaluation

We chose as target words: 'game', 'some'.
We will take the lists of 20 most similar words produced for each target word, mentioned above and printed down here, by each model → then we will have 4 lists, 2 words in each model (there are 2 models) → then we will judge the correctness of their similarities manually follow by 'topically related to the target word' and 'same semantic class as the target word'.

'game' most similar in word2vec:
['games', 'play', 'match', 'matchup', 'agame', 'ballgame', 'thegame', 'opener', 'matches', 'tournament', 'playing', 'league', 'Game', 'scrimmages', 'fourgame', 'scrimmage', 'postseason', 'playoffs', 'gme', 'season']
'some' most similar in word2vec:
['many', 'few', 'lot', 'several', 'plenty', 'those', 'lots', 'these', 'handful', 'all', 'Some', 'little', 'certain', 'couple', 'various', 'other', 'bunch', 'slew', 'numerous','smattering']

'game' most similar in ChatGPT:
["play", "match", "contest", "sport", "competition", "pastime", "amusement", "entertainment", "recreation", "activity", "fun", "diversion", "event", "exercise", "challenge", "gaming", "activity", "hobby", "pursuit", "session"]
'some' most similar in ChatGPT:
["a few", "several", "a number of", "a portion of", "part of", "a bit of", "a little", "a couple of", "certain", "a handful of", "somebody's", "somehow", "someone's", "something's", "someplace", "somewhat", "somewhere", "in some measure", "to some degree", "to some extent"]

Our judgment to the candidate word to be topically related to the target word:
'game' most similar in word2vec:
[1,1,1,0,1,1,1,0,1,1,1,1,1,0,1,0,1,1,0,1]
'game' most similar in ChatGPT:
[1,1,1,1,1,1,1,1,0,1,1,0,0,1,0,1,1,0,0,0]
'some' most similar in word2vec:
[0,1,0,1,0,1,0,1,0,0,1,1,1,1,0,1,1,0,0,0]
'some' most similar in ChatGPT:
[1,1,1,1,1,1,1,1,0,1,0,1,1,1,1,1,1,1,1,1]

Our judgement to the candidate words to be in same semantic class with the target word:
'game' most similar in word2vec:
[1,1,1,1,0,0,0,0,1,1,1,1,1,1,0,0,0,1,1,0,1]
'game' most similar in ChatGPT:
[1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,0]
'some' most similar in word2vec:
[1,0,0,1,1,0,1,0,1,0,1,0,1,0,1,0,0,0,0,1]
'some' most similar in ChatGPT:
[1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0]

```
map mesure for word2vec first judgement:
0.6827054304627834

map mesure for ChatGPT first judgement:
0.9308691864741916

map mesure for word2vec second judgement:
0.6670507506291821

map mesure for ChatGPT second judgement:
0.9917728519282956
```

Figure 6: MAP

We can see from the results that in this case ChatGPT is disticntly better than word2vec model.