GloVe Vectors

NLP Lab 04

Alex Hadley

Spring 2021

Similarity Questions

Table of Most Similar Words

Using a pretrained GloVe model, we can calculate cosine similarities between words. The table below shows ten words and the four words they are closest to in terms of the cosine distance between their GloVe vectors.

Word	1st	2nd	3rd	4th
red	yellow (0.90)	blue (0.89)	green (0.86)	black (0.84)
magenta	cyan (0.86)	fuchsia (0.82)	chartreuse (0.77)	shading (0.73)
flower	flowers (0.84)	fruit (0.79)	tree (0.75)	fruits (0.73)
plant	plants (0.90)	factory (0.76)	facility (0.71)	produce (0.71)
two	three (0.99)	four (0.98)	five (0.96)	six (0.96)
thousand	hundred (0.93)	2,000 (0.91)	1,000 (0.89)	4,000 (0.89)
one	another (0.95)	only (0.94)	same (0.94)	. (0.91)
jupiter	orbiting (0.77)	mars (0.77)	planets (0.77)	orbits (0.76)
mercury	carbon (0.66)	toxic (0.66)	helium (0.66)	oxygen (0.66)
oberlin	swarthmore (0.82)	berklee (0.79)	juilliard (0.77)	graduating (0.76)

Similar Word Discussion

There are a lot of interesting relationships that are revealed by the similarity relationships shown above.

It makes a lot of sense that "red" has a very similar vector compared to other colors. In particular, it is related to other simple, primary colors. On the other hand, "magenta" is similar to "cyan", which makes sense since CMYK is a common color scheme, but is also similar to longer and more specific color names, such as "chartreuse" and "fuchsia". It makes sense that "red", a simpler and more common color, is similar to other simple colors whereas "magenta" is similar to more complicated color words.

It makes sense that "flower" is similar to "flowers", its plural form, as well as other related words such as "tree" and "fruit". The fact that plural forms show up highlights the fact that the GloVe model is analyzing word tokens literally and not doing anything like lemmatization to try to match different forms of a word by their stem. Similarly, we see that "plant" and "plants" are closely related.

The word "plant" is interesting because it has multiple meanings. It could refer to a plant

organism, or a plant as in a factory. In appears that here "plant" is more similar to "factory" and "facility" than the organism. The fourth most common word, "produce", can actually have multiple meanings. It could be a noun referring to edible plants, or a verb meaning to manufacture. Plant organisms can be eaten as produce, and factories can produce items, so it could be relating to either sense of "plant". It would be intersting to do more research involving word senses to see which of these meanings contributes more to the high similarlity of "plant" and "produce".

It makes sense that "two" is similar to other small numbers, especially those close to it. On the other hand, we see that "thousand" is most similar to other multiples of itself ("2,000" and "4,000") rather than closer numbers like "1,001". This makes sense, since "thousand" is probably used more often in similar contexts as multiples of itself or other larger numbers like "hundred". It is also interesting that some of the words "thousand" is related to are numbers written in digits rather than in words. There is no single word for two thousand, so this result makes sense. The word "one" has different kinds of similar words than the other numbers. The most similar word to "one" is "another", which makes sense since "one another" is a common phrase. Similarly, the phrases "only one" and "same one" are common, making it so that "only" and "same" are similar words. However, I am confused about why "one" is so similar to ".", a period. The most common place to find a period is at the end of sentences, but I cannot think of a reason why "one" would be particularly likely to be near the end of sentences like a period.

We can see that "jupiter" is similar to other planet-related words like "orbiting", "planets", and "mars". This makes sense, since "jupiter" is normally only used to refer to the planet. (In theory it could also be used to refer to the Roman god, but that is less common.) Mercury is also a planet; however, we can see that "mercury" is more similar to words relating to the element. This illustrates that words with multiple meanings tend to have one meaning that is far more common.

Finally, "oberlin", a college, is similar to the names of other colleges, such as "swarthmore", as well as college-related words like "graduating". This makes sense. It is interesting that proper nouns are also included in the GloVe dataset. That makes sense, since I believe the authors trained their model on text from Wikipedia.

Discussion of Other Words Tried

Below is a table of some other words and their most similar words in terms of the cosine distance between their GloVe vectors.

Word	1st	2nd	3rd	4th
i	'd (0.96)	me (0.95)	maybe (0.93)	know (0.93)
alex	andy (0.79)	kevin (0.76)	williams (0.74)	matt (0.74)
sourdough	flatbread (0.70)	crusts (0.69)	crusty (0.69)	yeasted (0.68)
nyc	subway (0.69)	metro (0.64)	nightclub (0.62)	westside (0.62)
apple	blackberry (0.75)	chips (0.74)	iphone (0.74)	microsoft (0.73)
nlp	hagelin (0.70)	.760 (0.69)	inp (0.68)	+18 (0.67)

Word	1st	2nd	3rd	4th
claremont	irvine (0.73)	pitzer (0.70)	broomfield (0.69)	pomona (0.67)
trump	casino (0.68)	nows (0.67)	casinos (0.64)	hilton (0.64)
biden	gephardt (0.84)	cheney (0.83)	rodham (0.81)	pelosi (0.81)
corona	cruces (0.67)	janiero (0.66)	segundo (0.66)	colle (0.66)

For "i", it is interesting that the most similar word is "'d". This shows us that pieces of contractions are considered to be separate words, revealing something about how the dataset our GloVe model was trained on was tokenized. Next, I tried putting in my own name, Alex. The most similar words are also generic male first names, which makes sense. The word "sourdough" is related to other bread-related words, which is no surprise. Similarly, "nyc" is similar to other words used in relation to New York City.

I found it particularly interesting that the word "apple" is most similar to words relating to the technology company, not the fruit. Before Apple became such a large company, this would not have been the case. It is interesting that proper nouns can eclipse the meanings of ordinary words, and that the words most similar to another word can change over time.

Next, I tried putting "nlp", the abbreviation for natural language processing. The most similar word is "hagelin". I searched for this word online, and determined that it refers to John Hagelin, a US presidential candidate from the Natural Law Party (NLP). Notably, there is a Wikipedia article about him that uses the abbreviation "nlp" several times. Since the GloVe model was trained on text from Wikipedia, this makes sense. On the other hand, I could not find out what ".760", "inp", and "+18" have to do with the word "nlp". My guess is that maybe "nlp" was pretty uncommon in the dataset and it happened to show up near these other uncommon tokens, leading to them turning out similar in the model.

I also found that "claremont" was similar to "irvine", a location near Claremont, as well as "pitzer", as in Pitzer College! I believe Pomona is both a college and a nearby location, so "pomona" could be similar for either reason.

I tried the word "trump" because I was curious if the results would be different than we would expect in 2021. Sure enough, the most similar was "casino". While this might have previously been highly associated with Donald Trump, he would likely now have a different set of similar words. I believe this GloVe model was trained on Wikipedia articles from 2014 in particular, so this behavior makes sense. It is interesting how word meanings, especially those relating to current events, can change drastically over time. I also tried looking up "biden" since he only recently became president; however, since he has been in politics for a long time and was the vice president in 2014, the most similar words are in fact other political figures. I do wonder though how the most similar words to "biden" have changed since 2014 or will change in the coming years.

Finally, I looked up words similar to "corona". Obviously, the coronavirus pandemic had not happened yet in 2014 when this model was trained, so none of the most similar words are related to the virus or the pandemic. Furthermore, I tried looking up how some of the other words related to "corona", but it was almost impossible to do since any Google search with

"corona" in it now brings up millions of results relating to COVID-19. This is an interesting example of how a pretty obscure word can suddenly become common and get a new meaning.

Plotting / Visualization Questions

Analysis of Given Relations

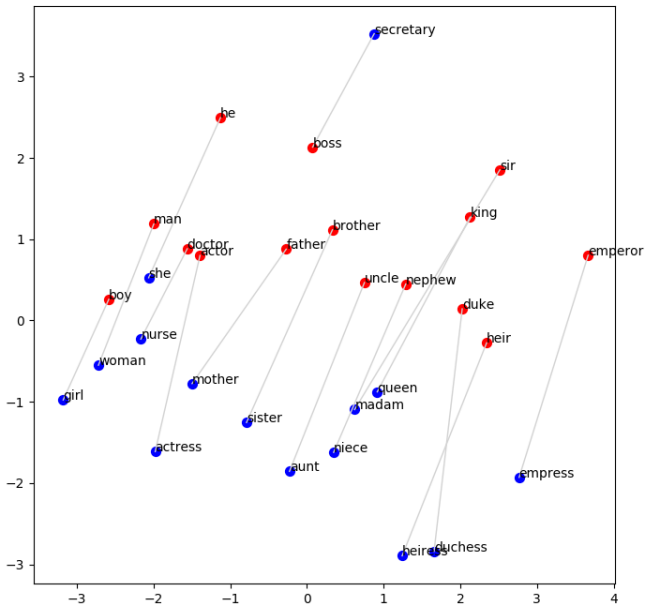

Figure 1: GloVe relationships between gender words

In Figure 1, the plot of the relationships between gender-related words, there is a pretty clear distinction between two separate categories. Also, it appears that the direction and distance of the vector needed to get from one word to the other are similar. Some pairs, such as "man" and "woman" or "father" and "mother", are just definitionally related to males or females. However, some pairs represent societal steriotypes about the roles filled by men or women. For example, the relationship between "doctor" and "nurse" appears very similar to the other definitionally gendered pairs. The relationship between "boss" and "secretary" appears as a bit of an outlier. "boss" is kind of close to the other male-related words, but secretary is on the other end of the graph. I am not sure exactly why this is. The slope of the line between the two vectors is similar to the other pairs. I would conclude that there might be gender sterotype being displayed between a boss and secretary role, but there also might be something different going on for this pair.

In Figure 2, the plot of relationships between opposite words, there is less of an obvious pattern. One reason for this disorganization could be that there is nothing particular about one category or the other in this case. Before, one side of the relationships referred to male

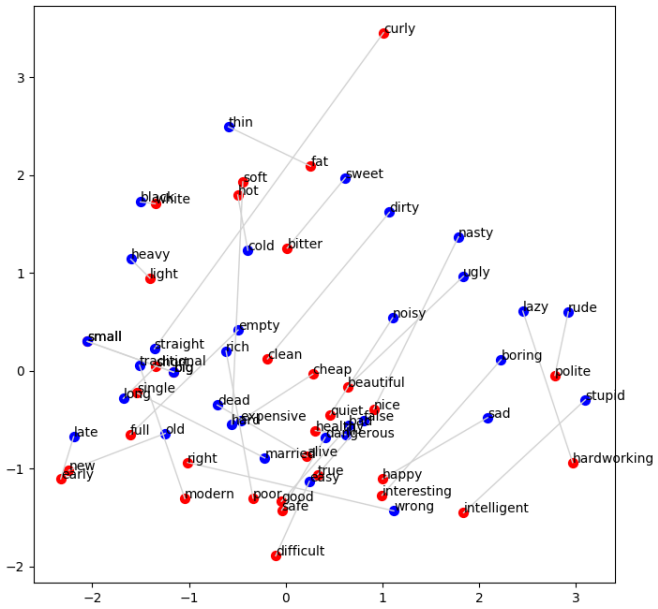

Figure 2: GloVe relationships between opposite words

words and the other referred to female words; however, opposites do not have a particular direction. There is also not an overall pattern regarding the direction or distance between vectors. Some of the differences do appear parallel, such as the group near the middle with a slope close to -1, suggesting that maybe there are some subcategories of opposite pairs that are more similar in nature.

In Figure 3, the plot of countries and their corresponding capitals, there are two clear groups, with capitals and countries clustered together. Furthermore, most countries and capitals appear to be connected by parallel lines of similar length, meaning that the relationships are similar in nature. This makes sense, since we would expect capitals to appear in contexts related to their particular countries, and to have similar relationships to their capitals in language.

In Figure 4, the plot showing the relationships between demonyms and their corresponding countries, it is a little hard to see what is going on. At first, it looks a little choatic; however, there are generally two overlapping groups, one of red dots representing countries and another of blue dots representing demonyms. Most of the relationships go are roughly parallel, averaging out to be horizontal; however, sometimes the country and corresponding demonym are closer together than others. For instance, “czech” is both the country and the denonym, so in that case the two points overlap! Perhaps pairs are closer together when the demonym is more similar to the spelling of the country. It also seems like while countries and corresponding demonyms are relatively close together, countries themselves are still spread out. This makes sense, since a country and demonym would be used together more often

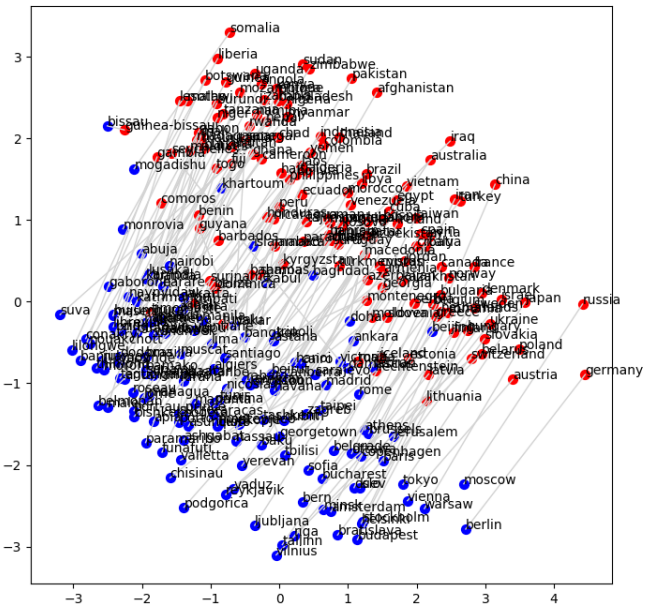

Figure 3: GloVe relationships between countries and capitals

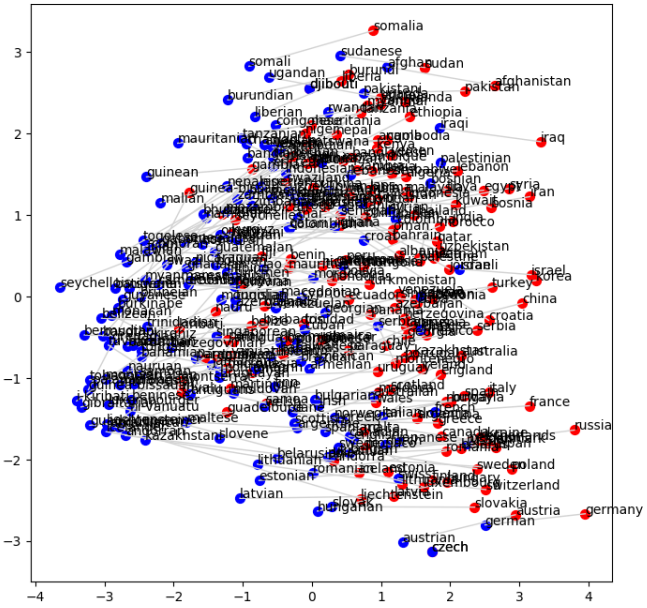

Figure 4: GloVe relationships between demonyms

than two particular countries.

Analysis of New Relations

I also analyzed a file of relationships that I created. In the file, I wrote around 20 meronym relationships, where the first word is a part of the thing in the second word. For example, "petal" and "flower" or "engine" and "car". The plot is shown in Figure 5.

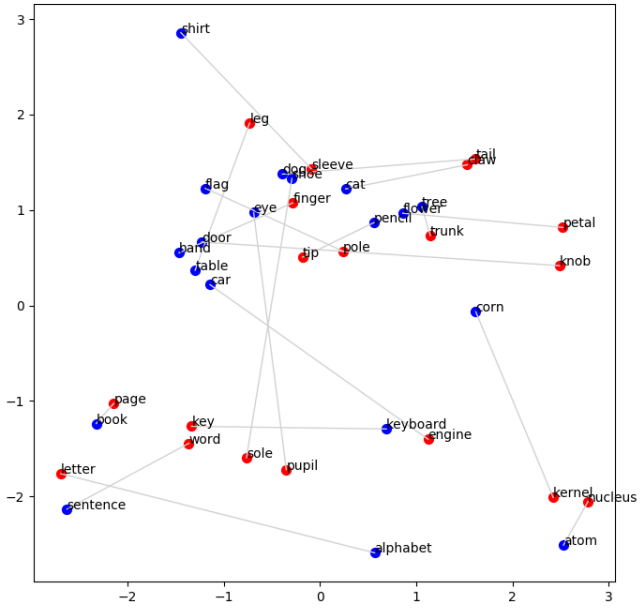

Figure 5: GloVe relationships between meronyms

I expected to possibly find a plot that looked like the gender plot, with the relationships between parts and wholes being of similar length and parallel. However, the resulting graph does not really show a particular pattern. In fact, I had trouble finding even a single pair of relationships that seem similar. This might have to do with the fact that my examples come from many different categories. For example, "shirt" and "sleeve" are related to clothing, whereas "atom" and "nucleus" are related to science. Part-whole relationships might just not be strong enough in natural language to show up as an overall pattern amoung relationships from many different domains.

Prediction Questions

Comparison of first word to second word in relations

The second column in the following table contains the frequency of how often the first word in each relationship was the most similar word to the second word in the relationship. The

third column contains frequencies of how often the first word was in the top 10 most similar words to the second. The third column contains the mean reciprocal rank (MRR), which is the average reciprocal rank of the first word in the list of words most similar to the second. MRR is a measure of how similar words in each set of relationships tended to be.

File	1st most similar	1st in top 10	MRR
gender.txt	0.083	0.583	0.269
opposites.txt	0.192	0.308	0.244
capitals.txt	0.097	0.354	0.169
demonyms.txt	0.148	0.528	0.265

Comparison of first word to vector sum in relations

The following table has the same statistics as the previous table. However, this time we are comparing the vector of the first word to the vector of the second word plus the average vector difference between the first and second words in each set of relationships. Essentially, this is using GloVe vectors to predict the first word from the second word by adding the average difference between the two vectors. We would expect this prediction method to work better if relationships in the set tend to be similar in length and direction.

File	1st most similar	1st in top 10	MRR
gender.txt	0.667	1	0.833
opposites.txt	0	0.429	0.108
capitals.txt	0.679	1	0.803
demonyms.txt	0.639	0.806	0.695

Discussion of Above Comparisons

First we will discuss the comparisons between the first and second words. In this case, higher frequencies mean that words in pairs for the particular set of relationships tend to be similar to each other. It makes sense that the similarity rates for gender.txt are relatively high, since many of the gendered words are related. For example, “king” and “queen” are both words for royalty, or “brother” and “sister” are both words for sibings. For demonyms.txt, I think it is important to keep in mind the directionality in which we are analyzing similarity. It makes sense that the corresponding country would be one of the first words to show up when looking at words similar to the demonym, explaining why the frequencies and MRR are relatively high; however, in the other direction I would expect that other words might be more similar to a country than its demonym, for example nearby countries. I am a little surprised that the frequencies for capitals.txt are so much lower than those for demonyms.txt, since in that case we are also looking in the direction from capital city to country. I also expected the rates of opposite.txt to be the lowest given that its plot was the most chaotic; however, the structure shown in the plot has more to do with the prediction portion.

Looking at the prediction statistics, we can see that indeed opposites.txt has the lowest

similarity frequencies and MRR. This makes sense given that there was no obvious pattern in its plot. Our method of prediction was to add the average difference vector between the two words, which works best when most of the relationships in the set have similar direction and magnitude. This explains why the prediction rates are pretty high for gender.txt, capitals.txt, and demonyms.txt. Their plots showed more similar relationship difference vectors between points, leading to a more accurate prediction using this method.

Discussion of Similarity Across Multiple Runs

Because the train data set, from which we calculated the average difference vector, and the test data set, which we made predictions on, were randomly assigned, it is important to do a few tests in order to get a sense for how much of our result is due to randomness and how much is consistent across trials. I conducted 12 trials, 3 for each text file. The table below shows results for the comparison of the first word to the second word in relations.

Trial	File	1st most similar	1st in top 10	MRR
1	gender.txt	0.250	0.667	0.409
2	gender.txt	0.250	0.667	0.405
3	gender.txt	0.167	0.667	0.354
4	opposites.txt	0.154	0.308	0.213
5	opposites.txt	0.154	0.308	0.211
6	opposites.txt	0.115	0.308	0.181
7	capitals.txt	0.097	0.398	0.180
8	capitals.txt	0.080	0.398	0.175
9	capitals.txt	0.080	0.398	0.174
10	demonyms.txt	0.141	0.472	0.251
11	demonyms.txt	0.134	0.535	0.256
12	demonyms.txt	0.134	0.542	0.258

When looking at these results, it is important to keep in mind that gender.txt contained 15 relationship pairs, opposites.txt contained 33 pairs, capital.txt contained 144 pairs, and demonyms contained 184 pairs. In correspondance with these size differences, can see than gender.txt changed the most between trials, whereas demonyms.txt very consistent. Overall, the statistics do not change too drastically between trials.

The following table shows comparisons of the first word to the vector sum prediction in relations.

Trial	File	1st most similar	1st in top 10	MRR
1	gender.txt	0	0.667	0.333
2	gender.txt	0.333	1	0.667
3	gender.txt	0.667	0.667	0.667
4	opposites.txt	0	0.286	0.106
5	opposites.txt	0	0.286	0.107

Trial	File	1st most similar	1st in top 10	MRR
6	opposites.txt	0	0.429	0.128
7	capitals.txt	0.643	1	0.813
8	capitals.txt	0.679	0.857	0.762
9	capitals.txt	0.607	0.929	0.697
10	demonyms.txt	0.778	0.861	0.816
11	demonyms.txt	0.583	0.722	0.621
12	demonyms.txt	0.583	0.806	0.674

Note that the test set was 20% of the total set. So for gender.txt, the test set used for these trials had only 3 words in it, which helps explain why we see such variation in the results. Similarly, opposites.txt only had a few words in the prediction set and likewise has some variation. It is a little surprising the demonyms.txt, with is larger sample size, changes from and MRR of 0.816 in trial 10 to below 0.7 in the final two trials. However, this can still be explained by the fact that 20% of the total relations is still around 37 words for this set, which is relatively small. Also, the inclusion or exclusion of particular outliers (such as "boss" and "secretary" in the gender category) might have a significant effect on these statistics between random samples.

Discussion of Extension Results

For my additional analysis, I performed the above experiments using the glove.6B.200d.txt model, which has 200 dimensional vectors as opposed to the 50 dimensional vectors we were using previously. The table below shows comparisons of the first word to the second word in relations.

File	1st most similar	1st in top 10	MRR
gender.txt	0.167	0.917	0.431
opposites.txt	0.154	0.462	0.211
capitals.txt	0.230	0.867	0.441
demonyms.txt	0.380	0.718	0.503

All of these statistics are higher. 200 dimensional vectors contain more information, so maybe we are more likely to pick up relationships between words than we were before. For example, the pairs in capitals.txt and demonyms.txt are definitionally related by country, so they should be very closely related. Maybe the higher dimensional vectors are better able to pick up these correspondances.

The table below shows comparisions of the first word to the vector sum prediction for relationships.

File	1st most similar	1st in top 10	MRR
gender.txt	0	0.667	0.333
opposites.txt	0	0.143	0.059
capitals.txt	0.464	1	0.732
demonyms.txt	0.528	0.944	0.729

As we saw earlier, the results for gender.txt and opposites.txt are a bit unreliable in this category due to the very low sample size. For capitals.txt and demonyms.txt, the MRR values are within the range of results from our multiple trials using 50 dimensional vectors. In order to see if there is actually an improvement in predictive power, we would have to do more trials using 200 dimensional vectors and compare the range of values to what we got before.