LSA of Conceptual Combination Stimuli

Rory Flemming

October 17, 2018

Word Pairs

```
# Import word pairs. The data is coded in the following way
# col 1- the first word of stimulus
# col 2- the second word of the stimulus
# col 3- 0 if 1st word is "Concrete", 1 if first word is "Abstract"
# col 4- order of stimulus presentation. 0 indicates the example pair in "training"
CC_stims <- read.csv(file='CC_stimuli_POA-5.csv',header=TRUE)
str(CC_stims)
```

```
## 'data.frame':    13 obs. of  4 variables:
##  $ word1 : Factor w/ 13 levels "chimney","frog",..: 3 2 6 5 11 10 12 8 7 13 ...
##  $ word2 : Factor w/ 13 levels "bliss","cactus",..: 8 9 4 11 10 6 7 2 1 13 ...
##  $ abs1st: int  1 0 1 1 0 0 1 1 0 1 ...
##  $ order : int  1 2 3 4 5 6 7 8 9 10 ...
```

Comparison Corpi

We will use three comparison corpi, for validation purposes:
* TASA - "Touchstone Applied Science Associates, Inc." A corpus of a broad set of topics and used to compile "The Educator's Word Frequency Guide." Built from >37.5k documents, and containing >92k different terms.
* EN_100k - The recommended space for computations in English. Absolutely massive (too big for me to venture downloading). ~2 Billion words, almost 5.4 million documents, with rows on 100k most frequent words. 5.4mill dimensions reduced to 300 via SVD.

```
load("TASA.rda")
TASA_mat = as.textmatrix(TASA)
```

LSA: Semantic Distance

The first two things I will look at are:
1) What are the magnitudes of the semantic distances between the first words and the second words of the stimulus pairs?
2) Are these cosine distances symmetric? (They should be)

```
# First, let's just get a read of the semantic distances of the pairs
for (i in 1:nrow(CC_stims)){ # For each CC stimulus
  # compute cosine distance of the words in the ith pair
  CC_stims$semantic_distance[i] = 1 - abs(Cosine(x=CC_stims$word1[i],
                                                 y=CC_stims$word2[i],tvectors=TASA_mat));
  # same computation, but switch their places
  CC_stims$semantic_distance2[i] = 1 - abs(Cosine(x=CC_stims$word2[i],
```

```
                                                    y=CC_stims$word1[i],tvectors=TASA_mat));
}
CC_stims # look at the results
```

```
##        word1   word2 abs1st order semantic_distance semantic_distance2
## 1    honesty  ladder      1     1         0.9713464          0.9713464
## 2       frog    luck      0     2         0.9849996          0.9849996
## 3      mercy    comb      1     3         0.9874850          0.9874850
## 4    justice  pillow      1     4         0.9761459          0.9761459
## 5      tiger paradox      0     5         0.9666753          0.9666753
## 6    thimble   glory      0     6         0.9546344          0.9546344
## 7  willpower    kite      1     7         0.9728395          0.9728395
## 8  reasoning  cactus      1     8         0.9837986          0.9837986
## 9     parrot   bliss      0     9         0.9479750          0.9479750
## 10    wisdom tractor      1    10         0.9820657          0.9820657
## 11    sponge purpose      0    11         0.9797330          0.9797330
## 12      hope   clock      1    12         0.8931780          0.8931780
## 13   chimney courage      0     0         0.9414438          0.9414438
```

The table shows us that the semantic distances of the stimuli range [0.893178, 0.987485], and that the distances, when computed by cosine distances are symmetric. Other distance metrics do allow for asymmetry. Really quick, I am going to download the EN_100k LSA space, and run this same analysis to see if we get similar results. Due to the size, more sophisticated analysis ought to be done on another machine, if it involves storing this data while manipulating it or other vaiables...

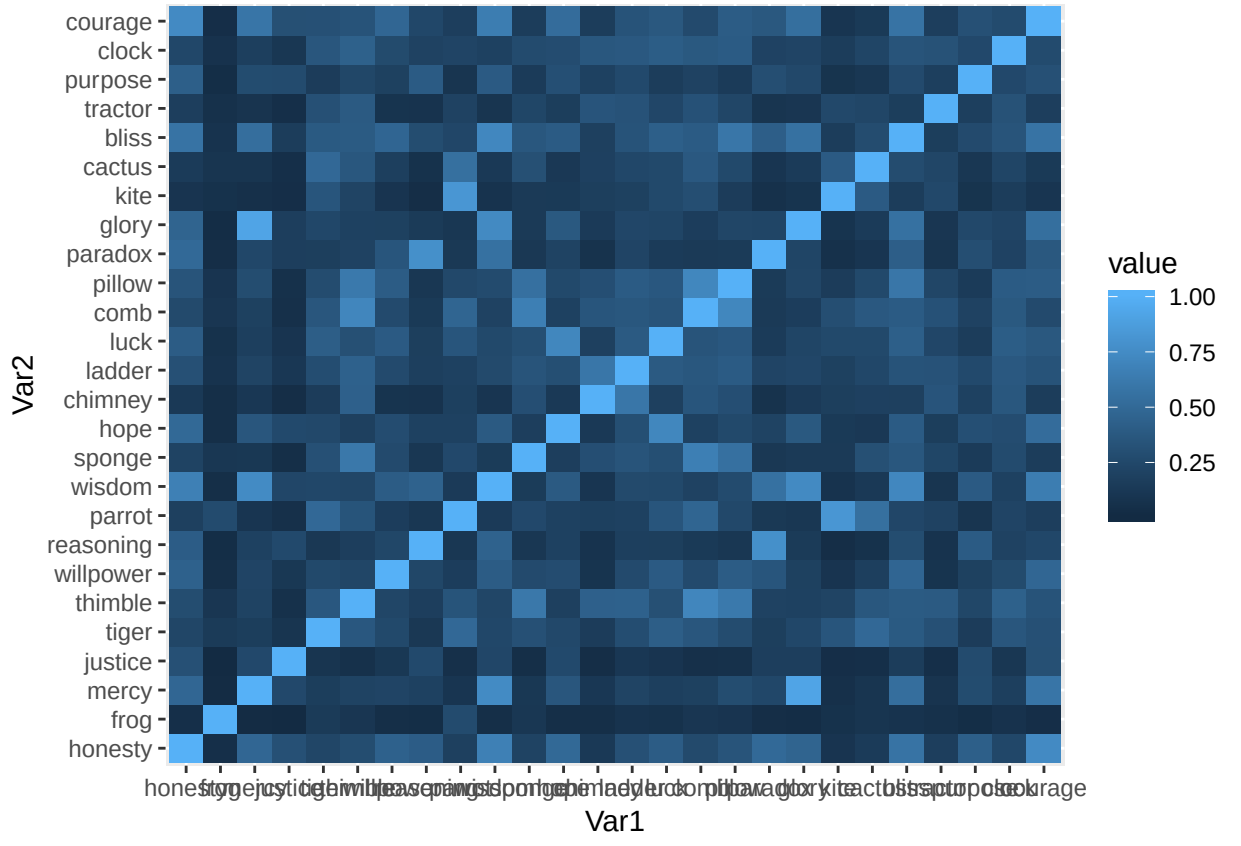
```
load('EN_100k_lsa.rda');
EN_100k = as.textmatrix(EN_100k_lsa);
for (i in 1:nrow(CC_stims)){ # For each CC stimulus
  # compute cosine distance of the words in the ith pair
  CC_stims$semantic_distance2[i] = 1 - abs(Cosine(x=CC_stims$word2[i],
                                                    y=CC_stims$word1[i],tvectors=EN_100k));
}
CC_stims
```

```
##        word1   word2 abs1st order semantic_distance semantic_distance2
## 1    honesty  ladder      1     1         0.9713464          0.6909368
## 2       frog    luck      0     2         0.9849996          0.9331300
## 3      mercy    comb      1     3         0.9874850          0.8122425
## 4    justice  pillow      1     4         0.9761459          0.9387678
## 5      tiger paradox      0     5         0.9666753          0.8277675
## 6    thimble   glory      0     6         0.9546344          0.8085583
## 7  willpower    kite      1     7         0.9728395          0.9120745
## 8  reasoning  cactus      1     8         0.9837986          0.9328778
## 9     parrot   bliss      0     9         0.9479750          0.7657528
## 10    wisdom tractor      1    10         0.9820657          0.9097992
## 11    sponge purpose      0    11         0.9797330          0.8549207
## 12      hope   clock      1    12         0.8931780          0.7198179
## 13   chimney courage      0     0         0.9414438          0.8411402
```

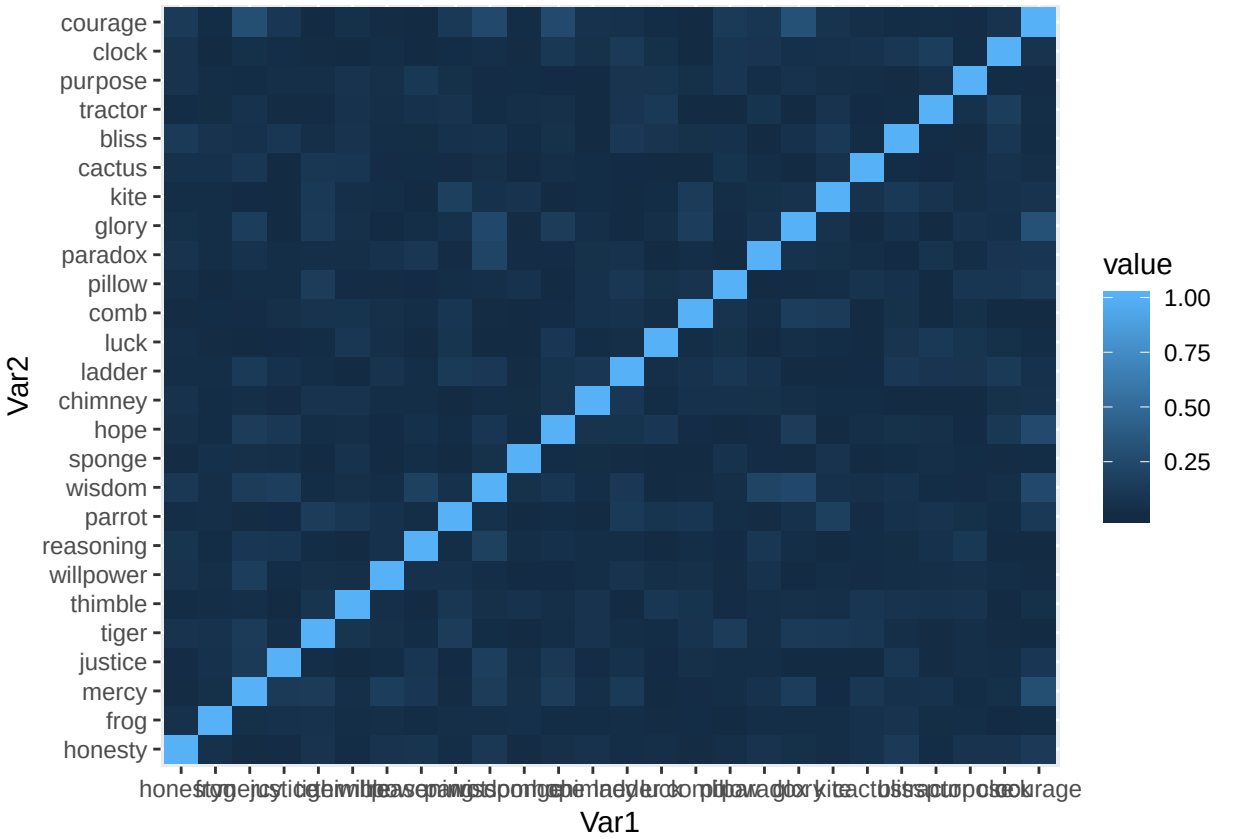
Woah!! We actually get way different, and richer results. There is more variability at the very lease. Maybe the TASA corpus is not appropriate for our word content. I would be more inclined to "trust" the EN_100k since it is so much larger and more general. Still, it is a tough quest to decide on an appropriate LSA

space... Before tossing this out, let's have a look at the results just from these two and see if there is some correlation between them...
What I'm going to try to do next is to create the covariance matrix for each of the stimulus words in each of the spaces, then compare them to one another. The point of doing this would be to see if we can correctly identify "concrete" vs "abstract" words in the stimulus set.

```
# Separate out the words and their label (Abs, concrete)
words = c(as.character(CC_stims$word1),as.character(CC_stims$word2));
labels = c(CC_stims$abs1st,(1-abs(CC_stims$abs1st)));

# Compute semantic simmilarity on all of the words
words_simmat_EN_100k = abs(multicos(words,words,EN_100k));
word_simmat_TASA = abs(multicos(words,words,TASA_mat));

melted_EN100k = melt(words_simmat_EN_100k);
melted_TASA = melt(word_simmat_TASA);

ggplot(melted_EN100k, aes(x=Var1,y=Var2,fill=value)) + geom_tile()
```




```
ggplot(melted_TASA, aes(x=Var1,y=Var2,fill=value)) + geom_tile()
```



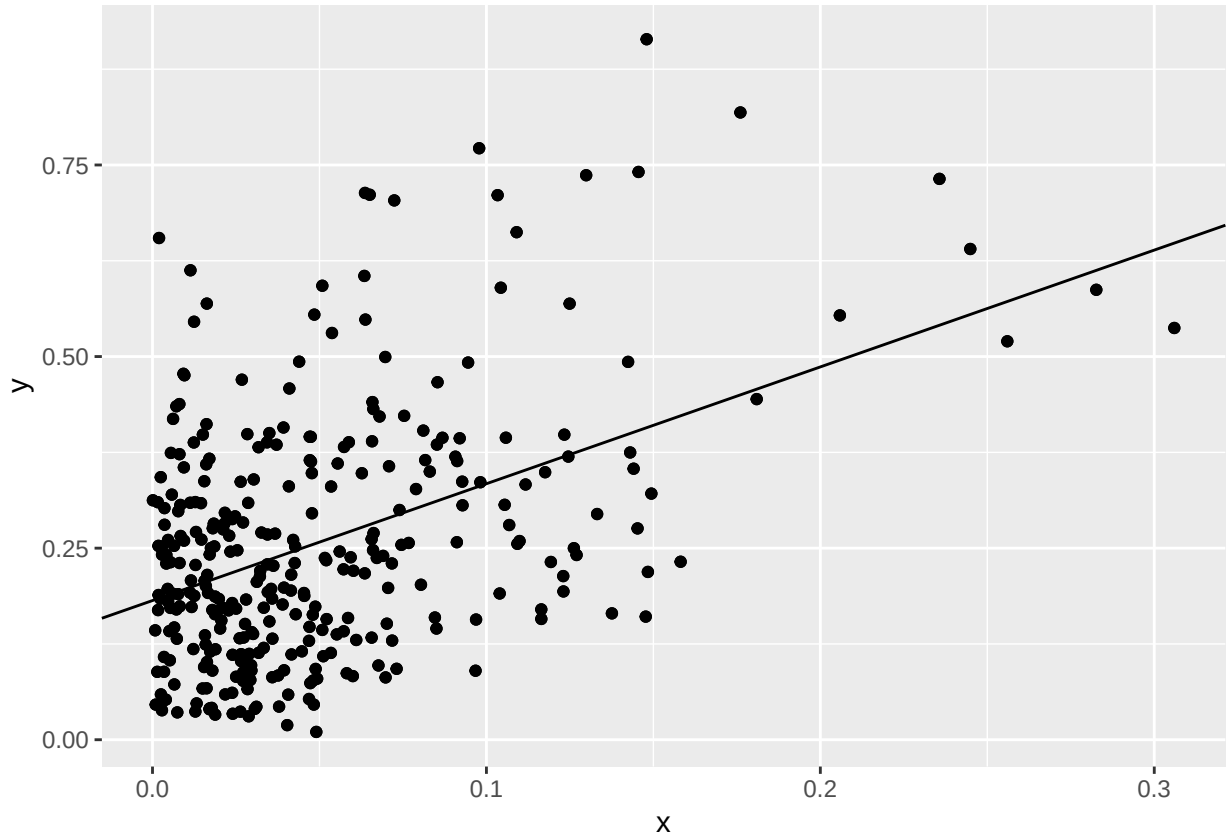
Now, these huge matrices are a bit hard to interpret... Now, how can we look at their simmilarity? I'm going to start super simple and just ask: do the two LSA spaces show a consistent pattern of correlation? In other words, is there a positive relationship between the covariances calculated in TASA and EN_100k? Here, I fit a simple lienar model:

Let's look at the results:

```
# Report the coefficients
TASA2EN100k$coefficients
```

```
## (Intercept)           x
##   0.1816351   1.5243273
```

```
#Plot model fit
ggplot(data=NULL,aes(x=x,y=y)) +geom_point() + geom_abline(slope=TASA2EN100k$coefficients[2],intercept =
```



I'm sure there is an easier way to visualize (and probably a better way to model) this, but from this we see that:
1) The TASA and EN_100k spaces predict *similar* covariances semantic distances for our stimulus set.
2) The EN_100k presents more variability in the distances, but both are "uninformative" in some regards where the other is informative. This contributes somewhat to the large residuals.