Imports

```
import gensim, logging
from gensim.models import Word2Vec
from nltk.corpus import brown, movie_reviews
```

Vectors from NLTK

```
f = '%(asctime)s : %(levelname)s : %(message)s'
logging.basicConfig(format=f, level=logging.INFO)

b = Word2Vec(brown.sents())
b_opt = Word2Vec(brown.sents(), size=150, window=10,
                                 min_count=2, workers=10,
                                 ns_exponent=0.75)

mr = Word2Vec(movie_reviews.sents())
```

Explore!

- What words are most different between different corpora

```
mr.most_similar('flop', topn=10)
```

- Play with different parameter settings (how small can embedding size get before it gets crappy, how does smaller window change nearest words, how does negative sampling exponent change things)
- Try it out on different datasets!
- Create a t-SNE (from sklearn.manifold import TSNE)

```
X = model.wv[model.wv.vocab]

tsne = TSNE(n_components=2)
X_tsne = tsne.fit_transform(X)

plt.scatter(X_tsne[:, 0], X_tsne[:, 1])
plt.show()
```

word2vec's samples

Generate the "normal" distribution over words from the Brown corpus and sample from words from that distribution

```
>>> from nltk.corpus import brown
>>> from nltk import FreqDist
>>> brown_words = FreqDist(brown.words())
>>> [prob_dist.generate() for _ in range(25)]
['worked', 'line', 'an', "''", 'visit', 'in', ',', '.', 'ho
```

Negative Sampling Distribution

Now create Word2Vec's negative sampling distribution and sample from it.

```
>>> neg_samp = FreqDist()
>>> for w in brown_words:
        neg_samp[w] = brow_words.freq(w) ** 0.75
>>> neg_dist = MLEProbDist(neg_samp)
>>> [neg_dist.generate() for _ in range(25)]
['vanished', 'applied', 'consonantal', 'allocations', 'typ
```

Negative Sampling Distribution

Now create Word2Vec's negative sampling distribution and sample from it.

```
>>> neg_samp = FreqDist()
>>> for w in brown_words:
        neg_samp[w] = brow_words.freq(w) ** 0.75
>>> neg_dist = MLEProbDist(neg_samp)
>>> [neg_dist.generate() for _ in range(25)]
['vanished', 'applied', 'consonantal', 'allocations', 'typ
```

What's different?

Exam Question

Let's say that you set the Word and Context vectors from Word2Vec to be the same matrix. How would this:

1. Affect the number of parameters?
2. Cause problems with the gradient?
3. Change the "story" of the model?

Exam Question

Let's say that you set the Word and Context vectors from Word2Vec to be the same matrix. How would this:

1. Affect the number of parameters? **Half as many**
2. Cause problems with the gradient?
3. Change the "story" of the model?

Exam Question

Let's say that you set the Word and Context vectors from Word2Vec to be the same matrix. How would this:

1. Affect the number of parameters? **Half as many**
2. Cause problems with the gradient? **If same word was predicted from a context, quadratic terms in gradient**
3. Change the "story" of the model?

Exam Question

Let's say that you set the Word and Context vectors from Word2Vec to be the same matrix. How would this:

1. Affect the number of parameters? **Half as many**
2. Cause problems with the gradient? **If same word was predicted from a context, quadratic terms in gradient**
3. Change the "story" of the model? **Parameters need to do "double duty": predict what will appear in a context and be those predictions**

Dataset

- Two types of words
  - Vehicles
  - Fruits
- Learn a representation with two dimensions
- Word2Vec skipgram negative sampling
- $\alpha = 0.1$ (bad choice in practice!)
- We'll do update for one positive and one negative sample
  - Note: much of word2vec magic is sampling negative words, you'll have to take my word for it

Word		
ambulance	-0.228	0.099
apple	0.078	0.217
backhoe	-0.086	0.138
banana	0.046	0.195
crane	-0.220	0.153
firetruck	0.039	-0.047
lemon	0.008	-0.043
strawberry	0.202	-0.081

Context		
ambulance	0.000	0.000
apple	0.000	0.000
backhoe	0.000	0.000
banana	0.000	0.000
crane	0.000	0.000
firetruck	0.000	0.000
lemon	0.000	0.000
strawberry	0.000	0.000

$\alpha = 0.1$

POS (banana vs lemon)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{lemon}}$

POS (banana vs lemon)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{lemon}} = 0.046 * 0.000 + 0.195 * 0.000$

POS (banana vs lemon)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{lemon}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$

POS (banana vs lemon)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{lemon}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 1.0 - \pi = 1.0 - \sigma(0.000) =$

POS (banana vs lemon)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{lemon}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 1.0 - \pi = 1.0 - \sigma(0.000) = 0.500$

POS (banana vs lemon)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{lemon}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 1.0 - \pi = 1.0 - \sigma(0.000) = 0.500$
- $\Delta w_{\text{banana}} = \alpha e \cdot c_{\text{lemon}} =$

POS (banana vs lemon)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{lemon}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 1.0 - \pi = 1.0 - \sigma(0.000) = 0.500$
- $\Delta w_{\text{banana}} = \alpha e \cdot c_{\text{lemon}} = 0.10 \cdot 0.500 \cdot (0.000, 0.000) =$

POS (banana vs lemon)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{lemon}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 1.0 - \pi = 1.0 - \sigma(0.000) = 0.500$
- $\Delta w_{\text{banana}} = \alpha e \cdot c_{\text{lemon}} = 0.10 \cdot 0.500 \cdot (0.000, 0.000) = (0.000, 0.000)$

POS (banana vs lemon)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{lemon}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 1.0 - \pi = 1.0 - \sigma(0.000) = 0.500$
- $\Delta w_{\text{banana}} = \alpha e \cdot c_{\text{lemon}} = 0.10 \cdot 0.500 \cdot (0.000, 0.000) = (0.000, 0.000)$
- $\Delta c_{\text{lemon}} = \alpha e \cdot w_{\text{banana}} =$

POS (banana vs lemon)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{lemon}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 1.0 - \pi = 1.0 - \sigma(0.000) = 0.500$
- $\Delta w_{\text{banana}} = \alpha e \cdot c_{\text{lemon}} = 0.10 \cdot 0.500 \cdot (0.000, 0.000) = (0.000, 0.000)$
- $\Delta c_{\text{lemon}} = \alpha e \cdot w_{\text{banana}} = 0.10 \cdot 0.500 \cdot (0.046, 0.195) =$

POS (banana vs lemon)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{lemon}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 1.0 - \pi = 1.0 - \sigma(0.000) = 0.500$
- $\Delta w_{\text{banana}} = \alpha e \cdot c_{\text{lemon}} = 0.10 \cdot 0.500 \cdot (0.000, 0.000) = (0.000, 0.000)$
- $\Delta c_{\text{lemon}} = \alpha e \cdot w_{\text{banana}} = 0.10 \cdot 0.500 \cdot (0.046, 0.195) = (0.002, 0.010)$

NEG (banana vs firetruck)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{firetruck}}$

NEG (banana vs firetruck)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{firetruck}} = 0.046 * 0.000 + 0.195 * 0.000$

NEG (banana vs firetruck)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{firetruck}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$

NEG (banana vs firetruck)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{firetruck}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 0.0 - \pi = 0.0 - \sigma(0.000) =$

NEG (banana vs firetruck)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{firetruck}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 0.0 - \pi = 0.0 - \sigma(0.000) = -0.500$

NEG (banana vs firetruck)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{firetruck}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 0.0 - \pi = 0.0 - \sigma(0.000) = -0.500$
- $\Delta w_{\text{banana}} = \alpha e \cdot c_{\text{firetruck}} =$

NEG (banana vs firetruck)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{firetruck}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 0.0 - \pi = 0.0 - \sigma(0.000) = -0.500$
- $\Delta w_{\text{banana}} = \alpha e \cdot c_{\text{firetruck}} = 0.10 \cdot -0.500 \cdot (0.000, 0.000) =$

NEG (banana vs firetruck)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{firetruck}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 0.0 - \pi = 0.0 - \sigma(0.000) = -0.500$
- $\Delta w_{\text{banana}} = \alpha e \cdot c_{\text{firetruck}} = 0.10 \cdot -0.500 \cdot (0.000, 0.000) = (-0.000, -0.000)$

NEG (banana vs firetruck)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{firetruck}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 0.0 - \pi = 0.0 - \sigma(0.000) = -0.500$
- $\Delta w_{\text{banana}} = \alpha e \cdot c_{\text{firetruck}} = 0.10 \cdot -0.500 \cdot (0.000, 0.000) = (-0.000, -0.000)$
- $\Delta c_{\text{firetruck}} = \alpha e \cdot w_{\text{banana}} =$

NEG (banana vs firetruck)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{firetruck}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 0.0 - \pi = 0.0 - \sigma(0.000) = -0.500$
- $\Delta w_{\text{banana}} = \alpha e \cdot c_{\text{firetruck}} = 0.10 \cdot -0.500 \cdot (0.000, 0.000) = (-0.000, -0.000)$
- $\Delta c_{\text{firetruck}} = \alpha e \cdot w_{\text{banana}} = 0.10 \cdot -0.500 \cdot (0.046, 0.195) =$

NEG (banana vs firetruck)

- $z = w_{\text{banana}}^{\top} \cdot c_{\text{firetruck}} = 0.046 * 0.000 + 0.195 * 0.000 = 0.000$
- $e = 0.0 - \pi = 0.0 - \sigma(0.000) = -0.500$
- $\Delta w_{\text{banana}} = \alpha e \cdot c_{\text{firetruck}} = 0.10 \cdot -0.500 \cdot (0.000, 0.000) = (-0.000, -0.000)$
- $\Delta c_{\text{firetruck}} = \alpha e \cdot w_{\text{banana}} = 0.10 \cdot -0.500 \cdot (0.046, 0.195) = (-0.002, -0.010)$

Word		
ambulance	-0.228	0.099
apple	0.078	0.217
backhoe	-0.086	0.138
banana	0.046	0.195
crane	-0.220	0.153
firetruck	0.039	-0.047
lemon	0.008	-0.043
strawberry	0.202	-0.081

Context		
ambulance	0.000	0.000
apple	0.000	0.000
backhoe	-0.002	-0.010
banana	0.000	0.000
crane	0.000	0.000
firetruck	-0.002	-0.010
lemon	0.005	0.019
strawberry	0.000	0.000

$\alpha = 0.1$

Much later . . .

Vectors are starting to take shape

Word		
ambulance	-0.906	0.107
apple	0.992	0.780
backhoe	-0.902	0.459
banana	1.286	0.573
crane	-1.119	0.399
firetruck	-0.830	0.094
lemon	0.750	-0.289
strawberry	1.174	-0.379

Context		
ambulance	-0.927	-0.090
apple	0.973	-0.923
backhoe	-0.984	-0.379
banana	0.634	-0.486
crane	-1.258	-0.188
firetruck	-1.224	-0.060
lemon	1.087	-0.081
strawberry	1.054	0.410

$\alpha = 0.1$

POS (firetruck vs backhoe)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{backhoe}}$

POS (firetruck vs backhoe)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{backhoe}} = -0.830 * -0.984 + 0.094 * -0.379$

POS (firetruck vs backhoe)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{backhoe}} = -0.830 * -0.984 + 0.094 * -0.379 = 0.780$

POS (firetruck vs backhoe)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{backhoe}} = -0.830 * -0.984 + 0.094 * -0.379 = 0.780$
- $e = 1.0 - \pi = 1.0 - \sigma(0.780) =$

POS (firetruck vs backhoe)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{backhoe}} = -0.830 * -0.984 + 0.094 * -0.379 = 0.780$
- $e = 1.0 - \pi = 1.0 - \sigma(0.780) = 0.314$

POS (firetruck vs backhoe)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{backhoe}} = -0.830 * -0.984 + 0.094 * -0.379 = 0.780$
- $e = 1.0 - \pi = 1.0 - \sigma(0.780) = 0.314$
- $\Delta w_{\text{firetruck}} = \alpha e \cdot c_{\text{backhoe}} =$

POS (firetruck vs backhoe)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{backhoe}} = -0.830 * -0.984 + 0.094 * -0.379 = 0.780$
- $e = 1.0 - \pi = 1.0 - \sigma(0.780) = 0.314$
- $\Delta w_{\text{firetruck}} = \alpha e \cdot c_{\text{backhoe}} = 0.10 \cdot 0.314 \cdot (-0.984, -0.379) =$

POS (firetruck vs backhoe)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{backhoe}} = -0.830 * -0.984 + 0.094 * -0.379 = 0.780$
- $e = 1.0 - \pi = 1.0 - \sigma(0.780) = 0.314$
- $\Delta w_{\text{firetruck}} = \alpha e \cdot c_{\text{backhoe}} = 0.10 \cdot 0.314 \cdot (-0.984, -0.379) = (-0.031, -0.012)$

POS (firetruck vs backhoe)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{backhoe}} = -0.830 * -0.984 + 0.094 * -0.379 = 0.780$
- $e = 1.0 - \pi = 1.0 - \sigma(0.780) = 0.314$
- $\Delta w_{\text{firetruck}} = \alpha e \cdot c_{\text{backhoe}} = 0.10 \cdot 0.314 \cdot (-0.984, -0.379) = (-0.031, -0.012)$
- $\Delta c_{\text{backhoe}} = \alpha e \cdot w_{\text{firetruck}} =$

POS (firetruck vs backhoe)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{backhoe}} = -0.830 * -0.984 + 0.094 * -0.379 = 0.780$
- $e = 1.0 - \pi = 1.0 - \sigma(0.780) = 0.314$
- $\Delta w_{\text{firetruck}} = \alpha e \cdot c_{\text{backhoe}} = 0.10 \cdot 0.314 \cdot (-0.984, -0.379) = (-0.031, -0.012)$
- $\Delta c_{\text{backhoe}} = \alpha e \cdot w_{\text{firetruck}} = 0.10 \cdot 0.314 \cdot (-0.830, 0.094) =$

POS (firetruck vs backhoe)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{backhoe}} = -0.830 * -0.984 + 0.094 * -0.379 = 0.780$
- $e = 1.0 - \pi = 1.0 - \sigma(0.780) = 0.314$
- $\Delta w_{\text{firetruck}} = \alpha e \cdot c_{\text{backhoe}} = 0.10 \cdot 0.314 \cdot (-0.984, -0.379) = (-0.031, -0.012)$
- $\Delta c_{\text{backhoe}} = \alpha e \cdot w_{\text{firetruck}} = 0.10 \cdot 0.314 \cdot (-0.830, 0.094) = (-0.026, 0.003)$

NEG (firetruck vs crane)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{crane}}$

NEG (firetruck vs crane)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{crane}} = -0.830 * -1.258 + 0.094 * -0.188$

NEG (firetruck vs crane)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{crane}} = -0.830 * -1.258 + 0.094 * -0.188 = 1.025$

NEG (firetruck vs crane)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{crane}} = -0.830 * -1.258 + 0.094 * -0.188 = 1.025$
- $e = 0.0 - \pi = 0.0 - \sigma(1.025) =$

NEG (firetruck vs crane)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{crane}} = -0.830 * -1.258 + 0.094 * -0.188 = 1.025$
- $e = 0.0 - \pi = 0.0 - \sigma(1.025) = -0.736$

NEG (firetruck vs crane)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{crane}} = -0.830 * -1.258 + 0.094 * -0.188 = 1.025$
- $e = 0.0 - \pi = 0.0 - \sigma(1.025) = -0.736$
- $\Delta w_{\text{firetruck}} = \alpha e \cdot c_{\text{crane}} =$

NEG (firetruck vs crane)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{crane}} = -0.830 * -1.258 + 0.094 * -0.188 = 1.025$
- $e = 0.0 - \pi = 0.0 - \sigma(1.025) = -0.736$
- $\Delta w_{\text{firetruck}} = \alpha e \cdot c_{\text{crane}} = 0.10 \cdot -0.736 \cdot (-1.258, -0.188) =$

NEG (firetruck vs crane)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{crane}} = -0.830 * -1.258 + 0.094 * -0.188 = 1.025$
- $e = 0.0 - \pi = 0.0 - \sigma(1.025) = -0.736$
- $\Delta w_{\text{firetruck}} = \alpha e \cdot c_{\text{crane}} = 0.10 \cdot -0.736 \cdot (-1.258, -0.188) = (0.093, 0.014)$

NEG (firetruck vs crane)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{crane}} = -0.830 * -1.258 + 0.094 * -0.188 = 1.025$
- $e = 0.0 - \pi = 0.0 - \sigma(1.025) = -0.736$
- $\Delta w_{\text{firetruck}} = \alpha e \cdot c_{\text{crane}} = 0.10 \cdot -0.736 \cdot (-1.258, -0.188) = (0.093, 0.014)$
- $\Delta c_{\text{crane}} = \alpha e \cdot w_{\text{firetruck}} =$

NEG (firetruck vs crane)

- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{crane}} = -0.830 * -1.258 + 0.094 * -0.188 = 1.025$
- $e = 0.0 - \pi = 0.0 - \sigma(1.025) = -0.736$
- $\Delta w_{\text{firetruck}} = \alpha e \cdot c_{\text{crane}} = 0.10 \cdot -0.736 \cdot (-1.258, -0.188) = (0.093, 0.014)$
- $\Delta c_{\text{crane}} = \alpha e \cdot w_{\text{firetruck}} = 0.10 \cdot -0.736 \cdot (-0.830, 0.094) =$

NEG (firetruck vs crane)

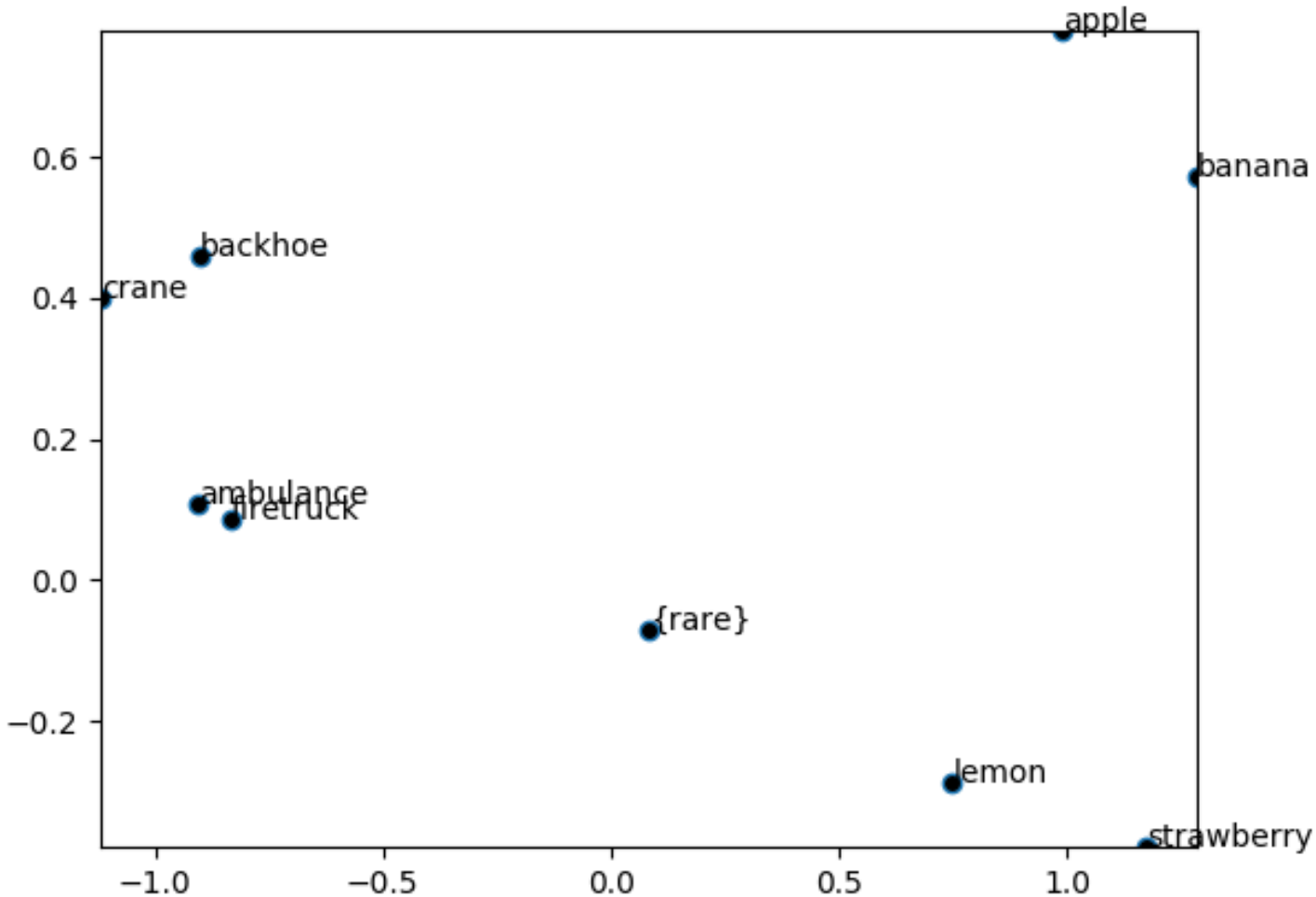
- $z = w_{\text{firetruck}}^{\top} \cdot c_{\text{crane}} = -0.830 * -1.258 + 0.094 * -0.188 = 1.025$
- $e = 0.0 - \pi = 0.0 - \sigma(1.025) = -0.736$
- $\Delta w_{\text{firetruck}} = \alpha e \cdot c_{\text{crane}} = 0.10 \cdot -0.736 \cdot (-1.258, -0.188) = (0.093, 0.014)$
- $\Delta c_{\text{crane}} = \alpha e \cdot w_{\text{firetruck}} = 0.10 \cdot -0.736 \cdot (-0.830, 0.094) = (0.061, -0.007)$

Word		
ambulance	-0.906	0.107
apple	0.992	0.780
backhoe	-0.902	0.459
banana	1.286	0.573
crane	-1.119	0.399
firetruck	-0.833	0.086
lemon	0.750	-0.289
strawberry	1.174	-0.379

Context		
ambulance	-0.927	-0.090
apple	0.973	-0.923
backhoe	-1.035	-0.373
banana	0.634	-0.486
crane	-1.196	-0.195
firetruck	-1.224	-0.060
lemon	1.110	-0.083
strawberry	1.054	0.410

$\alpha = 0.1$

Word Vectors



Context Vectors