FEATURE VECTORS

What is a vector? Consider the corpus "the dog chased the cat", "the cat chased the mouse", "the dog chased the squirrel", "the dog killed the chicken", "the cat killed the mouse", "the cat chased the cockroach". There are multiple stalks, here, but the ones of interest are the one for the dog:

```
the  dog chased the cat
the  dog chased the squirrel
the  dog killed the chicken
```

and the stalk for the cat:

```
the dog chased the  cat
                the  cat chased the mouse
                the  cat killed the mouse
                the  cat chased the cockroach
```

One old approach to data mining is to trim these down to 3-grams, and then compare them as feature vectors. These 3-gram feature vector for the dog is:

```
the  dog chased   ; 2 observations
the  dog killed   ; 1 observation
```

and the 3-gram stalk for the cat is:

```
chased the  cat           ; 1 observation
       the  cat chased    ; 2 observations
       the  cat killed    ; 1 observation
```

These are now explicitly vectors, as the addition of the observation count makes them so. The vertical alignment reminds us that they are also still stalks, and that the vector comes from collocations.

Recall how a vector is defined. One writes a vector $\vec{v}$ as a sum over basis elements $\hat{e}_i$with (usually real-number) coefficients a_i:

$$\vec{v} = \sum_i a_i \hat{e}_i$$

The basis elements $\hat{e}_i$ are unit-length vectors. Another common notation is the bra-ket notation, which says the same thing, but in a different way:

$$\vec{v} = \sum_i a_i |i\rangle$$

The bra-ket notation is slightly easier to use for this example. The above 3-gram collocations can be written as vectors. The one for dog would be

$$\overrightarrow{dog} = 2\,|the\ *\ chased\rangle + |the\ *\ killed\rangle$$

while the one for cat would be

$$\overrightarrow{cat} = |chased\ the\ *\rangle + 2\,|the\ *\ chased\rangle + |the\ *\ killed\rangle$$

Date: 11 November 2017.

The $*$ here is the wild-card; it indicates where "dog" and "cat" should go, but it also indicates how the basis vectors should be treated: the wild-card helps establish that dogs and cats are similar. The basis vectors indicate how dot products can be taken. Recall the defintion of a dot-product (the inner product). For $\vec{v}$ as above, and $\vec{w} = \sum_i b_i \hat{e}_i$, one has that

$$\vec{v} \cdot \vec{w} = \sum_i \sum_j a_i b_j \hat{e}_i \cdot \hat{e}_j = \sum_i \sum_j a_i b_j \delta_{ij} = \sum_i a_i b_i$$

where the Kronecker delta was used in the middle term:

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Thus, the inner prioduct of $\overrightarrow{cat}$ and $\overrightarrow{dog}$ can be computed:

$$\overrightarrow{cat} \cdot \overrightarrow{dog} = 0 \cdot 1 + 2 \cdot 2 + 1 \cdot 1 = 5$$

One common way to express the similarity of $\overrightarrow{cat}$ and $\overrightarrow{dog}$ is to compute the cosine similarity. The angle θ between two vectors is given by

$$\cos\theta = \vec{v} \cdot \vec{w} / |\vec{v}| \, |\vec{w}|$$

where $|\vec{v}|$ is the length of $\vec{v}$. Since $\left|\overrightarrow{cat}\right| = \sqrt{6}$ and $\left|\overrightarrow{dog}\right| = \sqrt{5}$ one finds that

$$\cos\theta = \frac{5}{\sqrt{30}} \approx 0.913$$

That is, dogs and cats really are similar.

If one was working with a dependency parse, as opposed to 3-grams, and if one used the Frobenius algebra notation such as that used by Kartsaklis in [1], then one would write the basis elements as a peculiar kind of tensor, and one might arrive at an expression roughly of the form

$$\overline{dog} = 2 \left(\overleftarrow{the} \otimes \overrightarrow{chased} \right) + 1 \left(\overleftarrow{the} \otimes \overrightarrow{killed} \right)$$

and

$$\overline{cat} = \left(\overleftarrow{chased} \otimes \overleftarrow{the} \right) + 2 \left(\overleftarrow{the} \otimes \overrightarrow{chased} \right) + 1 \left(\overleftarrow{the} \otimes \overrightarrow{killed} \right)$$

Ignoring the differnces in notation (ignoring that the quantities in parenthesis are tensors), one clearly can see that these are still feature vectors. Focusing on the vector aspect only, these represent the same information as the 3-gram feature vectors. They're the same thing. The difference between them is that the bra-ket notation was used for the 3-grams, while the tensor notation was used for the dependency parse. The feature vectors can also be written using the link-grammar-inspired notation:

```
dog: [the- & chased+]2 or [the- & killed+]1;
cat: [chased- & the-]1 or [the- & chased+]2 or [the- & killed+]1;
```

The notation is different, the meaning is the same. The above gives two feature vectors, one for dog, and one for cat. The fact that they just happen to look identical to the 3-gram feature vectors is only because a link-grammar parse of such short phrases look a lot like ordinary 3-grams. For more complex sentences, the 3-gram approach, and the dependency-parse approach give different feature vectors. The concept of feature vectors remain the same.

REFERENCES

[1] Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. A study of entanglement in a categorical framework of natural language. *Proceedings Quantum Physics and Logic, Electronic Proceedings in Theoretical Computer Science*, 172:249–260, 2014.