School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 2, 2017)
Workshop exercises: Week 3

1. Finish any remaining questions from last week, if necessary.

&

---

2. What is a **vector space model**, and why is it useful?

   - We treat each of the possible events that can occur as a separate (orthogonal) "dimension" in a Cartesian co-ordinate space. It isn't really important what this means, except that our representation of an instance can be expressed as a list of values according to the events.
   - In a text processing context, each "instance" is a document, and each "event" is (usually) the occurrence of an individual word. The values that these "features" can take are (usually) either binary (i.e. Is the word present in the document, or not?) or are frequencies (i.e. How many times is the word present in the document?).
   - Another important consideration is that the documents are usually pre-processed in a manner to reduce the number of possible "words": for example, by folding case, or removing punctuation, or removing very common words ("stopping"), and so on.
   - This representation is useful because it allows us to represent a complex document in a simplified manner, to expedite the calculation of similarity measures.

3. Consider the following collection of "documents" $\mathcal{C}$:

   (*i*) *It is what it is.*

   (*ii*) *Jean's hat is finer than Karl's hat.*

   (*iii*) *We are obsessing about gene issues.*

   Build a vector space model for this collection.

   - What the vector space looks like depends a lot on what pre-processing we do to these documents. In this case, we am going to fold case, remove punctuation, and also remove the `'s` clitics. We're going to leave "stopwords" in the representation, and we won't canonicalise the tokens further (more on this in later weeks).
   - Note that different choices in pre-processing would lead to different estimates of similarity.
   - Here's the word list (of 14 elements), which will correspond to each element in the vectors below:
     `about`, `are`, `finer`, `gene`, `hat`, `is`, `issues`, `it`, `jean`, `karl`, `obsessing`, `than`, `we`, `what`
     - (`i`): 0,0,0,0,0,2,0,2,0,0,0,0,0,1
     - (`ii`): 0,0,1,0,2,1,0,0,1,1,0,1,0,0
     - (`iii`): 1,1,0,1,0,0,1,0,0,0,1,0,1,0

4. Use a model to decide which of the sentences in $\mathcal{C}$ is most similar to the following sentence:

   (*iv*) *Karl is obsessed with genes.*

   - The first thing to do is to represent this document in the same format as the documents above. You'll notice that we have an immediate problem here, because three of these (five) words aren't in our word list above (because we haven't seen them yet!).
   - To make my life a little easier, I'll tack them on at the end, so that the vectors above just need 3 more zeroes on the right-hand side. Here's the representation of the four "documents" in table form: (We call this the "document–term matrix".)

	ab	ar	f	g	h	is	iss	it	j	k	obing	t	we	wh	obed	wi	gs
(*i*)	0	0	0	0	0	2	0	2	0	0	0	0	0	1	0	0	0
(*ii*)	0	0	1	0	2	1	0	0	1	1	0	1	0	0	0	0	0
(*iii*)	1	1	0	1	0	0	1	0	0	0	1	0	1	0	0	0	0
(*iv*)	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	1	1

Based on the following metrics:

(a) Jaccard similarity

- Recall that the definition of Jaccard similarity is:

$$\text{sim}(A, B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid}$$

- Here, we are using set intersection, so that we don't care about the frequencies of words, only their presence or absence.
- For (*i*), the two sentences only have a single word in common (`is`; you can confirm this by looking for elements that are non-zero in both vectors). The union of the two sentences is the count of all of words that occur in either sentence (no double counting!): this is 7 (the five words in the latter sentence, plus `it` and `what`). All in all, the similarity is $\frac{1}{7}$.
- For (*ii*), there are two words in common (`karin` and `is`). There are 9 words in the union, for a similarity of $\frac{2}{9}$.
- For (*iii*), there aren't any words shared between the two sentences, so the similarity is 0.
- All in all, (*ii*) is the most similar.

(b) the Dice coefficient

- Recall that the definition of Dice is:

$$\text{sim}(A, B) = \frac{2 * \mid A \cap B \mid}{\mid A \mid + \mid B \mid}$$

- This is quite similar to the formula above; let's just work through it quickly:
- For (*i*), we have $\frac{2*1}{3+5} = \frac{2}{8}$.
- For (*ii*), we have $\frac{2*2}{6+5} = \frac{4}{11}$.
- For (*iii*), we have $\frac{0}{6+5} = 0$.
- Again, (*ii*) is the most similar.

(c) Euclidean distance

- This time, we're talking about a distance, rather than a similarity. The most similar text is the one with the smallest distance. Recall the definition of Euclidean distance:

$$d(A, B) = \sqrt{\sum_i (A_i - B_i)^2}$$

- For (*i*), this is: $\sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (2-1)^2 + \cdots}$ $\sqrt{\cdots + (0-0)^2 + (2-0)^2 + (0-0)^2 + (0-1)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + \cdots}$ $\sqrt{\cdots + (1-0)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2} = \sqrt{10}$.
- Notice how the frequencies are important now. To save space, I will ignore all of the $(0-0)^2$ terms below:
- For (*ii*), this is $\sqrt{(1-0)^2 + (2-0)^2 + (1-1)^2 + (1-0)^2 + (1-1)^2 + (1-0)^2 + \cdots}$ $\sqrt{\cdots + (0-1)^2 + (0-1)^2 + (0-1)^2} = \sqrt{10}$
- For (*iii*), this is $\sqrt{(1-0)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2 + \cdots}$ $\sqrt{\cdots + (1-0)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2} = \sqrt{11}$
- So, (*i*) and (*ii*) are tied for the best similarity, but (*iii*) is actually very close this time (effectively, none of them are very similar `(-:`).

(d) Manhattan distance

- As above, but without the root and squaring of terms. The three distances are 8, 8, and 11 respectively.

(e) Cosine similarity

- Recall that the formula for cosine similarity is as follows:

$$cos(A, B) = \frac{\vec{a} \cdot \vec{b}}{| \vec{a} || \vec{b} |}$$

- The length of a vector is found by taking the square root of the sum of the squares of its entries (effectively finding the Euclidean distance between it and the zero vector, consisting entirely of zeroes). Leaving out the 0^2 terms, the vector lengths are as follows:

$$\begin{aligned}
| (i) | &= \sqrt{2^2 + 2^2 + 1^2} = \sqrt{9} \\
| (ii) | &= \sqrt{1^2 + 2^2 + 1^2 + 1^2 + 1^2 + 1^2} = \sqrt{9} \\
| (iii) | &= \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2} = \sqrt{6} \\
| (iv) | &= \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2} = \sqrt{5}
\end{aligned}$$

- We find the dot product by adding up the products of the values for each of the corresponding elements; I'll ignore $0 \cdot 0$ terms below:

$$\begin{aligned}
(\vec{i}) \cdot (\vec{iv}) &= 2 \cdot 1 + 2 \cdot 0 + 0 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 + 0 \cdot 1 + 0 \cdot 1 = 2 \\
(\vec{ii}) \cdot (\vec{iv}) &= 1 \cdot 0 + 2 \cdot 0 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 + 0 \cdot 1 + 0 \cdot 1 = 2 \\
(\vec{iii}) \cdot (\vec{iv}) &= 1 \cdot 0 + 1 \cdot 0 + 1 \cdot 0 + 0 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 + 1 \cdot 0 + 1 \cdot 0 + 0 \cdot 1 + 0 \cdot 1 + 0 \cdot 1 = 0
\end{aligned}$$

- Putting it all together:

$$\begin{aligned}
cos((\vec{i}), (\vec{iv})) &= \frac{(\vec{i}) \cdot (\vec{iv})}{| (\vec{i}) || (\vec{iv}) |} \\
&= \frac{2}{\sqrt{9}\sqrt{5}} \approx 0.298 \\
cos((\vec{ii}), (\vec{iv})) &= \frac{(\vec{ii}) \cdot (\vec{iv})}{| (\vec{ii}) || (\vec{iv}) |} \\
&= \frac{2}{\sqrt{9}\sqrt{5}} \approx 0.298 \\
cos((\vec{iii}), (\vec{iv})) &= \frac{(\vec{iii}) \cdot (\vec{iv})}{| (\vec{iii}) || (\vec{iv}) |} \\
&= \frac{0}{\sqrt{6}\sqrt{5}} = 0
\end{aligned}$$

- Once more, (i) and (ii) are tied as being the most similar (substantially more so than (iii) this time).

&

Consider the following dataset, representing a collection of documents:

ID	*apple*	*ibm*	*lemon*	*sun*	LABEL
A	4	0	1	1	FRUIT
B	5	0	5	2	FRUIT
C	2	5	0	0	COMP
D	1	2	1	7	COMP

5. If we wished to estimate the value of $P(\textit{apple}, \text{FRUIT})$ based on the data above, we might arrive at the value $\frac{2}{4}$, or the value $\frac{9}{36}$.

   (a) What do each of these probabilities correspond to? Explain the model that underlies each of these interpretations.

   - $\frac{2}{4}$ refers to the **document**–based model; i.e. how many of the 4 documents are FRUIT, and contain (at least one) *apple*? $\frac{9}{36}$ refers to the **term**–based model; i.e. how many of the tokens that we have seen (36 in total) are *apple* in FRUIT documents?

   (b) Which of these do you think would be more useful, if we wished to use these probabilities to help predict the labels of unknown documents?

   - Ultimately, this is an empirical question — which of these two methods gives us a better predictions *on some actual data*?
   - Generally speaking, though, the document–based model is preferred. This can be seen for a couple of reasons:
     - Our objective here is to predict the labels of documents; therefore, our model should be based around documents. If we wished to instead learn the most likely value for *apple* in a previously-unseen FRUIT document (which we sometimes wish to do!), then the term-based model would probably be better.
     - The term–based model is sensitive to the actual values that we have in our dataset; if there was some instance with disproportionately large values (say, the value of *apple* is 1000, for example), that will bias our estimates, so that the other instances don't contribute to the estimated probabilities.

6. Calculate the **entropy** of the distribution of the LABEL attribute, based on the data above.

   - Recall that the entropy (in bits) of a distribution is defined as:

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x)$$

   - Here, we are interested in the distribution of the LABEL attribute, which takes two values: FRUIT and COMP.
   - We begin by calculating the probabilities of the events: $P(\text{FRUIT}) = \frac{2}{4}$ and $P(\text{COMP}) = \frac{2}{4}$
   - Now, we simply plug into the formula:

$$\begin{aligned} H(\text{LABEL}) &= -[P(\text{FRUIT}) \log_2 P(\text{FRUIT}) + P(\text{COMP}) \log_2 P(\text{COMP})] \\ &= -[\frac{2}{4} \log_2 \frac{2}{4} + \frac{2}{4} \log_2 \frac{2}{4}] \\ &= -[(0.5)(-1) + (0.5)(-1)] = 1 \text{ bit} \end{aligned}$$