School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 2, 2019)

Workshop sample solutions: Week 3

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

---

2. Consider the following collection of "documents" *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 feature vector for all the documents in this collection.

   A vector space depends on how the pre-processing carried out.

   In this example, the pre-processing steps include case-folding (change all uppercases to lowercase), removing punctuations, and also removing " 's " clitics.

   The other possible options in pre-processing documents could be removing the "stop words" (words such as: and, the, a, an, …), however please note that we do NOT performing this step for this example.

   Please note that different choices in pre-processing would lead to different estimates of similarity.

   The following is the list of all the words/tokens (the vector space) in *C* in alphabetic order.

   about, are, finer, gene, hat, is, issues, it, jean, karl, obsessing, than, we, what

   We use this vector space to develop the feature vector for all the documents using a "document-term matrix":

	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

   Leading to the following feature vectors:

   (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>

3. Based on the following metrics decide which of the "documents" in *C* is most similar to (*iv*) "document"

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

   The first thing to do is to represent this document in the same format as the other documents in *C*. You'll notice that we have an immediate problem: three of these (five) words aren't in the space vector (because we haven't seen them yet!).

   To make our lives a little easier, we can append them at end of the space vector, so that the feature vectors would need another 3 more zeroes on the right-hand side. Accordingly, the

table would change to:

	about	are	finer	gene	hat	is	issues	it	jean	karl	Obsessing	than	we	what	Obssed	with	genes
(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

Leading to the following feature vectors:

(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>

Using these vectors would enable us to calculate the similarity between (iv) and the documents in C.

(a) Euclidean distance

Recall the definition of Euclidean distance:

$$d(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

The Euclidean distance between (i) and (iv) can be calculated as follows:

$$d(i,iv) = \sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (2-1)^2 + (0-0)^2 + (2-0)^2 + \sqrt{+(0-0)^2 + (0-1)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2} \sqrt{+(0-1)^2} = \sqrt{10}$$

Similarly, the distance between the rest of documents in *C* and (iv) can calculated as follows:

$$d(ii,iv) = \sqrt{(0-0)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2 + (2-0)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2 + \sqrt{+(1-0)^2 + (1-1)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2 + (0-1)^2} \sqrt{+(0-1)^2} = \sqrt{10}$$

$$d(iii,iv) = \sqrt{(1-0)^2 + (1-0)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2 + (0-1)^2 + (1-0)^2 + (0-0)^2 + \sqrt{+(0-0)^2 + (0-1)^2 + (1-0)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2 + (0-1)^2 + (0-1)^2} \sqrt{+(0-1)^2} = \sqrt{11}$$

So both (i) and (ii) have the distance of $\sqrt{10}$ from (iv), which mean they are "closer" or more similar to (iv) in compare with (iii) that shows the distance of $\sqrt{11}$ from (iv).

(b) Cosine similarity

Recall the definition of Cosine Similarity:

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

The length of a vector can be calculated by taking the square root of the sum of entries squared:

$$|\vec{a}| = \sqrt{\sum_{i=1}^{n} a_n^2}$$

The length of the four vectors can be calculated as follows (0^2 terms neglected):

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

The dot product could be calculated by adding up the products of the values for each of the corresponding elements (0 . 0 terms neglected)

$$\overrightarrow{(i)} \ . \ \overrightarrow{(iv)} = 2 \cdot 1 + 2 \cdot 0 + 0 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 + 0 \cdot 1 + 0 \cdot 1 = 2$$
$$\overrightarrow{(ii)} \ . \ \overrightarrow{(iv)} = 1 \cdot 0 + 2 \cdot 0 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 + 0 \cdot 1 + 0 \cdot 1 = 2$$
$$\overrightarrow{(iii)} \ . \ \overrightarrow{(iv)} = 1 \cdot 0 + 1 \cdot 0 + 1 \cdot 0 + 0 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 + 1 \cdot 0 + 1 \cdot 0 + 0 \cdot 1 + 0 \cdot 1 + 0 \cdot 1 = 0$$

Putting it all together we will have:

$$\cos(\overrightarrow{(i)}, \overrightarrow{(iv)}) = \frac{\overrightarrow{(i)} \ . \ \overrightarrow{(iv)}}{|\overrightarrow{(i)}||\overrightarrow{(iv)}|} = \frac{2}{\sqrt{9}\sqrt{5}} \approx 0.298$$

$$\cos(\overrightarrow{(ii)}, \overrightarrow{(iv)}) = \frac{\overrightarrow{(ii)} \ . \ \overrightarrow{(iv)}}{|\overrightarrow{(ii)}||\overrightarrow{(iv)}|} = \frac{2}{\sqrt{9}\sqrt{5}} \approx 0.298$$

$$\cos(\overrightarrow{(iii)}, \overrightarrow{(iv)}) = \frac{\overrightarrow{(iii)} \ . \ \overrightarrow{(iv)}}{|\overrightarrow{(iii)}||\overrightarrow{(iv)}|} = \frac{0}{\sqrt{6}\sqrt{5}} \approx 0$$

Based on these calculations we can conclude that (i) and (ii) are more similar to (iv) in comparison with (iii).

(c) Jaccard Similarity

Recall that the definition of Jaccard similarity is:

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

Here, we are using set intersection: in this method we don't consider frequencies of words, but only their presence or absence.

- In calculating the similarity between *(i)* and *(iv)*, we can see that the two sentences only have a single word in common (`is`), so their intersection is equal to 1. The union of the two sentences is the count of all of words that occur in either sentences (no double counting!). So the union of *(i)* and (iv) is 7 (the five words in the *(iv)*, plus '`it`' and '`what`' from *(i)*). Therefore, $sim((i), (iv)) = \frac{1}{7}$.
- In calculating the similarity between *(ii)* and *(iv)*, there are two words in common ('`karl`' and '`is`'). There are 9 words in the union of *(ii)* and *(iv)*, so $sim((i), (iv)) = \frac{2}{9}$.
- In calculating the similarity between *(iii)* and *(iv)*, there aren't any words shared between the two sentences, so the similarity is 0.

Putting it all together using Jaccard similarity, (ii) is the most similar document to (iv).

4. Explain the difference between *Distance* and *Similarity* calculations in question 3.

When using a Distance function (such as Euclidean distance or Manhattan distance), we are looking for the MINIMUM distance between the documents. In other words the documents that have the shorter/smaller distance from each other are considered more similar.

However, when using a Similarity function (such as Cosine similarity or Jaccard similarity), we are looking for the MAXIMUM similarity between the documents.

&

---

5. Consider we have two coins, one *fair* coin and another one with two *heads*. Both coins are in a bag. In our test trial, we randomly select a coin from the bag, toss it and check the results.

   (a) Using the Bayes Rules, calculate the prior and posterior probability of choosing the fair coin in this experiment. (In your calculation, consider both possibilities of observing a *head* or a *tail* in our test trial).

Recall Bayesian Rule from lectures:

$$P(A|B) = \frac{P(A).P(B|A)}{P(B)}$$

Where

- P(A) is the Prior Probability
- P(B|A) is our data likelihood
- P(B) is our evidence

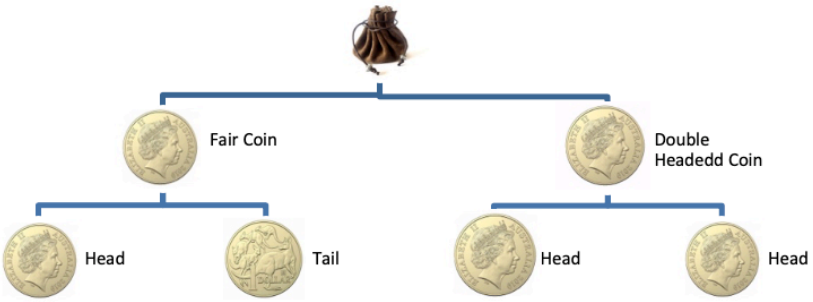

Figure 1

If we observe a **Tail** in our trial experiment, it is obvious that we have chosen the fair coin in our trial (why?).

Now let's calculate the probability of chossing the fair coin in the situation that we observe a **Head**.

$$P(Fair|Head) = \frac{P(Fair).P(Head|Fair)}{P(Head)}$$

Using the tree of possibilities (figure 1), we can see that P(Head) from all the possibilities (leaves of the tree) is 3 out of 4. We also know that we only have 2 coins so probability of choosing the fair one is 1 out of 2 and if we have chosen a fair coin the probability of observing a head is again 1 out of 2. Putting it all together, we will have:

$$P(Fair|Head) = \frac{\frac{1}{2}.\frac{1}{2}}{\frac{3}{4}} = \frac{1}{3}$$

(b) We repeat the experience again. Calculate the probability that we have chosen the fair coin (in both experiments) if we observe two *Heads* in a row.

Approach 1:

For analyzing the possibility of observing two heads in a row we need to extend our tree of possibilities one more level (figure 2).

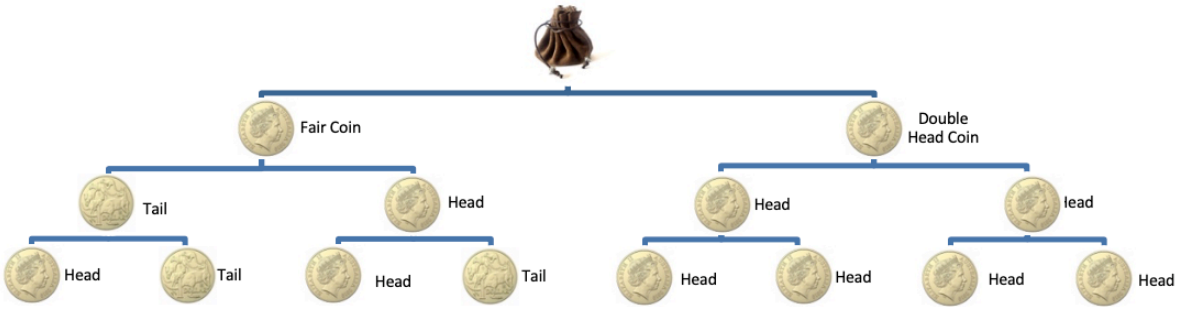

Figure 2

Using the Bayes Rules we can re-write the posterior probability of observing two Heads as follows:

$$P(Fair|Head_1, Head_2) = \frac{P(Fair).P(Head_1, Head_2|Fair)}{P(Head_1, Head_2)}$$

In this formula Probability of observing two heads in a row if we have chosen the fair coin is 1 out of 4 (why?). And the probability of having two heads in total is the number of the heads in all our leaves, which is 5 out of 8. Also, we have our Prior knowledge (that one of the two coins is fair). So putting it all together we will have:

$$P(Fair|Head_1, Head_2) = \frac{\frac{1}{2}.\frac{1}{4}}{\frac{5}{8}} = \frac{1}{5}$$

Approach 2:

Another way to approach this problem is to consider the outcome of our first experiment as a prior knowledge and calculate the posterior probability of our second experiment based on that. So basically in this approach we assume that in the first experiment we have observed a Head (prior knowledge) and we want to calculate the probability of choosing a Fair coin in the second experiment. So the tree will be trimmed as shown in figure 3.

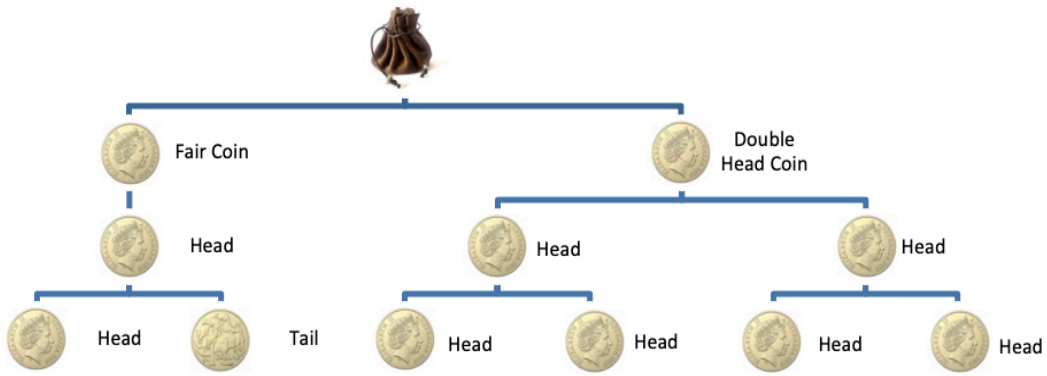

Figure 3

Based on this assumption, the posterior probability of choosing the *Fair coin when the first observation is Head* (Fair*) will be:

$$P(Fair^*|Head_2) = \frac{P(Fair^*).P(Head_2|Fair)}{P(Head_2)}$$

P(Fair*) is the outcome of calculations from part (a) (probability of choosing the fair coin given the observing a Head), which is $\frac{1}{3}$.

Using the tree, it can be observed that probability of observing a Head having a Fair coin ($P(Head_2|Fair)$) is $\frac{1}{2}$; and probability of observing a head in total (all the shown leaves) is $\frac{5}{6}$.

Putting it all together we will have:

$$P(Fair^*|Head_2) = \frac{\frac{1}{3}.\frac{1}{2}}{\frac{5}{6}} = \frac{1}{5}$$

6. Calculate the **entropy** for our first trial in question 5.

Recall the definition of entropy:

$$H(X) = -\sum_i p(x_i)\log_2 p(x_i)$$

In our first trial in question 5 the entropy of running the trial can be defined as follow:

$$H(X) = -[p(Head)\log_2 p(Head) + p(Tail)\log_2 p(Tail)]$$

$$= -\left[\frac{3}{4}\times\log_2\frac{3}{4} + \frac{1}{4}\times\log_2\frac{1}{4}\right]$$

$$= -[0.75\times(-0.42) + 0.25\times(-2)]$$

$$= 0.815 \quad \text{(less than 1 bit)}$$