ID 2088192

Question 20233

After inserting your student ID and the question number in the title, header and footer, write your answers between here and the statement of good academic conduct. Your ID and the question number will automatically appear on any subsequent pages.

(a)

(i) The IDF for each word are calculated as:

$$IDF(\text{challenge}) = \ln(6/1) \approx 1.79$$
$$IDF(\text{child}) = \ln(6/1) \approx 1.79$$
$$IDF(\text{educate}) = \ln(6/4) \approx 0.41$$
$$IDF(\text{health}) = \ln(6/4) \approx 0.41$$
$$IDF(\text{home}) = \ln(6/2) \approx 1.10$$
$$IDF(\text{minister}) = \ln(6/1) \approx 1.79$$
$$IDF(\text{need}) = \ln(6/1) \approx 1.79$$
$$IDF(\text{parent}) = \ln(6/2) \approx 1.10$$
$$IDF(\text{priority}) = \ln(6/2) \approx 1.10$$
$$IDF(\text{school}) = \ln(6/1) \approx 1.79$$
$$IDF(\text{system}) = \ln(6/3) \approx 0.69$$
$$IDF(\text{uk}) = \ln(6/2) \approx 1.10$$

(ii) The document lengths of the query Q and 6 texts in the corpus are:

$$\|Q\| = \sqrt{0.41^2 + 0.69^2} \approx 0.80$$
$$\|D_1\| = \sqrt{0.41^2 + 0.41^2 + 1.10^2 + 0.69^2} \approx 1.42$$
$$\|D_2\| = \sqrt{0.41^2 + 0.41^2 + 1.10^2 + 1.79^2} \approx 2.18$$
$$\|D_3\| = \sqrt{5.38^2 + 0.41^2 + 1.10^2 + 1.10^2} \approx 5.62$$
$$\|D_4\| = \sqrt{1.79^2 + 1.10^2 + 1.10^2 + 1.79^2} \approx 2.97$$
$$\|D_5\| = \sqrt{0.41^2 + 0.41^2 + 1.79^2 + 1.38^2 + 1.10^2} \approx 2.58$$
$$\|D_6\| = \sqrt{0.41^2 + 0.69^2 + 1.10^2} \approx 1.36$$

(continued)

Thus the TF-IDF similarity between Q and the documents are:

$$Sim(Q, D_1) = \frac{0.41^2 + 0.69^2}{0.80 \times 1.42} \approx 0.57$$
$$Sim(Q, D_2) = \frac{0.41^2}{0.80 \times 2.18} \approx 0.096$$
$$Sim(Q, D_3) = 0$$
$$Sim(Q, D_4) = 0$$
$$Sim(Q, D_5) = \frac{0.41^2 + 0.69 \times 1.38}{0.80 \times 2.58} \approx 0.54$$
$$Sim(Q, D_6) = \frac{0.41^2 + 0.69^2}{0.80 \times 1.36} \approx 0.59$$

(b)

(i) The d_∞ distance between data points and two centroids are:

$$d_\infty(x_1, c_1) = \max(|4-1|, |2-2|) = 3$$
$$d_\infty(x_2, c_1) = \max(|-1-1|, |2-2|) = 2$$
$$d_\infty(x_3, c_1) = \max(|-2-1|, |1-2|) = 3$$
$$d_\infty(x_4, c_1) = \max(|3-1|, |1-2|) = 2$$
$$d_\infty(x_5, c_1) = \max(|-2-1|, |2-2|) = 3$$

$$d_\infty(x_1, c_2) = \max(|4-2|, |2-1|) = 2$$
$$d_\infty(x_2, c_2) = \max(|-1-2|, |2-1|) = 3$$
$$d_\infty(x_3, c_2) = \max(|-2-2|, |1-1|) = 4$$
$$d_\infty(x_4, c_2) = \max(|3-2|, |1-1|) = 1$$
$$d_\infty(x_5, c_2) = \max(|-2-2|, |2-1|) = 4$$

Assign the data points to the closest centroid in d_∞ distance:

$$\min(d_\infty(x_1, c_1), d_\infty(x_1, c_2)) = 2$$
$$\min(d_\infty(x_2, c_1), d_\infty(x_2, c_2)) = 2$$
$$\min(d_\infty(x_3, c_1), d_\infty(x_3, c_2)) = 3$$
$$\min(d_\infty(x_4, c_1), d_\infty(x_4, c_2)) = 1$$
$$\min(d_\infty(x_5, c_1), d_\infty(x_5, c_2)) = 3$$

x_2, x_3, x_5 assigned to c_1, x_1, x_4 assigned to c_2.

Update the values of new centroid $\bar{c}_1$, $\bar{c}_2$:

$$\bar{c}_1 = \begin{bmatrix} [(-1)+(-2)+(-2)]/3 \\ (2+1+2)/3 \end{bmatrix} = \begin{bmatrix} -1.67 \\ 1.67 \end{bmatrix}$$
$$\bar{c}_2 = \begin{bmatrix} (4+3)/2 \\ (2+1)/2 \end{bmatrix} = \begin{bmatrix} 3.5 \\ 1.5 \end{bmatrix}$$

(c)

(i) The d_∞ distance between the data point x and centroids:

$$d_\infty(c_1, x) = \max(|1-4|, |2-2|) = 3$$
$$d_\infty(c_2, x) = \max(|2-4|, |1-2|) = 2$$
$$d_\infty(c_3, x) = \max(|3-4|, |2-2|) = 1$$

The closest centroid to x is c_3 , i.e. $i(x) = 3$.

The new values of three centroids are calculated as:

$$c^1_{new} = \begin{bmatrix}1\\2\end{bmatrix} + e^{\frac{-(3-1)^2}{10}} \times 0.5 \times (\begin{bmatrix}4\\2\end{bmatrix} - \begin{bmatrix}1\\2\end{bmatrix}) = \begin{bmatrix}2.01\\2\end{bmatrix}$$
$$c^2_{new} = \begin{bmatrix}2\\1\end{bmatrix} + e^{\frac{-(3-2)^2}{10}} \times 0.5 \times (\begin{bmatrix}4\\2\end{bmatrix} - \begin{bmatrix}2\\1\end{bmatrix}) = \begin{bmatrix}2.9\\1.45\end{bmatrix}$$
$$c^1_{new} = \begin{bmatrix}3\\2\end{bmatrix} + e^{\frac{-(3-3)^2}{10}} \times 0.5 \times (\begin{bmatrix}4\\2\end{bmatrix} - \begin{bmatrix}3\\2\end{bmatrix}) = \begin{bmatrix}3.5\\2\end{bmatrix}$$

(ii) (1) For $i(x) = j$: the function $h(j,k) \to 1$.

Rewrite the SOM update rule as:

$$c^j_{new} = c^j_{old} + \eta \times (x - c^j_{old})$$
$$= (1-\eta)c^j_{old} + \eta x$$

(2) For $i(x) \neq j$:

When $\sigma \to 0$, we easily have $\frac{-(j-k)^2}{\sigma} \to -\infty$.

Thus the neighbourhood indicator function $h(j,k) = e^{\frac{-(j-k)^2}{\sigma}} \to 0$.

Rewrite the SOM update rule as:

$c^j_{new} = c^j_{old}$, leaving out the zero part.

Do not write below this line

Statement of good academic conduct

By submitting this assignment, I understand that I am agreeing to the following statement of good academic conduct:

- I confirm that this assignment is **my own work** and I have not worked with others in preparing this assignment.
- I confirm this assignment was written by me and is in my own words, except for any materials from published or other sources which are clearly indicated and acknowledged as such by appropriate referencing.
- I confirm that this work is not copied from any other person's work (published or unpublished), web site, book or other source, and has not previously been submitted for assessment either at the University of Birmingham or elsewhere.
- I confirm that I have not asked, or paid, others to prepare any part of this work for me.
- I confirm that I have read and understood the University regulations on plagiarism (https://intranet.birmingham.ac.uk/as/registry/policy/conduct/plagiarism/index.aspx).