NLP Exam answer key

Yutong Wanyan

wanyan-yutong-exam.pdf

Short Answers

1. ① word2vec can better handle high dimensionality

   ② count vectorization is very labor-intensive.

2. B. Option C is redundant in terms of coding efficiency; Option A is not right for txt. file.

3. the features we get from n-gram model are more likely to contain meaningful natural languages.

4. This pattern doesn't specify word boundary. A more precise way: r'\bson\b'

5.

6. Recall. As a high recall indicates more actual positives are correctly predicted, so more relevant baby care products will be returned

7.

Naive Bayes:

(A).

(i). $P(y = \text{below}) = \frac{7}{10}$

(ii). $P(x_1 = \text{Green}, x_2 = \text{Wool}, x_3 = \text{Clothing} \mid y = \text{above})$

$= P(x_1 \mid y \text{ above}) \times P(x_2 \mid y \text{ above}) \times P(x_3 \mid y \text{ above})$

$= \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} = 0.074$

(iii).

$P(x_1, x_2, x_3) = P(x \cdots \mid y \text{ above}) \cdot P(y \text{ above})$

$+$

$P(x \cdots \mid y \text{ below}) \cdot P(y \text{ below})$

$P(x \cdots \mid y \text{ below}) = \frac{1}{7} \times \frac{1}{7} \times \frac{3}{7}$

$= \frac{3}{343}$

I didn't group Green and recycleable together as there is another color word "blue".

$P(y_{\text{below}}) = \frac{7}{10}$, $P(y \text{ above}) = \frac{3}{10}$

Therefore:

$P(x_1, x_2, x_3) = 0.074 \times \frac{3}{10} + \frac{3}{343} \times \frac{7}{10}$

$= 0.0283$

(iv). $P(y \text{ above} \mid x_1 x_2 x_3) = \frac{P(x_1 x_2 x_3 \mid y \text{ above}) \cdot P(y \text{ above})}{P(x_1 x_2 x_3)}$

$= \frac{0.074 \times \frac{3}{10}}{0.0283}$

$= 0.784$

Vectorization and Similarity

horror = thriller
animated = animation

(a).

	comedy	G-rated	M.B.J.	S.B.	NYC	animation	adventure	action	LA	romantic	M.D.	horror
idf	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	1	1
tf(A)	1	1	1	1	1	0	0	0	0	0	0	0
tf(B)	0	1	0	0	0	1	1	1	1	0	0	0
tf(C)	1	0	0	1	1	1	0	0	0	1	0	0
tf(D)	0	0	0	0	0	0	1	1	1	0	1	2
tf idf(A)	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0	0	0	0
tf idf(B)	0	$\frac{1}{2}$	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
tf idf(C)	$\frac{1}{2}$	0	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	1	0	0
tf idf(D)	0	0	0	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	1	2

(b). query: G-rated action comedy

q	1	1	0	0	0	0	0	1	0	0	0	0
tf idf(q)	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0	0	$\frac{1}{2}$	0	0	0	0

$$\|A\| = \sqrt{\frac{1}{2}^2 + \frac{1}{2}^2 + \frac{1}{2}^2 + \frac{1}{2}^2} = 1$$

$$\|B\| = \sqrt{\frac{1}{2}^2 + \frac{1}{2}^2 + \frac{1}{2}^2 + \frac{1}{2}^2 + \frac{1}{2}^2} = \frac{\sqrt{5}}{2}$$

$$\|q\| = \sqrt{\frac{1}{2}^2 + \frac{1}{2}^2 + \frac{1}{2}^2} = \frac{\sqrt{3}}{2}$$

$$\underset{\text{sim}}{\cos}(A, q) = \frac{\frac{1}{2}\times\frac{1}{2} + \frac{1}{2}\times\frac{1}{2}}{1 \times \frac{\sqrt{5}}{2}} = 0.4472$$

$$\underset{\text{sim}}{\cos}(B, q) = \frac{\frac{1}{2}\times\frac{1}{2} + \frac{1}{2}\times\frac{1}{2}}{1 \times \frac{\sqrt{3}}{2}} = 0.5774 > 0.4472$$

$\Rightarrow$ We recommend this to B.

N-gram Models

(A).
1. I buy banana cake to sell
2. I eat cake
3. he love to eat bananas
4. I love she

	START	I	buy	banana	cake	to	sell	eat	he	love	she	END
START	0	$\frac{3}{4}$	0	0	0	0	0	0	$\frac{1}{4}$	0	0	0
I	0	0	$\frac{1}{3}$	0	0	0	0	$\frac{1}{3}$	0	$\frac{1}{3}$	0	0
buy	0	0	0	1	0	0	0	0	0	0	0	0
banana	0	0	0	0	$\frac{1}{2}$	0	0	0	0	0	0	$\frac{1}{2}$
cake	0	0	0	0	0	$\frac{1}{2}$	0	0	0	0	0	$\frac{1}{2}$
to	0	0	0	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0
sell	0	0	0	0	0	0	0	0	0	0	0	1
eat	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0	0	0	0
he	0	0	0	0	0	0	0	0	1	0	0	0
love	0	0	0	0	0	$\frac{1}{2}$	0	0	0	0	$\frac{1}{2}$	0
she	0	0	0	0	0	0	0	0	0	0	0	1
END	0	0	0	0	0	0	0	0	0	0	0	0

(B). I love to eat cake.

$P(X_1 = I \mid X_0 = START) = \frac{3}{4}$

$P(X_2 = love \mid X_1 = I) = \frac{1}{3}$

$P(X_3 = to \mid X_2 = love) = \frac{1}{2}$

$P(X_4 = eat \mid X_3 = to) = \frac{1}{2}$

$P(X_5 = cake \mid X_4 = eat) = \frac{1}{2}$

$P(X_6 = END \mid X_5 = cake) = \frac{1}{2}$

$\frac{3}{4} \times \frac{1}{3} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$

$= 0.0156$

$\text{perplexity} = \frac{1}{\sqrt[N]{P(\text{sentence})}} = \frac{1}{\sqrt[4]{0.0156}}$

$= 2.83$

True / False

A. False. This situation is more likely to happen with word2vec; however, it can be possible if all three words have exact same term frequency in document.

B. True. TF-IDF will return the same vector regardless token order; it's only about term frequency.

E: False. Latin-1 is an extended-ASCII that doesn't provide flexibility in variable-length encoding; it can't correctly encode the character heart-shape because it's over Latin-1's encoding limit, so there is no way it can be more efficient.