NLP HW4. Answer key
Yulong Wangan.

1.

(a). (i) = Accuracy = $\frac{4+8}{15} = \frac{4}{5}$

(ii) = precision = $\frac{4}{5}$

(iii) = recall = $\frac{4}{6} = \frac{2}{3}$

(iv). $2 \times \frac{P \times R}{P+R}$

$= 2 \times \frac{\frac{4}{5} \times \frac{2}{3}}{\frac{4}{5} + \frac{2}{3}} = 0.727$

(b).

		predicted P	predicted N
actual	P	4	2
	N	1	8

(c). p(gender_actual = "women") = probability of an observation of being women actually.

P(predicted women | actual women) = probability of an observation being predicted as women given the observation is women actually.

(d). Recall should be prioritized because recall cares about how many predicted values are positive out of all actual positives. A high recall indicates most women clothings are categorized correctly.

2.

(A): (i): $\frac{1}{3}$

(ii): $\frac{2}{6} = \frac{1}{3}$

(iii): $\frac{2}{6} = \frac{1}{3}$

(iv): $P(y|x) = \frac{P(x|y)P(y)}{P(x)} = \frac{\frac{1}{3} \times \frac{1}{2}}{\frac{1}{3}} = \frac{1}{2}$

(B). if "love" and "movie" are independent, then P(love and movie) should equal to P(love) × P(movie).

P(love and movie) = $\frac{1}{6}$

P(love) × P(movie) = $\frac{1}{6} \times \frac{4}{6} = \frac{1}{9} \neq \frac{1}{6}$

∴ so "love" and "movie" are not independent.

3.

(a).

	trendy	jeans	old	blue	red	wool
idf	2	7/4	2	2	5/2	5/2
tf(A)	1	1	0	0	0	0
tf(B)	0	1	2	1	0	0
tf(C)	1	1	1	1	1	1
tfidf(A)	2	7/4	0	0	0	0
tfidf(B)	0	7/4	4	2	0	0
tfidf(C)	2	7/4	2	2	5/2	5/2

(b).

q	0	1	1	0	0	0
tf·idf(q)	0	3/4	2	0	0	0

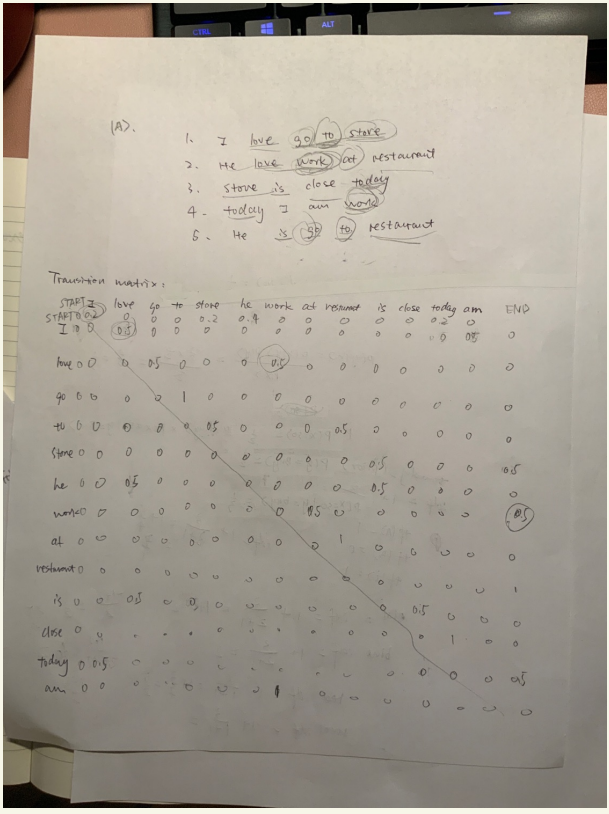
$$\cos(q, B) = \frac{\frac{3}{4}\times\frac{3}{4} + 2\times 4}{\sqrt{(\frac{3}{4})^2 + 2^2 + 4^2}\cdot\sqrt{(\frac{3}{4})^2 + 2^2}} \approx 0.867$$

$$\cos(q, C) = \frac{\frac{3}{4}\times\frac{3}{4} + 2\times 2}{\sqrt{(\frac{3}{4})^2 + 2^2 + 2^2 + (\frac{5}{2})^2 + 6^2}\cdot\sqrt{(\frac{3}{4})^2 + 2^2}} \approx 0.506 < 0.867$$

Therefore, we should recommend product B.

4.

(A).



(B).

$P(X_1 = I \mid X_0 = \text{start}) = 0.2$

$P(X_2 = \text{love} \mid X_1 = I) = 0.5$

$P(X_3 = \text{work} \mid X_2 = \text{love}) = 0.5$

$P(X_4 = \text{End} \mid X_3 = \text{work}) = 0.5$

$0.2 \times 0.5 \times 0.5 \times 0.5 = 0.025$

(C).

① stemming / lemmatization → reduce different forms of the same word/token

② removal of stopwords so the probability of target word following other words is precise.

③ Perplexity: we can't compare sentences with different lengths as longer sentence has smaller probability. We can apply perplexity formula $\frac{1}{\sqrt[N]{P(\text{sentence})}}$ to reduce the effect of sentence length.

(minimizing the perplexity ⟺ maximizing the probability of sentence to be in natural language)