Quiz 3

Name:

1. Most content-based recommender systems use simple retrieval models, such as keyword matching or the Vector Space Model (VSM) with basic TF-IDF weighting. Typically, in VSM, we estimate the relevance of a word t_k for a document d_j as

$$TF-IDF(t_k, d_j) = \frac{f_{k,j}}{max_{k'} f_{k',j}} \cdot \log_{10}\frac{N}{n_k}$$

where N denotes the number of documents in the corpus, and n_k denotes the number of documents in the collection in which the term t_k occurs at least once. $f_{k,j}$ represents the number of appearances of t_k in d_j.

Suppose that an online news service provider has a corpus with a million documents and two samples of them is shown as follows (2 lists of keywords preprocessed):

$$d_1 = \{\text{Interest, real, estate, rising, real, estate}\}$$
$$d_2 = \{\text{Feds, interest, rising, interest, rate, rising}\}$$

A user called Kim have read a news article

$$d_3 = \{\text{Lower, interest, rate, hotter, real, estate, market}\}$$

Note that n_k of each word is

word	n_k
Interest	10
Real	100
Estate	10
Rising	10
Feds	100
Rate	100
Lower	1000
Hotter	1000
Market	100000

Using the cosine similarity between TF-IDF vectors of documents, **which article among d1 and d2 would you recommend Kim?**

(sol)

TF-IDF("interest",d_1) = 1/2 $\log_{10}\frac{1000000}{10}$ = 2.5

TF-IDF("real",d_1) = 2/2 $\log_{10}\frac{1000000}{100}$ = 4

TF-IDF("estate", d_1) = 5

…

For <interest, real, estate, rising, feds, rate, lower, hotter, market>,

$$d_1 =< 2.5, 4, 5, 2.5, 0, 0, 0, 0\ 0 >$$

$$d_2 =< 5, 0, 0, 5, 2, 2, 0, 0, 0 >$$
$$d_3 =< 3, 4, 5, 0, 0, 4, 3, 3, 1 >$$

Because cosine similarity is

$$sim(d_i, d_j) = \frac{\sum_k w_{ki} \cdot w_{kj}}{\sqrt{\sum_k w_{ki}^2} \cdot \sqrt{\sum_k w_{kj}^2}}.$$

We can obtain the similarities of d_3 to the others as

- $sim(d_1, d_3) = 0.72$
- $sim(d_2, d_3) = 0.33$

Therefore, we recommend d_1.

2. Naive Bayes is a probabilistic approach to inductive learning and belongs to the general class of Bayesian classifiers. Suppose that Kim have *read* d_3 but *does not read* d_2 in the above question. Based on naïve Bayesian classifier model, **estimate the probability that Kim reads $\boldsymbol{d_1}$.**

Refer to the 14th page of *Content-based Recommendation Systems* by M. J. Pazzani1 and D. Billsus. In this question, a class is identified between *read* and *unread* articles.

(sol)

Note that we use the multinomial naïve Bayes formulation since it was shown to outperform the multivariate Bernoulli model.

$$P(c_j \mid d_i;\hat{\theta}) = \frac{P(c_j \mid \hat{\theta})P(d_i \mid c_j;\hat{\theta})}{P(d_i \mid \hat{\theta})}$$

where

$$P(d_i \mid c_j;\theta) = P(|d_i|)\prod_{t=1}^{|d_i|} P(w_t \mid c_j;\theta)^{N_{it}}$$

And for each P(w|c) is

$$P(w_t \mid c_j;\theta) = \frac{1+\sum_{i=1}^{|D|} N_{it}\, P(c_j \mid d_i)}{|V| + \sum_{s=1}^{|V|}\sum_{i=1}^{|D|} N_{is}\, P(c_j \mid d_i)}.$$

- $P(c_j|d_i)$ is 1 if the document d_i is classified as c_j, 0 otherwise
- N_{it} is the number of occurrences of the word w_t in d_i

Recall that

- $d_1 = \{\text{Interest, real, estate, rising, real, estate}\}$ → **?**
- $d_2 = \{\text{Feds, interest, rising, interest, rate, rising}\}$ → **unread**
- $d_3 = \{\text{Lower, interest, rate, hotter, real, estate, market}\}$ → **read**

We obtain the following probabilities:

- P(class = "unread") = 0.5
- P(class = "read") = 0.5
- P("interest" | "unread") = $\frac{1+2}{9+6}$ = 0.2
- P("interest" | "read") = $\frac{1+1}{9+7}$ = 0.125

- P("real"| "unread") = $\frac{1+0}{9+6}$ = 0.0667
- P("real"| "read") = $\frac{1+1}{9+7}$ = 0.125
- P("estate"| "unread") = $\frac{1+0}{9+6}$ = 0.0667
- P("estate"| "read") = $\frac{1+1}{9+7}$ = 0.125
- P("rising"| "unread") = $\frac{1+2}{9+6}$ = 0.2
- P("rising"| "read") = $\frac{1+0}{9+7}$ = 0.0625

$P(class = "unread" | d_1 = \{\text{Interest, real, estate, rising, real, estate}\})$ **→ ?**

P(class= "unread") * P("interest" | "unread") * P("real"| "unread")^2 * P("estate"| "unread")^2 * P("rising"| "unread") = 0.000000395852445
P(class = "read") * P("interest" | "read") * P("real"| "read")^2 * P("estate"| "read")^2 * P("rising"| "read") = 0.000000953674316

By Baye's theorem, $P(unread|d_1) = \frac{p(d_1|unread) \cdot p(unread)}{p(d_1|unread) \cdot p(unread) + p(d_1|read) \cdot p(read)}$

- $P(class = \text{unread}|d_1 = \{\text{Interest, real, estate, rising, real, estate}\}) = \frac{0.000000395852445}{0.000000395852445+0.000000953674316}$
- $P(class = \text{read}|d_1 = \{\text{Interest, real, estate, rising, real, estate}\}) = \frac{0.000000953674316}{0.000000395852445+0.000000953674316}$

3. The Rocchio's method is used for inducing linear, profile-style classifiers. This algorithm represents a class as vectors and recommends a document in a class with similar vectors. Rocchio's method computes a classifier $c_i = < \omega_{1i}, \dots, \omega_{|T|i} >$ for a class i where |T| denotes the size of vocabulary (i.e., it represents each class as a vector of |T|). Let us assume that Kim has read d_3 but does not d_2 similar to the above question. Thus, d_3 becomes a positive sample for *read articles* class and d_2 is a negative sample for the class. On the other hand, d_3 is a negative sample for *unread articles* class and d_2 is a positive sample for the class.

Compute a classifier $c_i = < \omega_{1i}, \dots, \omega_{|T|i} >$ for *read article* class (Use the TF-IDF vectors calculated in Question 1).

(sol)

$$\omega_{ki} = \beta \cdot \sum_{\{d_j \in POS_i\}} \frac{\omega_{kj}}{|POS_i|} - \gamma \cdot \sum_{\{d_j \in NEG_i\}} \frac{\omega_{kj}}{|NEG_i|}$$

From Question 1,

- $d_2 = < 5, 0, 0, 5, 2, 2, 0, 0, 0 >$ → NEG_{read}
- $d_3 = < 3, 4, 5, 0, 0, 4, 3, 3, 1 >$ → POS_{read}

Assume that β and γ are 1. Then,

$$c_{read} = < \frac{3}{1} - \frac{5}{1}, \frac{4}{1} - \frac{0}{1}, \frac{5}{1} - \frac{0}{1}, \frac{0}{1} - \frac{5}{1}, \frac{0}{1} - \frac{2}{1}, \frac{4}{1} - \frac{2}{1}, \frac{3}{1} - \frac{0}{1}, \frac{3}{1} - \frac{0}{1}, \frac{1}{1} - \frac{0}{1} >$$
$$= < -2, 4, 5, -5, -2, 2, 3, 3, 1 >$$