Information Retrieval

Probabilistic retrieval, Text classification

1. Suppose the document collection contains two documents:
   a. d1: Xyzzy reports a profit but revenue is down
   b. d2: Quorus narrows quarter loss but revenue decreases further

   Use the MLE unigram models from the documents and collection, with $\lambda = 1/2$. for the query 'revenue down', to rank the documents.
2. A set of documents

   (1) He moved from London, Ontario, to London, England.

   (2) He moved from London, England, to London, Ontario.

   (3) He moved from England to London, Ontario.

   Which of the documents have identical and different bag of words representations for (i) the Bernoulli model (ii) the multinomial model? If there are differences, describe them.
3. Estimates for term the are $P(X = \text{the}|c) \approx 0.05$ $P(U_{\text{the}} = 1|c) \approx 1.0$, for the multinomial and Bernoulli model respectively. Explain the difference.
4. Consider the following frequencies for the class *coffee* for four terms in the first 100,000 documents of Reuters-RCV1, Select two of these four terms based on (i) χ^2, (ii) mutual information, (iii) frequency.

term	N00	N01	N10	N11
brazil	98,012	102	1835	51
council	96,322	133	3525	20
producers	98,524	119	1118	34
roasted	99,824	143	23	10

5. Each document in the table has been classified as being related to China or not. Using this data, (i) estimate a multinomial Naive Bayes classifier, (ii) apply the classifier to the test document, (iii) estimate a Bernoulli NB classifier, (iv) apply the classifier to the test document. You need not estimate parameters that you don't need for classifying the test document.

docId	words	in c = China?
1	Taipei Taiwan	yes
2	Macao Taiwan Shanghai	yes
3	Japan Sapporo	no
4	Sapporo Osaka Taiwan	no
5	Taiwan Taiwan Sapporo	?

Solutions

1. $P(t|d) = \lambda P_{mle}(t|M_d) + (1-\lambda) P_{mle}(t|M_c)$
   $P(q|d1) = [(1/8 + 2/16)/2] \times [(1/8 + 1/16)/2] = 1/8 \times 3/32 = 3/256$
   $P(q|d2) = [(1/8 + 2/16)/2] \times [(0/8 + 1/16)/2] = 1/8 \times 1/32 = 1/256$
   $d1 > d2$.
2. (i) The three documents have identical bag of words representations.
   (ii) (1) and (2) are identical, the count of London in the representation of (3) is 1 instead of 2.
3. The numerical model calculates P(X=the)by the number of occurrences of the in the collection which is 5%(large as the is a stop word). The Bernoulli model , on the other hand , just calculates the probability by the number of documents which contain the word the which is 1.0 as all documents will contain it.
4. (i) brazil, roasted(ii) brazil, producer(iii) brazil, producer
5. (i) P (c) = P(c) = 1/2. The vocabulary has 7 terms: Japan, Macao, Osaka, Sapporo, Shanghai, Taipei, Taiwan. There are 5 tokens in the concatenation of all c documents. There are 5 tokens in the concatenation of all c documents. Thus, the denominators have the form (5+7).
   P(Taiwan|c) = (2+1)/(5+7) = 1/4, P(Taiwan|c') = (1+1)/(5+7) = 1/6,
   P(Sapporo|c) = (0+1)/(5+7) = 1/12, P(Sapporo|c') = (2+1)/(5+7) = 1/4
   (ii) We then get $P(c|d) \propto 1/2 \cdot (1/4)^2 \cdot 1/12 = 1/(2^7 \cdot 3) \approx 0.00260$ and $P(c'|d) \propto 1/2 \cdot (1/6)^2 \cdot (1/4) = 1/(2^5 \cdot 3^2) \approx 0.00347$. P(c|d)/P(c'|d) =3/4.Thus, the classifier assigns the test document to c=not China.
   (iii) Estimating parameters of a binomial Naive Bayes classifier:
   p(c=China)=0.5, p(c=not China)=0.5
   p(Taiwan|c=China)=(2+1)/(2+2)=3/4
   p(Sapporo|c=China)=p(Japan|c=China)=p(Osaka|c=China)=(0+1)/(2+2)=1/4
   p(Macao|c=China)=p(Shanghai|c=not China)=p(Osaka|c=not China) =
   (1+1)/(2+2)=1/2
   p(Sapporo|c=not China)=3/4
   p(Taiwan|c=not China) = p(Japan|c=not China)= p(Osaka|c=not China)=
   (1+1)/(2+2)=1/2
   p(Taipei|c=not China)=p(Shanghai|c=not China)=p(Macao|c=not China)=1/4
   (iv) p(test set|c=China)=(3/4)*(1/4)*(3/4)2*(1/2)3=27/211
   p(test set|c=not China)=(1/2)*(3/4)*(1/2)2*(3/4)3 = 82/211