Web Searching Mining Assignment3

student ID:110753201 Class:資科碩一 Name: 曹昱維

1

1-1

Priors: $P(c)=\frac{3}{4}, P(\bar{c})=\frac{1}{4}$
conditional probabilities：
$P(Chinese|c)=\frac{(5+1)}{(8+6)}=\frac{3}{7}, P(Tokyo|c)=P(Japan|c)=\frac{(0+1)}{(8+6)}=\frac{1}{14}$
$P(Chinese|\bar{c})=\frac{(1+1)}{(3+6)}=\frac{2}{9}, P(Tokyo|\bar{c})=P(Japan|\bar{c})=\frac{(1+1)}{(3+6)}=\frac{2}{9}$

1-2

$P(c|d_5)\propto\frac{3}{4}*\frac{3}{7}*\frac{3}{7}*\frac{3}{7}*\frac{1}{14}*\frac{1}{14}\approx 0.0003$
$P(\bar{c}|d_5)\propto\frac{1}{4}*\frac{2}{9}*\frac{2}{9}*\frac{2}{9}*\frac{2}{9}*\frac{2}{9}\approx 0.0001$
Therefore, the classifier assigns the test document to **c = China**

1-3

Priors: $P(c)=\frac{3}{4}, P(\bar{c})=\frac{1}{4}$
conditional probabilities：
$P(Chinese|c)=\frac{(3+1)}{(3+2)}=\frac{4}{5}, P(Tokyo|c)=P(Japan|c)=\frac{(0+1)}{(3+2)}=\frac{1}{5}$
$P(Chinese|\bar{c})=\frac{(1+1)}{(1+2)}=\frac{2}{3}, P(Tokyo|\bar{c})=P(Japan|\bar{c})=\frac{(1+1)}{(1+2)}=\frac{2}{3}$
$P(Beijing|\bar{c})=P(Macao|\bar{c})=P(Shanghai|\bar{c})=\frac{(0+1)}{(1+2)}=\frac{1}{3}$

1-4

$P(c|d_5)\propto P(c)*P(Chinese|c)*P(Japan|c)*P(Tokyo|c)*(1-P(Beijing|c))*$
$(1-P(Macao|c))*(1-P(Shanghai|c))$
$=\frac{3}{4}*\frac{4}{5}*\frac{1}{5}*\frac{1}{5}*(1-\frac{2}{5})*(1-\frac{2}{5})*(1-\frac{2}{5})\approx 0.005$
$P(\bar{c}|d_5)\propto P(\bar{c})*P(Chinese|\bar{c})*P(Japan|\bar{c})*P(Tokyo|\bar{c})*(1-P(Beijing|\bar{c}))*$
$(1-P(Macao|\bar{c}))*(1-P(Shanghai|\bar{c}))$
$=\frac{1}{4}*\frac{2}{3}*\frac{2}{3}*\frac{2}{3}*(1-\frac{1}{3})*(1-\frac{1}{3})*(1-\frac{1}{3})\approx 0.022$
Therefore, the classifier assigns the test document to $barc$
= Not China

2

2-1

$\chi^2(t,c)=\frac{(N_{11}+N_{10}+N_{01}+N_{00})*(N_{11}*N_{00}-N_{01}*N_{10})^2}{(N_{11}+N_{01})(N_{11}+N_{10})(N_{00}+N_{10})(N_{00}+N_{01})}$
$\chi^2(brazil,c)=\frac{(51+1835+102+98012)*(51*98012-102*1835)^2}{(51+102)(51+1835)(98102+1835)(98102+102)}\approx 817.45132$
$\chi^2(council,c)=\frac{(20+3525+133+96322)*(20*96322-133*3525)^2}{(20+133)(20+3525)(96322+3525)(96322+133)}\approx 40.67412$
$\chi^2(producer,c)=\frac{(34+1118+119+98524)*(34*98524-119*1118)^2}{(34+119)(34+1118)(98524+1118)(98524+119)}\approx 596.06999$
$\chi^2(roasted,c)=\frac{(10+23+143+99824)*(10*99824-23*143)^2}{(10+143)(10+23)(99824+23)(99824+143)}\approx 1964.29329$
Therefore we selected **brazil, roasted**

2-2

$$I(U;C) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}}+\frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} +\frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}}+\frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}}$$

$$I(brazil;C) = \frac{98012}{100000} * log_2(\frac{(100000 * 98012)}{(98012+102)(98012+1835)}) + \frac{102}{100000} * log_2(\frac{100000 * 102}{(102+98012)(51+102)})$$
$$+ \frac{1835}{100000} * log_2(\frac{100000 * 1835}{(51+1835)(1835+98012)}) + \frac{51}{100000} * log_2(\frac{100000 * 51}{(51+1835)(51+102)})$$
$$\approx 0.0015536892$$

$$I(council;C) = \frac{96322}{100000} * log_2(\frac{100000 * 96322}{(96322+133)(96322+3525)}) + \frac{133}{100000} * log_2(\frac{100000 * 133}{(133+96322)(20+133)})$$
$$+ \frac{3525}{100000} * log_2(\frac{100000 * 3525}{(20+3525)(3525+96322)}) + \frac{20}{100000} * log_2(\frac{100000 * 20}{(20+3525)(20+133)})$$
$$\approx 0.0001774273$$

$$I(producers;C) = \frac{98524}{100000} * log_2(\frac{100000 * 98524}{(98524+119)(98524+1118)}) + \frac{119}{100000} * log_2(\frac{100000 * 119}{(119+98524)(34+119)}$$
$$+ \frac{1118}{100000} * log_2(\frac{100000 * 1118}{(34+1118)(1118+98524)}) + \frac{34/100000}{*} log_2(\frac{100000 * 34}{(34+1118)(34+119)})$$
$$\approx 0.0010479995$$

$$I(roasted;C) = \frac{99824}{100000} * log_2(\frac{100000 * 99824}{(99824+143)(99824+23)}) + \frac{143/100000}{*} log_2(\frac{100000 * 143}{(119+99824)(10+143)})$$
$$+ \frac{23}{100000} * log_2(\frac{100000 * 23}{(10+23)(23+99824)}) + \frac{34}{100000} * log_2(\frac{100000 * 10}{(10+23)(10+143)})$$
$$\approx 0.0006484759$$

Therefore we selected **brazil, producer**

2-3

TFIDF would use the number of documents in the class c that contain the term t, so we need to compare the values of N_{11}.
Therefore we selected **brazil, producer**

3

3-1

macro-averaged precision = $\frac{\frac{80}{90}+\frac{20}{60}}{2} \approx 0.605$

3-2

micro-averaged precision = $\frac{80+20}{80+10+20+40} \approx 0.66$

4

4-1

The decision boundary/surface form :
$\vec{w}^T\vec{x} = b$
A vector $\vec{x}$ is on the decision boundary if it has equal distance to the two class centroids:
$|\vec{\mu}(c_1) - \vec{x}| = |\vec{\mu}(c_2) - \vec{x}|$
$\therefore (\vec{\mu}(c_1) - \vec{x})^2 = (\vec{\mu}(c_2) - \vec{x})^2$
$\Rightarrow (\vec{\mu}(c_1) - \vec{\mu}(c_2))\vec{x} = 0.5 * (\vec{\mu}(c_1)^2 - \vec{\mu}(c_2)^2)$
So the decision boundary of Rocchio classifier :
$(\vec{\mu}(c_1) - \vec{\mu}(c_2))\vec{x} = 0.5 * (\vec{\mu}(c_1)^2 - \vec{\mu}(c_2)^2)$

4-2

Rocchio classification does not handle multimodal classes correctly. Multimodal classes have more than two label for a data point while Rocchio classification can only deal with one label for a data point.

5

5-1

$< \vec{a}, \vec{x} >= 4, < \vec{b}, \vec{x} >= 16, < \vec{c}, \vec{x} >= 28$
Therefore,most similar is $\vec{c}$

5-2

$\frac{<\vec{a},\vec{x}>}{|\vec{a}||\vec{x}|} = \frac{4}{\sqrt{\frac{5}{2}}*2\sqrt{2}} \approx 0.8944$
$\frac{<\vec{b},\vec{x}>}{|\vec{b}||\vec{x}|} = \frac{16}{4\sqrt{2}*2\sqrt{2}} = 1$
$\frac{<\vec{c},\vec{x}>}{|\vec{c}||\vec{x}|} = \frac{28}{10*2\sqrt{2}} \approx 0.9899$
Therefore,most similar is $\vec{b}$

5-3

$d(\vec{a}, \vec{x}) = 1.5811, d(\vec{b}, \vec{x}) = 2.8284, d(\vec{c}, \vec{x}) = 7.2111$
Therefore,most similar is $\vec{a}$