(a) ① case-folding : change upper case to lower case

② removing punctuations and non-alphanumeric chars.

more: canonization, Remove of stop word, remove of page metadata, spliting into tokens.

(b) ① /[A-Z]/ → corresponding lower case

② /\W/ → ∅

(c) australia → 1(4) → 5(1) → 6(1)

election → 1(4) → 2(1) → 3(4) → 4(1) → 5(1) → 6(2)

federal → 1(2) → 3(3) → 4(2) → 5(1) → 6(2)

vote → 4(2) → 6(5)

ausrlian → 5(1)

(d) f_t, and pointers (given a token, point to the inverted index)

(e) $w_{1,t} = \langle 4, 4, 2, 0, 0 \rangle$

$w_{2,t} = \langle 0, 1, 0, 0, 0 \rangle$

$w_{3,t} = \langle 0, 4, 3, 0, 0 \rangle$

$w_{4,t} = \langle 0, 1, 2, 2, 0 \rangle$

$w_{5,t} = \langle 1, 1, 1, 0, 1 \rangle$

$w_{6,t} = \langle 1, 2, 2, 5, 0 \rangle$

$w_{q,t} = \langle 2, 1, 0, 0, 0 \rangle$

$\cos(1,q) = \frac{8+4}{\sqrt{34}\cdot\sqrt{5}} \approx 0.920$

$\cos(2,q) = \frac{1}{\sqrt{5}} \approx 0.447$

$\cos(3,q) = \frac{4}{5\cdot\sqrt{5}} \approx 0.358$

$\cos(4,q) = \frac{1}{3\sqrt{5}} \approx 0.149$

$\cos(5,q) = \frac{3}{2\sqrt{5}} \approx 0.671$

$\cos(6,q) = \frac{4}{\sqrt{34}\cdot\sqrt{5}} \approx 0.307$

∴ the rank is $(d_1, d_5, d_2, d_3, d_6, d_4)$

(f) i: means the document is useful to the user, which can resolve the user's information need; we discover this by checking whether the user click into it or not.

ii P@2 = 0

(g) ? $1 + \log_2 f_{d,t} = \log_2 2 f_{d,t} > \log_2 f_{d,t}$

the relative weight of $f_{d,t}$ increase means the terms appear in a document frequently gets a higher weight so the documents which contain more query term may get a higher rank

(h) Spelling Correction:

(i) parsing

(ii) a standard dictionary should be included and we need to use approximate string matching to compare the misspell word to the words in the dictionary to find the correct word.

(iii) Some document with misspell word would also be considered, like document 5.

(a) $P(\text{FACEBOOK} \mid ele=N, fed=N, aust=N, vot=N, ausr=Y)$

$= P(ele=N \mid FB) \cdot P(fed=N \mid FB) \cdot P(aust=N \mid FB) \cdot P(vot=N \mid FB) \cdot P(ausr=Y \mid FB) \cdot P(FB)$

$= \varepsilon \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot 1 \cdot \varepsilon \cdot \frac{1}{3} = \frac{\varepsilon^2}{12}$

$P(\text{TWITTER} \mid ele=N, fed=N, aust=N, vot=N, ausr=Y)$

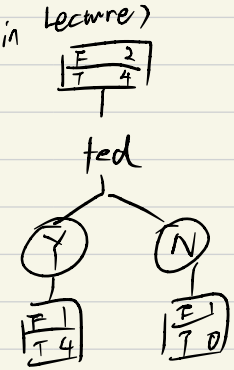
$= P(ele=N \mid TW) \cdot P(fed=N \mid TW) \cdot P(aust=N \mid TW) \cdot P(vot=N \mid TW) \cdot P(ausr=Y \mid TW) \cdot P(TW)$

$= \varepsilon \cdot \varepsilon \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{2}{3} = \frac{\varepsilon^2}{24}$

$\because \frac{\varepsilon^2}{12} > \frac{\varepsilon^2}{24}$

$\therefore$ classify it to FACEBOOK

(b) The solution is based on GR (GINI not in lecture)



$H(R) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) = 0.918$

$IG(ele \mid R) = 0$

$SI(ele \mid R) = 0$

$GR(ele \mid R) = 0$

$IG(fed \mid R) = H(R) - \left(-\left(\frac{5}{6}\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right)\right)\right)$

$= 0.918 - 0.602 \approx 0.316$

$SI(fed \mid R) = -\left(\frac{5}{6}\log\frac{5}{6} + \frac{1}{6}\log\frac{1}{6}\right)$

$= 0.650$

$GR(fed \mid R) = 0.486$

最后比 GR，选 GR 最大的做 root.

(c) i. training data not enough, model is under fit

ii. NB not change : $P(aust=Y \mid TWITTER) = P(aust=N \mid TWITTER) = \frac{1}{2}$

$P(aust=Y \mid FACEBOOK) = P(aust=N \mid FACEBOOK) = \frac{1}{2}$

(d) ele, fed, aust, vot, ausr is conditional independent given the label. The assumption is invalid, for example aust and ausr is relevant since aust may be the misspelling form of ausr

(e) NO, the dataset is small, and none of the training data is similar to the test data. Bagging is mainly to reduce sampleing bias.

(f) $\{ele=Y\} \rightarrow \{fed=Y\}$ support : $\frac{5}{6}$

confident : $\frac{5}{6}$

(a) ① Training data not enough
② Not effective to calcalate distance in Boolean attribute.

Change: crawl more data and record f_t of the the data instead of whether the term appear.

(b) the data can be seperated into two side by a hyperplane, data on one side belongs to FB, other side belongs to TW

Since the data is already linear-seperatable, we don't need to use kernel method to change the data, but soft-margin is still need to avoid being too sesceptible to noise.

(c) f_t, get more information, now we can get the frequency of terms, with frequency. we can calculate distance more accurately.

(f) use cross-validation, (i) the dataset is the data with label (ii) use part of the data set for testing, the rest of it for training, each time, and iterate the process, and calculate the calculate the average f1 score of them at the end, f1 score formula:

$$\text{f1 score} = \frac{2\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

(iii) cross-validation