Probabilistic Models: Spring 2014
Document Classification Example, Solutions

We are given the following **corpus** and topics. Only the words in **bold** are in the vocabulary.

Topic	Text
Fantasy	The hobbit tricked the **troll**. He hid from the **dragon**. The **dragon** set the **town** on **fire**. The dwarf killed the **dragon** and became **king**.
Technology	Many people use a **fire wall** to increase their **security**. The **security forum** helps people configure their **fire wall** to prevent hackers from setting their computers on **fire**.
High Seas	The **pirate** sailed his **ship** into **town**. The **pirate** scaled the **wall** and took the **king** prisoner on the **ship**. He later set the **town** on **fire**.
Technology	A **troll** lives in this **forum**. Do not feed the **troll**; he believes he is **king** of the **forum** and will set any thread on **fire**.
Fantasy	The **king** beyond the **wall** attacked a **town**. A **pirate** works for a different **king**. Yet another **king** has a **dragon** that set a **town** on **fire**.

1. Convert the documents into their bag of words representation. Use this order for the words: dragon, fire, forum, king, pirate, security, ship, town, troll, wall.

Table 1: The bag of words representation of each document, $\mathbf{n}$, and the number of times each word j occurs in documents of each topic k, $\sum_{i \in \mathbf{z}_k} \mathbf{n}_{i,j}$. The "Total" gives $\sum_{s=1}^{d} \sum_{i \in \mathbf{z}_k} \mathbf{n}_{i,s}$ for each topic.

	Bag of words					Topic		
Topic	Fantasy	Technology	High Seas	Technology	Fantasy	Fantasy	Technology	High Seas
dragon	3	0	0	0	1	4	0	0
fire	1	3	1	1	1	2	4	1
forum	0	1	0	2	0	0	3	0
king	1	0	1	1	3	4	1	1
pirate	0	0	2	0	1	1	0	2
security	0	2	0	0	0	0	2	0
ship	0	0	2	0	0	0	0	2
town	1	0	2	0	2	3	0	2
troll	1	0	0	2	0	1	2	0
wall	0	2	1	0	1	1	2	1
Total						16	14	9

2. Construct the naive Bayes classifier for the corpus.

   - Prior probabilities for the topics

$$Pr(C = \text{Fantasy}) = \frac{N_{\text{Fantasy}} + 1}{N + T} = \frac{3}{8}$$
$$Pr(C = \text{Technology}) = \frac{N_{\text{Technology}} + 1}{N + T} = \frac{3}{8}$$
$$Pr(C = \text{High Seas}) = \frac{N_{\text{High Seas}} + 1}{N + T} = \frac{2}{8}$$

   - Conditional probabilities for the words given the topics
     We can calculate, for example,

$$\begin{aligned} Pr(w_t = \text{dragon}|C = \text{Fantasy}) &= \frac{1 + \sum_{i \in \mathbf{Z}_{\text{Fantasy}}} \mathbf{n}_{i,\text{dragon}}}{d + \sum_{s=1}^{d} \sum_{i \in \mathbf{Z}_{\text{Fantasy}}} \mathbf{n}_{i,s}} \\ &= \frac{1+4}{10+16} \\ &= \frac{5}{26}. \end{aligned}$$

     The rest of the conditional probabilities are calculated similarly.

	Topic		
Word	Fantasy	Technology	High Seas
dragon	$\frac{5}{26}$	$\frac{1}{24}$	$\frac{1}{19}$
fire	$\frac{3}{26}$	$\frac{5}{24}$	$\frac{2}{19}$
forum	$\frac{1}{26}$	$\frac{4}{24}$	$\frac{1}{19}$
king	$\frac{5}{26}$	$\frac{2}{24}$	$\frac{2}{19}$
pirate	$\frac{2}{26}$	$\frac{1}{24}$	$\frac{3}{19}$
security	$\frac{1}{26}$	$\frac{3}{24}$	$\frac{1}{19}$
ship	$\frac{1}{26}$	$\frac{1}{24}$	$\frac{3}{19}$
town	$\frac{4}{26}$	$\frac{1}{24}$	$\frac{3}{19}$
troll	$\frac{2}{26}$	$\frac{3}{24}$	$\frac{1}{19}$
wall	$\frac{2}{26}$	$\frac{3}{24}$	$\frac{2}{19}$

3. Calculate the likelihood, or **conditional distributions**, for each document in the corpus ($Pr(\mathbf{n}_i|C = z_i)$).

4. Calculate the posterior probability, or **classification distribution**, for the following unlabeled documents ($Pr(C = k|\mathbf{n}_i)$).

Topic	Text
?	The red **king** and his **troll** attacked the **town** by **ship**. Somehow, the red **king** still set the **town** on **fire**.
?	The **forum** is on **fire** with discussion of a **pirate** ship which bypassed the **security** of a cruise **ship**. The **pirate** uploaded a video to the **forum**; naturally, the cruise **ship** was on **fire**.

Some useful equations

$$
\begin{aligned}
N &\coloneqq \text{the number of documents} \\
T &\coloneqq \text{the number of topics} \\
N_k &\coloneqq \text{the number of documents from topic } k \\
\mathbf{n}_{i,j} &\coloneqq \text{the number of times word } j \text{ appears in document } i \\
z_i &\coloneqq \text{the topic of document } i \\
\mathbf{Z}_k &\coloneqq \text{the indices of all documents from topic } k
\end{aligned}
$$

$$
\begin{aligned}
Pr(C = k) &= \frac{N_k + 1}{N + T} \\
Pr(w_t = j | C = k) &= \frac{1 + \sum_{i \in \mathbf{Z}_k} \mathbf{n}_{i,j}}{d + \sum_{s=1}^{d} \sum_{i \in \mathbf{Z}_k} \mathbf{n}_{i,s}} \\
P(\mathbf{n}_i | C = k) &= P(\text{drawing } \mathbf{n}_i \text{ one way} | C = k) \times \text{number of ways to draw } \mathbf{n}_i \\
Pr(C = k | \mathbf{n}_i) &= \frac{Pr(\mathbf{n}_i |\ C = k) \times Pr(C = k)}{Pr(\mathbf{n}_i)}
\end{aligned}
$$