7 Text Classification

Exercise VII.1

In an automated email router of a company, we want to make the distinction between three kind of emails: technical (about computers), financial, and the rest ("irrelevant"). For this we plan to use a Naive Bayes approach.

① What is the main assumption made by Naive Bayes classifiers? Why is it "Naive"?

We will consider the following three messages:

> The Dow industrials tumbled 120.54 to 10924.74, hurt by GM's sales forecast and two economic reports. Oil rose to $71.92.

from `www.wsj.com/`

> BitTorrent Inc. is boosting its network capacity as it prepares to become a centralized hub for legal video content. In May, BitTorrent announced a deal with Warner Brothers to distribute its TV and movie content via the BT platform. It has now lined up IP transit for streaming videos at a few gigabits per second

from `slashdot.org/`

> Intel will sell its XScale PXAxxx applications processor and 3G baseband processor businesses to Marvell for $600 million, plus existing liabilities. The deal could make Marvell the top supplier of 3G and later smartphone processors, and enable Intel to focus on its core x86 and wireless LAN chipset businesses, the companies say.

from `www.linuxdevices.com/`

② What pre-processing steps (before actually using the Naive Bayes Classifier) do you consider applying to the input text?

③ For the first text, give an example of the corresponding output of the pre-processor.

continues on back ☞

Suppose we have collected the following statistics[2] about the word frequencies within the corresponding classes, where "0.00…" stands for some very small value:

	technical	financial	irrelevant
$<number>	0.01	0.07	0.05
Dow	0.00…	0.08	0.00…
GM	0.00…	0.03	0.00…
IP	0.03	0.00…	0.00…
Intel	0.02	0.02	0.00…
business	0.01	0.07	0.04
capacity	0.01	0.00…	0.00…
chipset	0.04	0.01	0.00…
company	0.01	0.04	0.05

	technical	financial	irrelevant
deal	0.01	0.02	0.00…
forecast	0.00…	0.03	0.01
gigabit	0.03	0.00…	0.00…
hub	0.06	0.00…	0.01
network	0.04	0.01	0.00…
processor	0.07	0.01	0.00…
smartphone	0.04	0.04	0.01
wireless	0.02	0.01	0.00…

④ For each of the above three texts, in what category will it be classified, knowing that on average 50% of the emails happen to be technical, 40% to be financial and 10% to be of no interest.

You can assume that all the missing information is irrelevant (i.e. do not impact the results).

Provide a full explanation of all the steps and computations that lead to your results.

We now want to specifically focus on the processing of compounds such as "network capacity" in the second text.

⑥ How are the compounds handled by a Naive Bayes classifier if no specific pre-processing of compounds is used?

⑦ What changes if the compounds are handled by the NL pre-processor?

Discuss this situation (NL pre-processing handling compounds) with respect to the Naive Bayes main assumption.

⑧ Outline how you would build a pre-processor for compound words.

[2]Note that this is only partial information, statistics about other words not presented here have also been collected.

Exercise VII.2

You are responsible for a project aiming at providing on-line recommendations to the customers of a on-line book selling company.

The general idea behind this recommendation system is to cluster books according to both customers and content similarities, so as to propose books similar to the books already bought by a given customer. The core of the recommendation system is a clustering algorithm aiming at regrouping books likely to be appreciate by the same person. This clustering should not only be achieved based on the purchase history of customers, but should also be refined by the content of the books themselves. It's that latter aspect we want to address in this exam question.

① Briefly explain how books could be clustered according to similar content. Give the main steps and ideas.

② The chosen clustering algorithm is the dendrogram. What other algorithms could you propose for the same task? Briefly review advantages and disadvantages of each of them (including dendrograms). Which one would you recommend for the targeted task?

③ Consider the following six "documents" (toy example):

d_1 "*Because cows are not sorted as they return from the fields to their home pen, cow flows are improved.*"

d_2 "*He was convinced that if he owned the fountain pen that he'd seen in the shop window for years, he could write fantastic stories with it. That was the kind of pen you cannot forget.*"

d_3 "*With this book you will learn how to draw humans, animals (cows, horses, etc.) and flowers with a charcoal pen.*"

d_4 "*The cows were kept in pens behind the farm, hidden from the road. That was the typical kind of pen made for cows.*"

d_5 "*If Dracula wrote with a fountain pen, this would be the kind of pen he would write with, filled with blood red ink. It was the pen she chose for my punishment, the pen of my torment. What a mean cow!*"

d_6 "*What pen for what cow? A red pen for a red cow, a black pen for a black cow, a brown pen for a brown cow, ... Understand?*"

and suppose (toy example) that they are indexed only by the two words: *pen* and *cow*.

(a) Draw their vector representations.

(b) Give the definition of the cosine similarity. What vector's feature(s) is it sensible to?

(c) What is the result of the dendrogram clustering algorithm on those six documents, using the cosine similarity and single linkage?
Explain all the steps.

Hint: $5/\sqrt{34} < 3/\sqrt{10} < 4/\sqrt{17}$.