ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE – LAUSANNE
POLITECNICO FEDERALE – LOSANNA
SWISS FEDERAL INSTITUTE OF TECHNOLOGY – LAUSANNE

Faculté Informatique et Communication
Introduction to Natural Language Processing (Ms; CS-431)
Chappelier, J.-C. & Rajman, M.

CS-431 Hands On Text Classification
Solutions

J.-C. Chappelier M. Rajman

v. 2021118 – 1

QUESTION I [3 pt]

The Naïve Bayes algorithm is used in the framework of a sentiment analysis application to determine, for any input tweet, which, among a predefined set of sentiments, best corresponds to the mood expressed in the tweet.

Does the performed tweet classification task have to be supervised in this case?

[✔] yes [ ] no [ ] it depends on the implementation

Let us assume that only two sentiments are considered ("joyful" and "sad") and that typically 70% of the tweets are "joyful".

To which sentiment would the Naïve Bayes algorithm associate a tweet indexed by only two terms w_1 and w_2 , if:

- 10% of the occurrences of indexing terms in "joyful" tweets and 20% of the occurrences of indexing terms in "sad" tweets are w_1 ; while
- 30% of the occurrences of indexing terms in "joyful" tweets and 25% of the occurrences of indexing terms in "sad" tweets are w_2?

[ ] sad [✔] joyful [ ] undecidable

$$P(\text{sad}) \times P(w_1|\text{sad}) \times P(w_2|\text{sad}) = 0.3 \times 0.2 \times 0.25 = 1.5 \cdot 10^{-2}$$
$$P(\text{joyful}) \times P(w_1|\text{joyful}) \times P(w_2|\text{joyful}) = 0.7 \times 0.1 \times 0.3 = 2.1 \cdot 10^{-2}$$

QUESTION II [2 pt]

Consider the following matrix of measures over a set of three items:

0	5	2
5	0	2
2	2	0

What type(s) of measure is this matrix compatible with?

[✔] A dissimilarity only. $5 > 2 + 2$

[ ] A dissimilarity and a distance/metric.

[ ] None of the two

QUESTION III [4 pt]

You're working on an email classification software (and have some corpus).

In order to better understand your corpus, you plan to cluster it using dendrograms. To do so:

- you represent each email body by the empirical probability distribution over the tokens it contains (simply estimated by their relative frequencies);
- and make use of the Hellinger distance.

What is the distance between the following two email bodies:

email 1: *ski sun money sun*

email 2: *sun ibm sun apple money sun money sun*

The non-zero components of two vectors are (order: *ski, sun, money, ibm, apple*):

email 1: $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}, 0, 0)$ email 2: $(0, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$

The Hellinger distance between the two is then:

$$\sqrt{\frac{1}{4} + \frac{2}{8}} = \frac{1}{\sqrt{2}}$$

QUESTION IV [3 pt]

You run the dendrogram clustering algorithm using complete linkage. At some point, it reaches a state where what remains to be clustered are the two clusters, G_1 and G_2, that have already been build so far, and two email bodies, B_1 and B_2. Here are the distances between each of them:

	B_1	B_2	G_1	G_2
B_1	0	0.7	0.6	0.2
B_2	0.7	0	0.5	0.3
G_1	0.6	0.5	0	0.4
G_2	0.2	0.3	0.4	0

Draw the dendrogram corresponding to the final clustering.

The two closest ones are B_1 and G_2 which will thus be merged in a new cluster, let's say G_3, the distances of which with the other two are (complete linkage):

- with B_2: 0.7
- with G_1: 0.6

The new closest group thus consist of B_2 and G_1, ending up in the following tree: