



Министерство науки и высшего образования
Российской Федерации

Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Московский государственный технический университет
имени Н. Э. Баумана (национальный
исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

Факультет «Информатика и системы управления»

ОТЧЕТ по лабораторной работе №7

по дисциплине

«Информационный поиск и извлечение информации из
текстов»

Студент группы ИУ9-21М

_____ С.С. Погосян
(подпись, дата)

Руководитель

_____ Н.В. Лукашевич
(подпись, дата)

1. Постановка задачи

Домашняя задача 1

- Система рубрикации должна классифицировать поток документов по двум рубрикам.
- Эксперт отнес к первой рубрике 75 документов, ко второй рубрике – 50 документов.
- Система отнесла:
 - - к первой рубрике 100 документов, из них 50 правильно.
 - - ко второй рубрике 40 документов, из них 30 правильно.
- Найти макро-характеристики качества классификации (точность, полноту, F-меру) - и микро-характеристики (точность, полноту, F-меру).

Домашняя задача 2

- Даны документы и их классы C1 и C2
- $D1=(X1, X2, X3)$ C1
- $D2=(X1, X2, X4)$ C1
- $D3=(X4, X5, X6)$ C2
- Определить класс документа на основе метода наивного Байеса
- $D4 (X1, X4, X5)$

2. Решение

<https://github.com/legion15q/sem2/blob/master/num7-8/%D0%9E%D1%82%D1%87%D0%B5%D1%82/Rwru6ysY4z4.jpg>

https://github.com/legion15q/sem2/blob/master/num7-8/%D0%9E%D1%82%D1%87%D0%B5%D1%82/yQLvY_KZvfw.jpg

2.1. Задача 1

$$1) TP + FP = 100; \quad TP = 50; \quad FP = 50; \quad FN = 10$$

$$\text{Precision} = \frac{50}{100} = \frac{5}{10} = \frac{1}{2}$$

$$\text{Recall} = \frac{50}{75} = \frac{2}{3}$$

$$F_{1\text{score}} = \frac{2 \cdot Pr \cdot Rec}{Pr + Rec} = \frac{2 \cdot \frac{1}{2} \cdot \frac{2}{3}}{\frac{1}{2} + \frac{2}{3}} = \frac{\frac{4}{6}}{\frac{7}{6}} = \frac{4}{7}$$

$$2) \text{Precision} = \frac{30}{40} = \frac{3}{4}$$

$$\text{Recall} = \frac{30}{50} = \frac{3}{5}$$

$$FP = 10; \quad FN = 50; \quad TP = 30$$

$$F_{1\text{score}} = \frac{2 \cdot Pr \cdot Rec}{Pr + Rec} = \frac{2 \cdot \frac{3}{4} \cdot \frac{3}{5}}{\frac{3}{4} + \frac{3}{5}} = \frac{2}{3} = 0,66$$

$$\text{Macro}F_1 = \frac{1}{N} \sum_{i=0}^N F_{1\text{score}} = \frac{1}{2} \left(\frac{2}{3} + \frac{4}{7} \right) = 0.61$$

$$\text{MacroPrecision} = \frac{PR_1 + PR_2}{2} = \frac{\frac{3}{4} + \frac{1}{2}}{2} = \frac{5}{8} = 0,625$$

$$\text{MacroRecall} = \frac{Rec_1 + Rec_2}{2} = \frac{\frac{2}{3} + \frac{3}{5}}{2} = \frac{19}{30} = 0,63$$

$$\text{MicroRecall} = \frac{TP_1 + TP_2}{TP_1 + TP_2 + FN_1 + FN_2} = \frac{50 + 30}{50 + 30 + 10 + 50} = 0.64$$

$$\text{MicroPrecision} = \frac{TP_1 + TP_2}{TP_1 + TP_2 + FP_1 + FP_2} = \frac{50 + 30}{50 + 50 + 10 + 30} = 0.57$$

$$\text{Micro}F_{1\text{score}} = \frac{2Pr_{\text{micro}}Rec_{\text{micro}}}{Pr_{\text{micro}} + Rec_{\text{micro}}} = 0.603$$

2.2. Задача 2

$$P(C_1) = \frac{N(C=C_1)}{N} = 0,667$$

$$P(C_2) = \frac{N(C=C_2)}{N} = 0,33$$

$$P(x_1 | C_1) = 0,25; \quad P(x_1 | C_2) = 0,11$$

$$P(x_4 | C_1) = 0,16; \quad P(x_4 | C_2) = 0,22$$

$$P(x_5 | C_1) = 0,083; \quad P(x_5 | C_2) = 0,22$$

$$\begin{aligned} P(\tilde{C}_1) &= P(C_1) \cdot P(x_1 | C_1) \cdot P(x_4 | C_1) \cdot P(x_5 | C_1) = \\ &= 0,667 \cdot 0,25 \cdot 0,16 \cdot 0,083 = 0,0022 \end{aligned}$$

$$\begin{aligned} P(\tilde{C}_2) &= P(C_2) \cdot P(x_1 | C_2) \cdot P(x_4 | C_2) \cdot P(x_5 | C_2) = \\ &= 0,33 \cdot 0,11 \cdot 0,22 \cdot 0,22 = 0,0017 \end{aligned}$$

$$\arg \max \{P(\tilde{C}_1), P(\tilde{C}_2)\} = 1 \Rightarrow D_4(x_1, x_4, x_5) \in C_1$$