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)

Москва — 2021

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

2.1. Задача 1

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

$$\text{Precision } = \frac{50}{100} = \frac{5}{10} = \frac{1}{2}$$
$$\text{Recall} = \frac{50}{75} = \frac{2}{3}$$
$$F_{1\text{score}} = \frac{2 \cdot Pr \cdot Rec}{Pr + Rec} = \frac{2 \cdot \frac{1}{2} \cdot \frac{2}{3}}{\frac{1}{2} + \frac{2}{3}} = \frac{\frac{4}{6}}{\frac{7}{6}} = \frac{4}{7}$$
$$\text{Macro}F_1 = \frac{F_{1\text{score}} + Pr + Rec}{3} = (\text{Среднее арифмитическое}) =$$
$$= \frac{\frac{4}{7} + \frac{1}{2} + \frac{2}{3}}{3} = \frac{73}{126} = 0,579$$
$$2)\ \text{Precision } = \frac{30}{40} = \frac{3}{4}$$
$$\text{Recall} = \frac{30}{50} = \frac{3}{5}$$
$$FD = 10; \quad FN = 50; \quad TP = 30$$
$$F_{1\text{score}} = \frac{2 \cdot Pr \cdot Rec}{Pr + Rec} = \frac{2 \cdot \frac{3}{4} \cdot \frac{3}{5}}{\frac{3}{4} + \frac{3}{5}} = \frac{2}{3} = 0,66$$
$$\text{Macro}F_1 = \frac{F_{1\text{score}} + Pr + Rec}{3} = \frac{\frac{2}{3} + \frac{3}{4} + \frac{3}{5}}{3} = \frac{7}{10} = 0,7$$
$$\text{MacroPrecision} = \frac{PR_1 + PR_2}{2} = \frac{\frac{3}{4} + \frac{1}{2}}{2} = \frac{5}{8} = 0,625$$
$$\text{MacroRecall} = \frac{Rec_1 + Rec_2}{2} = \frac{\frac{2}{3} + \frac{3}{5}}{2} = \frac{19}{30} = 0,63$$
$$\text{MicroPrecision} = \frac{TP_1 + TP_2}{TP_1 + TP_2 + FP_1 + FD_2} = \frac{50 + 30}{50 + 30 + 50 + 10} =$$
$$= \frac{8}{14} = \frac{4}{7} = 0,57$$
$$\text{MicroRecall} = \frac{TP_1 + TP_2}{TP_1 + TD_2 + FN_1 + FN_2} = \frac{50 + 30}{50 + 30 + 10 + 50} =$$
$$= \text{MicroPrecision} = 0,57 = \text{Micro}F_1$$

2.2. Задача 2

$$P\left(C_1\right) = \frac{N(C=C_1)}{N} = 0,667$$
$$P\left(C_2\right) = \frac{N(C=C_2)}{N} = 0,33$$
$$P\left(x_1 \mid C_1\right) = 0,25; \quad P\left(x_1 \mid C_2\right) = 0,11$$
$$P\left(x_4 \mid C_1\right) = 0,16; \quad P\left(x_4 \mid C_2\right) = 0,22$$
$$P\left(x_5 \mid C_1\right) = 0,083; \quad P\left(x_5 \mid C_2\right) = 0,22$$
$$P(\tilde{C_1}) = P\left(C_1\right) \cdot P\left(x_1 \mid C_1\right) \cdot C\left(x_4 \mid C_1\right) \cdot P\left(x_5 \mid C_1\right) =$$
$$= 0,667 \cdot 0,25 \cdot 0,16 \cdot 0,083 = 0,0022$$
$$P(\tilde{C_2}) = P\left(C_2\right) \cdot P\left(x_1 \mid C_2\right) \cdot P\left(x_4 \mid C_2\right) \cdot P\left(x_5 \mid C_2\right) =$$
$$= 0,33 \cdot 0,11 \cdot 0,22 \cdot 0,22 = 0,0017$$
$$\arg\max\left\{P(\tilde{C_1}), P(\tilde{C_2})\right\} = 2 \Rightarrow D_4\left(x_1, x_4, x_5\right) : C_2$$