Выбор признаков для классификации текстов

Мера полезности признака

- взаимная информация
- критерий χ^2
- частота

Взаимная информация

$$I(T,C) = \frac{N_{11}}{N} log_2 \frac{N \cdot N_{11}}{N_{1\bullet} \cdot N_{\bullet 1}} + \frac{N_{01}}{N} log_2 \frac{N \cdot N_{01}}{N_{0\bullet} \cdot N_{\bullet 1}} +$$

$$+ \frac{N_{10}}{N} log_2 \frac{N \cdot N_{10}}{N_{1\bullet} \cdot N_{\bullet 0}} + \frac{N_{00}}{N} log_2 \frac{N \cdot N_{00}}{N_{0\bullet} \cdot N_{\bullet 0}} +$$

N_{tc} – Количество документов «класса c»/«вне класса c»,
в которых «присутствует слово t»/«отсутствует слово t»

$$t,c \in \{0,\ 1\}$$

Критерий χ^2

$$X^2(T,C) = \frac{(N_{11}+N_{10}+N_{01}+N_{00})\times(N_{11}N_{00}-N_{10}N_{01})^2}{(N_{11}+N_{01})\times(N_{11}+N_{10})\times(N_{10}+N_{00})\times(N_{01}+N_{00})}$$

α	χ^2
0,1	2,71
0,05	3,84
0,01	6,63

Если $X^2(T,C) > \chi^2_\alpha$, то слово и класс зависимы

Частота

N_{11}

Алгоритм

- Вычисление меры близости каждого слова и класса
- Выбираем k слов, имеющих наибольшие значения

Лексикон

Термы	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt
веществ	1	1	1	1	0	1	1	0
вид	1	1	1	1	1	1	0	1
влаг	1	1	1	0	1	1	1	1
влажн	1	1	1	1	1	1	1	1
вод	1	1	1	1	1	1	1	1
воздух	1	1	0	1	1	1	1	1
дерев	0	1	1	1	0	0	1	1
животн	1	0	0	0	0	0	0	0
корн	0	0	1	1	1	1	1	1
лист	1	0	0	1	1	1	1	0
осадк	1	1	0	1	1	1	1	1
плод	1	0	0	0	1	1	1	1
поверхн	1	1	1	1	0	1	1	1
полив	0	1	0	0	1	1	1	1
почв	1	1	1	1	1	1	1	1
почвен	1	1	1	0	1	0	1	0
растен	1	1	1	1	1	1	1	1
содержан	1	1	1	1	0	1	1	0
температур	1	1	1	1	1	1	1	1

Для класса C_1

Термы	N_{11}	N_{01}	N_{10}	N_{00}
веществ	3	0	3	2
вид	3	0	4	1
влаг	3	0	4	1
влажн	3	0	5	0
вод	3	0	5	0
воздух	2	1	5	0
дерев	2	1	3	2
животн	1	2	0	5
корн	1	2	5	0
лист	1	2	4	1
осадк	2	1	5	0
плод	1	2	4	1
поверхн	3	0	4	1
полив	1	2	4	1
почв	3	0	5	0
почвен	3	0	2	3
растен	3	0	5	0
содержан	3	0	3	2
температур	3	0	5	0

Для класса C_2

Термы	N_{11}	N_{01}	N_{10}	N_{00}
веществ	1	0	5	2
вид	1	0	6	1
влаг	0	1	7	0
влажн	1	0	7	0
вод	1	0	7	0
воздух	1	0	6	1
дерев	1	0	4	3
животн	0	1	1	6
корн	1	0	5	2
лист	1	0	4	3
осадк	1	0	6	1
плод	0	1	5	2
поверхн	1	0	6	1
полив	0	1	5	2
почв	1	0	7	0
почвен	0	1	5	2
растен	1	0	7	0
содержан	1	0	5	2
температур	1	0	7	0

Для класса C_3

Термы	N_{11}	N_{01}	N_{10}	N_{00}
веществ	2	2	4	0
вид	3	1	4	0
влаг	4	0	3	1
влажн	4	0	4	0
вод	4	0	4	0
воздух	4	0	3	1
дерев	2	2	3	1
животн	0	4	1	3
корн	4	0	2	2
лист	3	1	2	2
осадк	4	0	3	1
плод	4	0	1	3
поверхн	3	1	4	0
полив	4	0	1	3
почв	4	0	4	0
почвен	2	2	3	1
растен	4	0	4	0
содержан	2	2	4	0
температур	4	0	4	0

Мера полезности признака: взаимная информация

Слова	C_1	C_2	C_3	минимум	максимум
веществ	0,204	0,056	0,311	0,056	0,311
вид	0,092	0,026	0,138	0,026	0,138
влаг	0,092	0,544	0,138	0,092	0,544
влажн	0	0	0	0	0
вод	0	0	0	0	0
воздух	0,199	0,026	0,138	0,026	0,199
дерев	0,003	0,092	0,049	0,003	0,092
животн	0,199	0,026	0,138	0,026	0,199
корн	0,467	0,056	0,311	0,056	0,467
лист	0,159	0,092	0,049	0,049	0,159
осадк	0,199	0,026	0,138	0,026	0,199
плод	0,159	0,199	0,549	0,159	0,549
поверхн	0,092	0,026	0,138	0,026	0,138
полив	0,159	0,199	0,549	0,159	0,549
почв	0	0	0	0	0
почвен	0,348	0,199	0,049	0,049	0,348
растен	0	0	0	0	0
содержан	0,204	0,056	0,311	0,056	0,311
температур	0	0	0	0	0

Мера полезности признака: критерий χ^2

Слова	C_1	C_2	C_3	минимум	максимум
веществ	1,60	0,38	2,67	0,38	2,67
вид	0,69	0,16	1,14	0,16	1,14
влаг	0,69	8,00	1,14	0,69	8,00
влажн	0	0	0	0	0
вод	0	0	0	0	0
воздух	1,90	0,16	1,14	0,16	1,90
дерев	0,04	0,69	0,53	0,04	0,69
животн	1,90	0,16	1,14	0,16	1,90
корн	4,44	0,38	2,67	0,38	4,44
лист	1,74	0,69	0,53	0,53	1,74
осадк	1,90	0,16	1,14	0,16	1,90
плод	1,74	1,90	4,80	1,74	4,80
поверхн	0,69	0,16	1,14	0,16	1,14
полив	1,74	1,90	4,80	1,74	4,80
почв	0	0	0	0	0
почвен	2,88	1,90	0,53	0,53	2,88
растен	0	0	0	0	0
содержан	1,60	0,38	2,67	0,38	2,67
температур	0	0	0	0	0

Мера полезности признака: частота

Слова	C_1	C_2	C_3
веществ	3	1	2
вид	3	1	3
влаг	3	0	4
влажн	3	1	4
вод	3	1	4
воздух	2	1	4
дерев	2	1	2
животн	1	0	0
корн	1	1	4
лист	1	1	3
осадк	2	1	4
плод	1	0	4
поверхн	3	1	3
полив	1	0	4
почв	3	1	4
почвен	3	0	2
растен	3	1	4
содержан	3	1	2
температур	3	1	4