Darbo užduotis

Sukurti programą SPAMui klasifikuoti panaudojant Bajeso teoremą. Ištirti priklausomybę tarp programoje naudojamų nustatymų ir klasifikatoriaus darbo efektyvumo (*false positive*, *true positive*).

Darbo vykdymo rekomendacijos

Skaičiuojant kiekvienos leksemos (simbolių seka iš a..Z, 0..9, $, ', "; Visi kiti simboliai - yra skyrikliai tarp leksemų) pasirodymų skaičių kiekviename duomenų rinkinyje patartina naudoti *hash* lenteles.

Šaltiniai:

1) http://www.paulgraham.com/spam.html
2) http://en.wikipedia.org/wiki/Bayesian_spam_filtering

Darbo atlikimo schema (viena iš galimų)

1) Sudaromi 2 failų katalogai su SPAMu ir NESPAMu (tai apsimokymo duomenys). Tarkime, SPAM katalogo failuose yra 300 žodžių, NESPAM – 250 žodžių.
2) Sudaroma duomenų struktūra (pvz. *hash* lentelė), į kurią įrašomi kiekvienos leksemos pasirodymo kiekiai SPAM ir NESPAM failuose, pvz. leksema „africa“ apsimokymo duomenyse – SPAM kataloge sutinkama 50 kartų, NESPAM kataloge – 15 kartų; leksema „earn“ – 200 ir 33 kartų atitinkamai

	SPAM	NESPAM
africa	50	15
earn	200	33

3) Apskaičiuojama kiekvienos leksemos spamiškumo tikimybė, t.y. tikimybę, kad failas su šiuo žodžiu yra SPAMas. Naudojama formulė:

$$P(S \mid W) = \frac{P(W \mid S)}{P(W \mid S) + P(W \mid H)}$$

čia P(W|S) – tikimybė, kad leksema W yra SPAMe, P(W|H) – tikimybė, kad leksema W yra NESPAMe. Mūsų atveju:

W	P(W|S)	P(W|H)	P(S|W)
africa	$\frac{50}{300} = 0.167$	$\frac{15}{250} = 0.06$	$\frac{0.167}{0.167 + 0.06} = 0.736$
earn	$\frac{200}{300} = 0.667$	$\frac{33}{250} = 0.132$	$\frac{0.667}{0.667 + 0.132} = 0.834$

Apsimokymo duomenų leksemų spamiškumo tikimybes P(S|W) tikslinga saugoti atskirai, bet ne perskaičiuoti kiekvieną kartą, kai reikia atlikti failo SPAMo klasifikacijos analizę.

4) Pateiktas SPAM klasifikacijai naujo failo turinys suskaidomas į leksemas. Nustatoma kiekvienos leksemos spamiškumo tikimybė (žr. ankstesnį punktą). Žodžiams, kurie sutinkami primą kartą, priskiriama spamiškumo tikimybė – 0,4.
   Pvz.: naujo failo, pateikto analizei, turinys: africa earn zzz.
5) Iš analizuojamo failo pasirenkamas tam tikras leksemų skaičius *N* (pvz. 15-20), kurių spamiškumo tikimybės yra maksimaliai nutolusios neutralios (pvz. 0,5)
6) Įvertinama tikimybė, kad failo pasirinktos leksemos rodo į jo priklausomumą SPAMui. Naudojama formulė

$$p = \frac{p_1 \cdot p_2 \cdots p_N}{p_1 \cdot p_2 \cdots p_N + (1 - p_1)(1 - p_2) \cdots (1 - p_N)},$$

čia p_i – pasirinktos leksemos spamiškumo tikimybė. Mūsų atveju, iš failo išrenkame 2 leksemas – „africa“ ir „earn“:

$$p = \frac{0.736 \cdot 0.834}{0.736 \cdot 0.834 + (1 - 0.736) \cdot (1 - 0.834)} = 0.933$$

Pastabos: Jei turime situacija

	SPAM	NESPAM
top	11	0
bottom	0	44

Tuomet šių leksemų spamiškumas įvertinamas

W	P(W|S)	P(W|H)	P(S|W)
top	$\frac{11}{300} = 0.037$		0.99, jei P(W|H)=0
bottom		$\frac{44}{250} = 0.176$	0.01, jei P(W|S)=0