T-61.281 Luonnollisen kielen tilastollinen käsittely
Vastaukset 4, ke 14.2.2007, 12:15–14:00 — Tiedonhaku
Versio 1.0

1. Muutetaan tehtävässä annettu taulukko sellaiseen muotoon, joka paremmin sopii ensimmäisten mittojen laskemiseksi.

kone 1	relevantit	ei relevantit
valitsi	4 *tp*	6 *fp*
ei valinnut	2 *fn*	9988 *tn*

kone 2	relevantit	ei relevantit
valitsi	6 *tp*	4 *fp*
ei valinnut	0 *fn*	9990 *tn*

Taulukko 1: *Muokatut taulukot. Taulukkoihin on merkattu myös tp (True Positives, oikein hyväksytyt), fp (False Positives, väärin hyväksytyt), fn (False Negatives, väärät hylätyt) ja fp (True Negatives, oikeat hylätyt)*

Seuraavassa taulukossa on annettu mittojen määritelmät ja sijoitettu luvut.

mitta	measure	määritelmä	kone 1	kone 2	Kuinka suuri osa
tarkkuus	precision	$\frac{tp}{tp+fp}$	$\frac{4}{4+6} = 40\%$	$\frac{6}{6+4} = 60\%$	löytyneistä relevantteja
saanti	recall	$\frac{tp}{tp+fn}$	$\frac{4}{4+2} = 67\%$	$\frac{6}{6+0} = 100\%$	relevanteista löytyi
hajoama	fallout	$\frac{fp}{fp+tn}$	$\frac{6}{6+9988} = 0.06\%$	$\frac{4}{4+9990} = 0.04\%$	ei-relevanteista palautettiin
täsmäävyys	accuracy	$\frac{tp+tn}{N}$	$\frac{4+9988}{10000} = 99.92\%$	$\frac{6+9990}{10000} = 99.96\%$	luokiteltiin oikein
virhe	error	$\frac{fp+fn}{N}$	$\frac{6+2}{10000} = 0.08\%$	$\frac{4}{10000} = 0.04\%$	luokiteltiin väärin

Taulukko 2: *Tulokset. Huomataan, että vain tarkkuuden ja palautuksen tulosprosentti liikkuu helposti miellettävällä alueella.*

F-mitta määritellään tarkuuden ja palautuksen avulla:

$$\frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

missä P on tarkkuus ja R on palautus. α säätää näiden välistä painotusta. Jos valitaan $\alpha = 0.5$ saadaan

$$\frac{2PR}{P+R}$$

Ensimmäiselle koneelle saadaan näinollen $F_1 = 50\%$ ja toiselle $F_2 = 75\%$.

Interpoloimatonta keskitarkkuutta laskiessa katsotaan tarkkuutta aina kun löydetään relevantti dokumentti ja keskiarvoistetaan näiden tarkkuuksien yli. Relevantit dokumentit, joita ei palautettu, lasketaan mukaan tarkkuudella 0.

$$\begin{aligned}
\mathrm{UAP}_1 &= \frac{1}{6}\left(\frac{1}{1}+\frac{2}{2}+\frac{3}{4}+\frac{4}{10}+0+0\right) = 53\% \\
\mathrm{UAP}_2 &= \frac{1}{6}\left(\frac{1}{1}+\frac{2}{2}+\frac{3}{3}+\frac{4}{5}+\frac{5}{7}+\frac{6}{9}\right) = 86\%
\end{aligned}$$

2. Tehtävänannossa annettiin sanojen dokumenttifrekvenssit: $df_1 = 21$ ja $df_2 = 500$. Lisäksi tiedetään että kokoelmafrekvenssit ovat $cf_1 = 101$ ja $cf_2 = 700$. Kaikenkaikkiaan kokoelmassa on $N = 10000$ dokumenttia. Käänteinen dokumenttifrekvenssi määritellään $IDF_i = \log_2 \frac{N}{df_i}$, joten sanalle w_1 se on $\log_2 \frac{10000}{21} = 8.9$ ja sanalle w_2 se on $\log_2 \frac{10000}{500} = 4.3$.

   Residuaalisen käänteisen dokumenttifrekvenssin (RIDF) kohdalla kirjan ensimmäisessä painoksessa on runsaasti virheitä. RIDF:n kantava idea perustuu seuraavanlaiselle päättelylle: Voimme mallintaa sanan esiintymistä Poisson-jakaumalla p. Tämä toimii hyvin sanoille, jotka ovat suhteellisen tasaisesti jakautuneet korpuksessa. Sisällöllisesti merkittävät sanat esiintyvät yleensä ryhmissä asiaa käsittelevän dokumentin sisällä, ja Poisson-jakauma antaa siis tällöin väärän ennusteen sanojen yleisyydestä. RIDF:ssä mitataan käänteisen dokumenttifrekvenssin ja Poisson-jakauman välistä eroa. Mitä suurempi ero, sitä enemmän sana kuvaa dokumentin sisältöä.

   Tässä siis Poisson-jakauman käyttölogiikka on seuraava: Approksimodaan, että dokumentissä esiintyy sana w_i keskimäärin $\lambda = \frac{cf_i}{N}$ kertaa. Todennäköisyys sille, että jossain tietyssä dokumentissä sana w_1 esiintyy k kertaa saadaan Poisson-jakaumasta

$$Poisson(k; \lambda) = e^{-\lambda} \frac{\lambda^k}{k!}$$

   RIDF määritellään siis

$$\mathrm{RIDF} = \mathrm{IDF} - \log_2 \left(\frac{1}{1 - Poisson(0, \lambda)} \right)$$

   Tässä siis Poisson-jakaumasta otetaan todennäköisyys, että dokumentissä esiintyy haluttu sana vähintään kerran $(1 - Poisson(0, \lambda))$.

   Sievennellään RIDF:n lauseketta:

$$\begin{aligned}
\mathrm{RIDF} &= \mathrm{IDF} - \log_2 \left(\frac{1}{1 - Poisson(0, \lambda)} \right) \\
&= \log_2 \frac{N}{df_i} + \log_2 (1 - Poisson(0, \lambda) \\
&= \log_2 \frac{N(1 - e^{-\frac{cf_i}{N}(\frac{\lambda}{0!})^0})}{df_i} \\
&= \log_2 \frac{N(1 - e^{-\frac{cf_i}{N}})}{df_i}
\end{aligned}$$

   Sijoitellaan kaavaan luvut:

$$\begin{aligned}
\mathrm{RIDF}_1 &= \log_2 \frac{10000(1 - e^{-\frac{101}{10000}})}{21} = 2.3 \\
\mathrm{RIDF}_2 &= \log_2 \frac{10000(1 - e^{-\frac{700}{10000}})}{500} = 0.44
\end{aligned}$$

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
Schumacher	0	1	0	1	0	0	0
rata	1	1	1	0	0	1	0
formula	1	0	1	1	0	0	0
kolari	0	0	1	1	0	0	0
galaksi	0	0	0	0	1	1	0
tähti	0	0	1	0	0	1	1
planeetta	0	0	0	0	0	1	1
meteoriitti	0	0	0	0	1	0	0

Taulukko 3: *Dokumentti–sana-matriisi*

Huomataan, että RIDF painotti sanaa w_1 2.5 kertaa enemmän kuin IDF. Molempien menetelmien mielestä w_1 on relevantimpi hakutermi kuin w_2.

3. Pyydetty dokumentti–sanamatriisi on esitetty taulukossa 3.

SVD-hajotelmassa (Singular Value Decomposition) puretaan $m \times n$ ($m > n$) matriisi A palasiksi:

$$A = USV^T$$

Tässä U on ortogonaalinen $m \times n$ matriisi, S on diagonaalinen $n \times n$ matriisi ja V ortogonaalinen $n \times n$ matriisi. Dimensiota voidaan pienentää ottamalla S:stä vain d suurinta singulaariarvoa ja matriiseista U ja V niitä vastaavat singulaarivektorit. Esitys on neliöllisen virheen mielessä optimaalisin.

Lasketut matriisit on esitetty taulukoissa 4, 5 ja 6. (Matlab palauttaa oletuksena hieman erilaiset matriisit kuin mitä yllä mainittiin: Nyt U on $m \times m$, S $m \times n$ ja V $n \times n$. Käytännön merkitystä asialla ei ole; singulaariarvot, niiden määrä ja tulos dimension pudotuksen jälkeen ovat samoja.)

Tiputetaan sisäinen dimensio kahteen jättämällä U ja V -matriiseista muut dimensiot pois ja ottamalla S-matriisista vain kaksi suurinta ominaisarvoa. Nyt dokumenttien samankaltaisuutta voi verrata matriisilla $B = SV^T$. Jos matriisin sarakeet skaalataan yhden pituisiksi, on helppo laskea korrelaatioita rivien välillä. Tällainen skaalattu matriisi on esitetty taulukossa 7 ja siitä lasketut korrelaatiot taulukossa 8. Sanojen samankaltaisuutta voitaisiin verrata matriisista $W = US$. Korrelaatiomatriisista huomataan, että formula-artikkelit ja tähtitiedeartikkelit korreloivat sisäisesti paljon enemmän kuin ristiin. Alunperin täysin korreloimattomata dokumentit d_5 ja d_7 korreloivat nyt selvästi. Olemme projisoineet datan 2-ulotteiseen avaruuteen ja samantyyppiset artikkelit ovat päätyneet lähekkäin tähän alempiulotteiseen avaruuteen.

Lopuksi vielä pieni varoitus: kirjan kappaleessa 15 on runsaasti pikkuvirheitä, kannattaa tarkastaa kirjan errata (http://www-nlp.stanford.edu/fsnlp/errata.html).

	dim_1	dim_2	dim_3	dim_4	dim_5	dim_6	dim_7	dim_8
Schumacher	-0.200	-0.336	0.290	0.115	0.823	0.007	0.121	-0.243
rata	-0.590	0.007	0.184	0.686	-0.232	-0.183	0.025	0.243
formula	-0.435	-0.464	-0.040	-0.225	-0.333	0.609	0.045	-0.243
kolari	-0.317	-0.361	-0.108	-0.494	0.071	-0.438	-0.285	0.485
galaksi	-0.200	0.400	0.602	-0.242	-0.053	0.028	-0.563	-0.243
tähti	-0.464	0.376	-0.408	-0.213	0.034	-0.345	0.275	-0.485
planeetta	-0.257	0.476	-0.234	-0.070	0.363	0.530	-0.007	0.485
meteoriitti	-0.026	0.116	0.534	-0.336	-0.132	-0.048	0.713	0.243

Taulukko 4: U

2.949	0	0	0	0	0	0
0	2.107	0	0	0	0	0
0	0	1.459	0	0	0	0
0	0	0	1.311	0	0	0
0	0	0	0	1.183	0	0
0	0	0	0	0	0.638	0
0	0	0	0	0	0	0.460
0	0	0	0	0	0	0

Taulukko 5: S

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
dim_1	-0.348	-0.217	0.099	0.352	-0.478	0.669	0.152
dim_2	-0.268	-0.156	0.325	0.611	0.499	-0.275	0.316
dim_3	-0.613	-0.210	-0.255	-0.187	-0.390	-0.559	0.130
dim_4	-0.323	-0.551	0.098	-0.460	0.474	0.279	-0.261
dim_5	-0.077	0.245	0.779	-0.440	-0.157	-0.030	0.328
dim_6	-0.512	0.598	0.099	0.124	0.094	0.048	-0.587
dim_7	-0.244	0.404	-0.440	-0.216	0.335	0.290	0.583

Taulukko 6: V

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
dim_1	-0.913	-0.924	-0.971	-0.634	-0.400	-0.768	-0.646
dim_2	-0.407	-0.384	-0.238	-0.773	0.917	0.640	0.764

Taulukko 7: *Skaalattu B*

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
d_1	1.000						
d_2	1.000	1.000					
d_3	0.984	0.988	1.000				
d_4	0.894	0.882	0.800	1.000			
d_5	-0.008	0.018	0.171	-0.455	1.000		
d_6	0.441	0.464	0.594	-0.008	0.894	1.000	
d_7	0.279	0.304	0.446	-0.180	0.958	0.985	1.000

Taulukko 8: *Korrelaatiot*