CAIM, examen parcial

5 de novembre de 2020. Temps: 1 hora 30 minuts

Exercici 1 (1 punt)

En el model vectorial els documents i les consultes (*queries*) es representen mitjançant vectors de pesos. Quins mètodes hem vist al llarg del curs que permeten reduir el nombre de posicions del vector amb zeros respecte la representació original del document o consulta (per tal de millorar la qualitat de les respostes)?

(la vostra resposta s'espera que sigui breu; només cal indicar els noms dels mètodes; no cal explicar en què consisteixen)

Resposta: En les consultes: la regla de Rocchio i altres mètodes d'expansió de consultes (*query expansion*). En els documents o consultes: diferents mètodes d'enriquiment que consisteixen en afegint termes semànticament relacionats (sinònims, hiperònims,...). Aquests mètodes bàsicament redueixen el nombre de components que són zero del vector tot mantenint-ne la dimensió. Es pot interpretar que tècniques com ara la lematització, *stemming* o eliminació de *stop words* també redueixen el nombre de zeros però cal tenir en compte que aquests mètodes en realitat redueixen el nombre de components del vectors (i ho fan per a qualsevol consulta o document, no sobre un document o consulta concret).

Nota: Els diferents mètodes de compressió (Elias γ, codis unaris autodelimitats,...) no són una resposta vàlida perquè (a) s'apliquen sobre les *posting list* (no sobre el vector de tf-idf) i, a més, (b) no reduirien el nombre components amb zeros sinó el nombre de bits a zero de la representació binària de les components. Usar diccionaris per a implementar els vectors no redueix el nombre de posicions amb zero del vectors, sinó que significa canviar el vector per una estructura de dades diferent.

Exercici 2 (2.5 punts)

En un país avançat, on els diners públics no es malversen ni en recerca militar ni en reprimir la dissidència, s'ha decidit de subministrar un test ràpid de COVID a tots els alumnes d'una universitat. La universitat té 3000 estudiants i el test dona un 5% de casos positius. El test falla en un 0.5% dels alumnes sobre el total d'alumnes de la universitat, indicant que tenen COVID quan en realitat no és el cas. Un examen mèdic infal·lible determina que un 6% dels alumnes de la universitat tenen COVID.

1. Empleneu la matriu de confusió (*confusion matrix*).

2. Mesureu l'eficàcia del test usant els següents indicadors:

$$Sensibilitat = \frac{tp}{tp + fn}$$

$$Especificitat = \frac{tn}{tn + fp}.$$

3. Doneu la fórmules per a sensibilitat i especificitat equivalents però expressades en funció dels conjunts R i A vistos a la teoria.

Resposta:

1. A partir de la informació de les dades de l'enunciat obtenim

$$\begin{aligned} |R| = tp + fn &= 3000 \cdot 6\% = 180 \\ |A| = tp + fp &= 3000 \cdot 5\% = 150 \\ fp &= 3000 \cdot 0.5\% = 15. \end{aligned}$$

Això ens permet deduir

$$\begin{aligned} tp &= 150 - fp = 135 \\ fn &= 180 - tp = 45 \\ tn &= 3000 - (tp + fp + fn) = 2805. \end{aligned}$$

Per tant, la matriu de confusió és

		Resultat del test	
		Positiu	Negatiu
Realitat	Sí	tp 135	fn 45
	No	fp 15	tn 2805

2.

$$\begin{aligned} Sensibilitat &= \frac{tp}{tp + fn} = \frac{135}{135 + 45} = \frac{3}{4}. \\ Especificitat &= \frac{tn}{tn + fp} = \frac{2805}{2805 + 15} = \frac{187}{188}. \end{aligned}$$

3.

$$\begin{aligned} Sensibilitat &= \frac{|R \cap A|}{|R|} \\ Especificitat &= \frac{|\overline{R \cup A}|}{|\overline{R}|} = \frac{|\Delta \setminus R \cup A|}{|\Delta \setminus R|}, \\ &= \frac{|\overline{R} \cap \overline{A}|}{|\overline{R}|} \end{aligned}$$

on $\overline{X}$ és el complementari del conjunt X i Δ és tot el conjunt de persones sobre les que s'aplica el test (el conjunt de tots els "documents"). En l'enunciat $|\Delta| = 3000$.

Exercici 3 (3 punts)

Tots coneixeu la llei de Zipf que relaciona la freqüència d'una paraula amb el seu rang. Una altra llei de G. K. Zipf indica que $p(f)$, la proporció de tipus (termes diferents) d'un text que tenen freqüència f, segueix

$$p(f) = cf^{-\beta}, \tag{1}$$

on c és una constant de normalització. En els texts reals tenim $\beta \approx 2$.

1. Obteniu una fórmula exacta per a c.
2. Obteniu una fórmula aproximada per a c suposant que el text és prou llarg.
3. Obteniu una fórmula aproximada per a c suposant que el text és prou llarg i $\beta = 2$.
4. Estimeu la proporció de tipus d'un text amb freqüència 1 assumint $\beta = 2$.
5. Estimeu la proporció de tipus d'un text amb freqüència superior a 2 assumint $\beta = 2$.

Resposta:

1. En un text amb N aparicions de tipus (és a dir, N *tokens*), tenim que per definició de $p(f)$, cal que se satisfaci la següent condició de normalització

$$\sum_{f=1}^{N} p(f) = 1. \tag{2}$$

Substituint $p(f) = cf^{-\beta}$ en la condició de normalització, s'obté

$$c = \frac{1}{\sum_{f=1}^{N} f^{-\beta}}.$$

2. Quan $N \to \infty$ i $\beta > 1$,

$$c \approx \frac{1}{\zeta(\beta)},$$

on

$$\zeta(\beta) = \sum_{f=1}^{\infty} f^{-\beta}$$

és la funció zeta de Riemann.

3. $\zeta(\beta) = \pi^2/6$ i per tant $c \approx 6/\pi^2$.

4. La proporció de tipus amb freqüència 1 és $p(1)$. Quan $\beta = 2$ i $N \to \infty$ tenim $p(1) \approx 6/\pi^2 \approx 0.607$.

5. La proporció de tipus d'un text amb freqüència més gran que 2 és

$$\begin{aligned} p(f > 2) &= 1 - p(1) - p(2) \\ &= 1 - c - \frac{c}{4} \\ &= 1 - \frac{5}{4}c. \end{aligned}$$

   Quan $\beta = 2$ i $N \to \infty$ obtenim

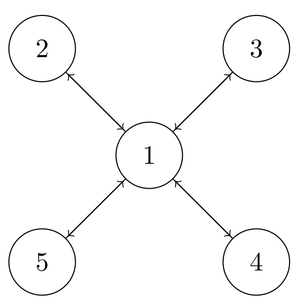
$$\begin{aligned} p(f > 2) &\approx 1 - \frac{15}{2\pi^2} \\ &\approx 0.240. \end{aligned}$$

Exercici 4 (3.5 punts)

Volem calcular els PageRank dels vèrtexs d'un graf estrella de n vèrtexs que consisteix en

- Un vèrtex central i $n - 1$ vèrtexs més.
- $n-1$ arestes del vèrtex central a la resta de vèrtexs i $n-1$ arestes de cada vèrtex no central al vèrtex central.

Per $n = 5$ aquest graf estrella és



El vèrtex central és el vèrtex 1 i la resta de vèrtexs s'etiqueten amb nombres entre 2 i n com en la figura. p_i és el pes de PageRank del vèrtex i-èssim. Volem saber-ne el següent:

1. Per $n = 5$ i $\lambda = 0$, els pesos p_1, p_2, p_3, p_4, p_5.

2. Per $n = 5$ i $\lambda = 1$, els pesos p_1, p_2, p_3, p_4, p_5.

3. Per n i λ qualssevol, fórmules per als pesos p_1,...,p_i,...,p_n en funció de n i λ.

4. Per $n = 5$ i $\lambda = 3/4$, els pesos p_1, p_2, p_3, p_4, p_5.

Resposta:

1. Amb $\lambda = 0$ el passejant aleatori camina sobre un graf complert i per tant, $p_1 = p_2 = p_3 = p_4 = p_5 = 1/5$.

2. Amb $\lambda = 1$ el passejant aleatori camina exclusivament sobre el graf de la figura. Per tant, $p_1 = 1/2$ i $p_2 + p_3 + p_4 + p_5 = 1/2$. Per simetria tenim que
$$p_2 = p_3 = ... = p_n, \tag{3}$$
que ens dona finalment $p_2 = p_3 = p_4 = p_5 = (1/2)/4 = 1/8$.

3. Per simetria, $p_2 = p_3 = ... = p_n$. Aplicant
$$p_i = \frac{1-\lambda}{n} + \lambda \sum_{j \to i} \frac{p_j}{k_j^{out}}$$
obtenim n equacions, que són
$$p_1 = \frac{1-\lambda}{n} + \lambda \sum_{i=2}^{n} p_i \tag{4}$$
i
$$p_i = \frac{1-\lambda}{n} + \lambda \frac{p_1}{n-1} \tag{5}$$
per $2 \leq i \leq n$, confirmant l'equació 3. Aplicant la condició de normalització obtenim l'equació $n+1$, que és
$$\sum_{i=1}^{n} p_i = 1.$$
Gràcies a la darrera equació
$$\sum_{i=2}^{n} p_i = 1 - p_1 \tag{6}$$
i per tant l'equació 4 esdevé
$$p_1 = \frac{1-\lambda}{n} + \lambda(1 - p_1),$$
que ens dona finalment
$$p_1 = \frac{1}{1+\lambda}\left(\frac{1-\lambda}{n} + \lambda\right). \tag{7}$$

p_i per $2 \leq i \leq n$ el podem obtenir de dues formes. La primera rau en el fet que les equacions 6 i 3 impliquen

$$p_1 = \frac{1 - p_1}{n - 1}. \tag{8}$$

Aplicant-hi l'equació 7, obtenim finalment

$$p_i = \frac{n + \lambda - 1}{(n-1)n(\lambda + 1)}. \tag{9}$$

La segona consisteix en aplicar 7 sobre 5.

4. A partir de la resposta a l'apartat següent amb $n = 5$ i $\lambda = 3/4$ obtenim $p_1 = 16/35 \approx 0.457$, $p_2 = p_3 = p_4 = p_5 = 19/140 \approx 0.136$.