CAIM, examen parcial

30 de novembre de 2016. Temps: 1 hora 50 minuts

Exercise 1 (1 point) Explain what pseudo-relevance feedback is (what it is for and what it consists of). Your answer should not contain more than 5 sentences.

Answer: Pseudo-relevance feedback is a method to improve the answer from a query without feedback from the user. It consists of calculating the k documents that resemble a query the most and then applying a query expansion technique to build a new query (e.g., Rocchio's rule or forming a new query based on terms in the top k documents). The process can be repeated various times. The user only gets the answer of the last query.

Exercise 2 (2 points) Suppose a vector model where the term weights are defined with the term frequency-inverse document frequency (tf-idf) scheme and the similarity between two documents is determined with the help of the cosine similarity. Recall that the tf-idf weight of a term i in a document d is defined as

$$tf_{d,i} \cdot idf_i$$

where $tf_{d,i}$ is the term frequency weight and idf is the inverse document frequency. Suppose that the term frequency weight is defined as

$$tf_{d,i} = \frac{f_{d,i}}{\max_j f_{d,j}},$$

The cosine similarity between the vectors of two documents, D_1 and D_2, is

$$sim(D_1, D_2) = \frac{D_1 \cdot D_2}{\sqrt{D_1 \cdot D_1}\sqrt{D_2 \cdot D_2}}.$$

Now suppose that we change the calculation of $tf_{d,i}$ (the weight of a term i in document d) using a new formula:

$$tf_{d,i} = \frac{f_{d,i}}{\sum_{e,j} f_{e,j}}.$$

Notice that now $tf_{d,i}$ is the relative frequency of the term over the whole ensemble of documents. With the new formula, D_1 becomes D_1' and D_2 becomes D_2'. Show the relationship between $sim(D_1', D_2')$ and $sim(D_1, D_2)$ using a mathematical argument.

Answer: We have $sim(D_1', D_2') = sim(D_1, D_2)$. The reason is that the cosine similarity measures the angle between the vectors D_1 and D_2 and that the new formula for $tf_{d,i}$ will only change the length of the vectors. A precise mathematical argument follows. Suppose that the new values of D_1 and D_2 are D'_1 and D'_2 respectively. First, notice that

$$D'_1 = \beta_1 D_1$$
$$D'_2 = \beta_2 D_2,$$

where

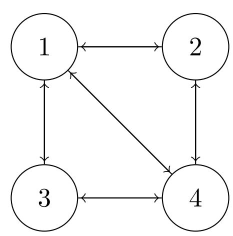
$$\beta_d = \frac{\max_j f_{d,j}}{\sum_{e,j} f_{e,j}}.$$

Then it is easy to see that

$$\begin{aligned} sim(D'_1, D'_2) &= \frac{D'_1 \cdot D'_2}{\sqrt{D'_1 \cdot D'_1}\sqrt{D'_2 \cdot D'_2}} \\ &= \frac{\beta_1\beta_2 D_1 \cdot D'_2}{\sqrt{\beta_1^2 D_1 \cdot D_1}\sqrt{\beta_2^2 D_2 \cdot D_2}} \\ &= \frac{D_1 \cdot D_2}{\sqrt{D_1 \cdot D_1}\sqrt{D_2 \cdot D_2}} \\ &= sim(D_1, D_2) \end{aligned}$$

as we wanted to prove.

Exercise 3 (3.5 points) Consider a small web defined by the following graph



1. Give the PageRank weights of every node for $\lambda = 0$.
2. Give the Google matrix for this system with a damping factor λ.
3. Give the PageRank equations and the PageRank weights of each node as a function of λ.
4. Give the PageRank weights for $\lambda = 1/2$.

Answer:

1. The Google matrix is defined as

$$G = \lambda M + \frac{1-\lambda}{4} J.$$

When $\lambda = 0$, M has no influence and the whole G defines a complete directed graph. In that graph, all nodes are topologically equivalent and thus $p_1 = p_2 = p_3 = p_4$. The constraint $\sum_{i=1}^{4} p_i = 1$ gives $p_1 = p_2 = p_3 = p_4 = 1/4$.

2. The adjacency matrix

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

gives the transition matrix

$$M = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 0 & 1/2 \\ 1/2 & 0 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}.$$

Then

$$\begin{aligned} G &= \lambda M + \frac{1-\lambda}{4} J \\ &= \begin{pmatrix} \alpha & \beta & \beta & \beta \\ \gamma & \alpha & \alpha & \gamma \\ \gamma & \alpha & \alpha & \gamma \\ \beta & \beta & \beta & \alpha \end{pmatrix}. \end{aligned}$$

with

$$\alpha = \frac{1-\lambda}{4}$$
$$\beta = \frac{3+\lambda}{12}$$
$$\gamma = \frac{\lambda+1}{4}$$

3. $\vec{p} = G^T \vec{p}$ gives four PageRank equations (the equation for p_2 and p_3 is the same):

$$p_1 = \frac{1-\lambda}{4} p_1 + \frac{\lambda+1}{4}(p_2 + p_3) + \frac{3+\lambda}{12} p_4 \tag{1}$$
$$p_2 = p_3 = \frac{3+\lambda}{12}(p_1 + p_4) + \frac{1-\lambda}{4}(p_2 + p_3) \tag{2}$$
$$p_4 = \frac{3+\lambda}{12} p_1 + \frac{\lambda+1}{4}(p_2 + p_3) + \frac{1-\lambda}{4} p_4 \tag{3}$$

The fifth PageRank equation is given by the normalization constraint, namely

$$\sum_{i=1}^{4} p_i = 1.$$

Notice that not only $p_2 = p_3$ but also $p_1 = p_4$ (swapping 1 and 4 in Eq. 1 yields Eq. 3 and the other way around). These identities allow one to express the normalization constraint as

$$2p_1 + 2p_2 = 1$$

obtaining

$$p_1 = \frac{1}{2} - p_2. \tag{4}$$

The fact that $p_2 = p_3$ allows one to transform the Eq. 1 into

$$p_1 = \frac{3-\lambda}{6} p_1 + \frac{\lambda+1}{2} p_2$$

Combining Eqs. 4 and 5, one gets

$$\begin{aligned} p_2 = p_3 &= f(3) \\ p_1 = p_4 &= f(1) \\ f(x) &= \frac{\frac{\lambda}{x} + 1}{4\left(\frac{2}{3}\lambda + 1\right)} \end{aligned}$$

after some algebra.

4.

$$\begin{aligned} p_2 = p_3 &= \frac{7}{32} = 0.21875 \\ p_1 = p_4 &= \frac{9}{32} = 0.28125 \end{aligned}$$

Exercise 4 (3.5 points) Suppose a posting list that consists of a sequence of n docid-frequency pairs, i.e.

$$x_1, y_1, ..., x_i, y_i, ...x_n, y_n$$

where x_i is the i-th docid and y_i is the frequency of occurrence of the term in document x_i. For instance, the sequence of integers

$$5, 2, 9, 7$$

indicates that the term appears two times in document 5 and 7 times in document 9.

1. We have compressed a posting list following the format above and obtained the following string of bits

$$0001101100101111111000100110101000010011 0$$

Decode the bit string to obtain the original posting list assuming that

   (a) Frequencies have been coded using unary self-delimiting codes as a sequence of 1's ending by a 0.

   (b) Docids have been coded using gap compression and Elias γ codes (the unary self-delimiting code within the Elias γ code is a sequence of 0's ending by a 1).

The first element of the bit string is an Elias γ code representing the number 13.

2. Provide an exact formula for B, the number of bits of a posting list that has been compressed with the procedure above. Provide an upper bound for B as a function of n, x_1, x_n and $\langle y \rangle$, the average value of y_i in the posting list.

Answer:

1. Segmenting the sequence, one gets a list of pairs

$$(0001101, 10), (010, 11111110), (00100, 110), (1, 0), (1, 0), (0001001, 10)$$

that encodes the list

$$(13, 2), (2, 8), (4, 3), (1, 1), (1, 1), (9, 2).$$

Undoing gap compression we finally obtain

$$(13, 2), (15, 8), (19, 3), (20, 1), (21, 1), (30, 2).$$

2. The length of the compressed list is

$$B = S + G + F,$$

where if S is the number of bits used for x_1, G is the number of bits used for the gaps and F is the number of bits used to code for the frequencies. Suppose that $e(z)$ and $u(z)$ are, respectively, the number of bits used to represent number z with an Elias γ code and a unary self-delimiting code. We have

$$\begin{aligned} F &= e(x_1) \\ G &= \sum_{i=2}^{n} e(x_i - x_{i-1}) \\ F &= \sum_{i=1}^{n} u(y_i). \end{aligned}$$

In general, a natural number z needs $\lfloor \log_2 z \rfloor + 1$ bits to represent all bits up to the most significant bit. Elias γ codes consist of a 1 bit surrounded by $\lfloor \log_2 z \rfloor$ zeros to its left and $\lfloor \log_2 z \rfloor$ bits to its right. Therefore, $e(z) = 2\lfloor \log_2 z \rfloor + 1$. Applying the definition of $e(z)$, we obtain

$$S = 2\lfloor \log_2 x_1 \rfloor + 1$$
$$G = n - 1 + 2H$$

with

$$H = \sum_{i=2}^{n} \lfloor \log_2 (x_i - x_{i-1}) \rfloor.$$

Obviously $u(z) = z$ and then

$$F = \sum_{i=1}^{n} y_i = n \langle y \rangle$$

Merging all the results above we obtain

$$B = 2\left(\lfloor \log_2 x_1 \rfloor + \sum_{i=2}^{n} \lfloor \log_2 (x_i - x_{i-1}) \rfloor \right) + n(\langle y \rangle + 1).$$

We want to bound B above involving only n, x_1, x_n and $\langle y \rangle$ as parameters. As only H involves additional information, our goal is to provide upper bounds of H that involve only the desired parameters. We consider two possibilities (only one suffices to get the maximum score in this exercise). *Example 1 (easy).* It is obvious that $x_i - x_{i-1} \leq x_n - x_1$ for $2 \leq i \leq n$. A tighter bound is

$$x_i - x_{i-1} \leq x_n - n + 2 - x_1$$

for $2 \leq i \leq n$. This is the largest gap between x_1 and x_2 when all the values from x_2 till x_n are consecutive numbers. With the tighter bound for the gap, one obtains

$$H \quad \leq \quad (n-1)\lfloor \log x_n - n + 2 - x_1 \rfloor$$

and finally

$$B \leq \lfloor 2(\log_2 x_1 \rfloor + (n-1)\lfloor \log(x_n - n + 2 - x_1) \rfloor + n(\langle y \rangle + 1).$$

Example 2 (not that easy). Knowing that $\lfloor \log z \rfloor \leq \log z$ and that in turn

$\log z \leq z - 1$ ($z - 1$ is the tangent at $z = 1$ of $\log z$) we obtain

$$\begin{aligned} H &\leq \sum_{i=2}^{n}(x_i - x_{i-1} - 1) \\ &= \sum_{i=2}^{n} x_i - \sum_{i=1}^{n-1} x_i - (n-1) \\ &= \sum_{i=1}^{n} x_i - x_1 - \left(\sum_{i=1}^{n} x_i - x_n\right) - (n-1) \\ &= x_n - x_1 - n + 1 \end{aligned}$$

and finally

$$B \leq 2(\lfloor \log_2 x_1 \rfloor + x_n - x_1 + 1) + n(\langle y \rangle - 1).$$