Interogarea setului de date

1 Concepte de bază

Procesul de regăsire a informațiilor este un algoritm care preia o interogare Q și un set de documente D_1, D_2, ...,D_n și identifică un coeficient de similaritate între fiecare document și interogarea Q. Acest model implica construirea vectorilor care reprezintă termeni din documente și un al vector care reprezintă termenii din interogare. Apoi o metodă de măsurare a apropierii(similarității) dintre fiecare vector de document și interogare trebuie aleasă.

2 Normalizarea datelor

Vectorii care au fost creați pentru a reprezenta in modelul Vector Space Model toate documentele de intrare trebuie normalizați folosind una dintre formulele prezentate în cursul 4.

Pentru acest exerciți recomand folosirea uneia din următoarele metode de normalizare:

- Binară: Utilizarea valorilor de „0” și „1”, adică 0 dacă un cuvânt nu este în document și 1 dacă cuvântul exilată în document indiferent de numărul de apariții
- Logaritmică pentru care se aplică formula

$$TF(d,a) = \begin{cases} 0, & if\ n(d,a) = 0 \\ 1 + \log\left(1 + \log\left(n(d,a)\right)\right), & otherwise \end{cases}$$

- Nominală: pentru care se aplică formula

$$TF(d,a) = \frac{n(d,a)}{max_\tau\, n\,(d,\tau)}$$

- Suma 1: pentru care se aplică formula

$$TF(d,a) = \frac{n(d,a)}{\sum_{i=0}^{k} n(d,a_i)}$$

- Invers Document Freqvency (IDF)

$$IDF(a) = \log\left(\frac{|D|}{|D_a|}\right)$$

În toate formulele de mai sus $n(d,a)$ -reprezintă numărul de apariții a cuvântului ”a” în documentul ”d”, $max_\tau\, n\,(d,\tau)$ reprezintă cuvântul care apare de cele mai frecvent în document, $|D|$ reprezintă numărul de documente din setul de date iar $|D_a|$ reprezintă numărul de documente din setul de date

care conțin termenul ”*a*”. Frecvența unui termen se va calcula folosind formula

$$TF(d,a) = TF(d,a) * IDF(a)$$

3 Metode de calcul a similarității între 2 vectori

Există mai multe metode de calcul a similarității între doi vectori. Fiecare metodă se alege în funcție de domeniu aplicabilității metodei respective. Printre cele mai cunoscute și folosite metode sunt calculul distanței Euclidiene sau a cosinusului unghiului dintre doi vectori.

Prezentăm în acest laborator doar trei metode:

1. Calculul similarității folosind distanta Euclidiană:

$$d(\vec{d}, \vec{d}') = \sqrt{\sum_{i=0}^{k} (d_i - d'_i)^2}$$

   unde *n* reprezintă nr. de trăsături caracteristice iar *x* și *x*' reprezintă cei doi vectori de intrare.

2. Calculul similarității folosind distanta Manhattan:

$$d(\vec{d}, \vec{d}') = \sum_{i=0}^{k} (|d_i - d'_i|)$$

3. Calculul similarității folosind cosinusul unghiului între 2 vectori:

$$sim(\overrightarrow{d_1}, \overrightarrow{d_2}) = \frac{\overrightarrow{d_1} \cdot \overrightarrow{d_2}}{\|\overrightarrow{d_1}\| \|\overrightarrow{d_2}\|}, \text{ unde}$$

$$\overrightarrow{d_1} \cdot \overrightarrow{d_2} = \sum_{i=0}^{k} d_{1i} d_{2i} \qquad \|d_1\| = \sqrt{\overrightarrow{d_1} \cdot \overrightarrow{d_1}}$$

4 Pași pentru găsirea documentelor relevante

1. Reprezentăm documentele prin vectori de frecvente de cuvinte (cum este prezentat în laboratorul anterior).
2. Normalizăm fiecare vector separat folosind o metodă de normalizare descrisă mai sus.
3. Citim o interogare (o propoziție) pe care o introduce utilizatorul de la tastatură (sau dintr-un fișier).
4. Aplicăm la interogarea citită aceleași etape de preprocesare ca și la cuvintele citite din fișiere.
5. Reprezentăm interogarea citită printr-un vector de dimensiunea dicționarului global creat în

etapa de reprezentare a documentelor.

6. Calculăm o similaritate intre interogarea citită și fiecare document din setul de date creat.
7. Ordonam documente astfel încât pe primele poziții se fie documentele cele mai similare.
8. Afișăm numele documentelor cele mai similare pe ecran si valoarea similarității obținută.