Structuri de date: Tema 3
Grafuri de cuvinte

Tudor Berariu
tudor.berariu@gmail.com

20 mai 2015

1 Pe scurt...

Pentru rezolvarea acestei teme va trebui construit un graf de cuvinte (pe baza unui text dat). Se va parcurge apoi acel graf pentru a construi fraze ce încep și se termină cu două cuvinte date.

2 Graful de cuvinte

Se dă un text în limba română. Se va parsa acest text și se vor separa cuvintele (fără semne de punctuație, cratime, etc.). Fiecare cuvânt va deveni nod într-un graf de cuvinte. Între fiecare două cuvinte consecutive din text va exista o muchie. Costul unei muchii va reprezenta un scor pentru acea secvență de două cuvinte calculat după cum urmează.

Pentru două cuvinte $word_A$ și $word_B$ vom calcula o matrice de contingență:

	$V = word_B$	$V \neq word_B$
$U = word_A$	O_{11}^{AB}	O_{12}^{AB}
$U \neq word_A$	O_{21}^{AB}	O_{22}^{AB}

unde:

- O_{11}^{AB} reprezintă numărul total de apariții ale sintagmei $word_1\ word_2$ în text.

- O_{12}^{AB} reprezintă numărul total de apariții ale unor sintagme în care $word_1$ este urmat de alt cuvânt în afară de $word_2$.
- O_{21}^{AB} reprezintă numărul total de apariții ale unor sintagme în care $word_2$ este precedat de alt cuvânt în afară de $word_1$.
- O_{22}^{AB} reprezintă numărul total de apariții ale unor sintagme în care primul cuvânt nu este $word_1$, iar cel de-al doilea nu este $word_2$.

Desigur, $O_{11}^{AB} + O_{12}^{AB} + O_{21}^{AB} + O_{22}^{AB} = M$ (numărul total de secvențe consecutive de două cuvinte).

Pe baza tabelei de contingență a două cuvinte $word_A$ și $word_B$ se calculează o măsură[1] pentru sintagma (colocația) $word_A word_B$:

$$\begin{aligned} OddsRatio(word_A, word_B) &= \log\left(\frac{O_{11}^{AB} \cdot O_{22}^{AB}}{O_{12}^{AB} \cdot O_{21}^{AB}}\right) \\ &= \log\left(O_{11}^{AB}\right) + \log\left(O_{22}^{AB}\right) \\ &\quad - \log\left(O_{12}^{AB}\right) - \log\left(O_{21}^{AB}\right) \end{aligned}$$

Deoarece este posibil ca unii termeni să fie zero, în practică se folosește:

$$\begin{aligned} OddsRatio^*(word_A, word_B) &= \log\left(\frac{\left(O_{11}^{AB} + 0.5\right) \cdot \left(O_{22}^{AB} + 0.5\right)}{O_{12}^{AB} \cdot O_{21}^{AB}}\right) \\ &= \log\left(O_{11}^{AB} + 0.5\right) + \log\left(O_{22}^{AB} + 0.5\right) \\ &\quad - \log\left(O_{12}^{AB} + 0.5\right) - \log\left(O_{21}^{AB} + 0.5\right) \end{aligned}$$

Deoarece măsura $OddsRatio(word_A, word_B)$ este o măsură a cât de puternică este alăturarea dintre cele două cuvinte, costul unui arc între nodurile $word_A$ și $word_B$ va fi:

$$\begin{aligned} Cost(word_A, word_B) &= 1 + \max_{(word_X, word_Y)} OddsRatio^*(word_X, word_Y) - \\ &\quad OddsRatio^*(word_A, word_B) \end{aligned}$$

Pentru două cuvinte ce nu au apariții consecutive în text nu va exista muchie între vârfurile corespunzătoare acestora.

[1] `http://www.collocations.de/AM/`

3 Cerințe

3.1 Cerința 1 : Construirea grafului de cuvinte

Se dă un text în limba română. Să se extragă cuvintele și să se construiască graful de cuvinte, calculându-se scorul muchiilor conform formulei din secțiunea 2.

Toate simbolurile `!.,?":();_*\n!£$%^&` vor fi eliminate, iar cratima va fi considerat separator între cuvinte.

Testarea se va face prin verificarea costurilor pentru câteva muchii alese la întâmplare.

3.2 Cerința 2: Construirea unei fraze

Se dau două cuvinte $word_{start}$ și $word_{end}$. Să se construiască o frază corespunzătoare drumului de cost minim între cele două noduri.

Pentru rezolvarea acestei cerințe se va implementa o coadă de priorități.

3.3 Bonus: Fraze de lungime fixă

Se dă un cuvânt $word_{end}$ și un număr n. Să se construiască fraza de lungime n ce se încheie cu $word_{end}$ de cost minim. Atenție: același cuvânt poate apărea de mai multe ori.

4 Trimiterea temei

Pentru trimiterea temei se va trimite o arhivă cu:

- toate fișierele sursă,
- `Makefile` care produce executabilul `words`.

Executabilul `words` va primi două argumente: numele fișierului de intrare (cu testele) și numele fișierului de ieșire.

4.1 Fișierul de intrare

Fișierul de intrare va conține pe prima linie numele fișierului cu textul pe baza căruia se va construi graful de cuvinte.

Pe linia a doua se va găsi un număr L. Pe următoarele L linii se vor găsi câte două cuvinte ce formează o sintagmă. Pentru fiecare pereche de cuvinte trebuie scris în fișierul de ieșire costul arcului dintre ele (cerința 1).

Pe linia $2 + L + 1$ se va găsi un număr M. Pe următoarele M linii se vor găsi câte două cuvinte. Pentru fiecare dintre acestea trebuie afișată fraza corespunzătoare drumului de cost minim dintre acestea (cerința 2).

Pe linia $2+L+1+M+1$ se va găsi un număr N. Pe următoarele N linii se vor găsi câte un număr n și un cuvânt w. Pentru fiecare dintre acestea se vor scrie în fișierul de ieșire propozițiile de cost minim de dimensiune n ce se termină cu w.

Vezi exemplu în Secțiunea5.2.

4.2 Fișierul de ieșire

Fișierul de ieșire va avea L linii pe care se va găsi câte o valoare reală corespunzătoare costului arcului descris la linia $3 \leq l \leq 2 + L$.

Următoarele M linii corespund frazelor construite pentru drumul de cost minim între cuvintele date pe linia $2+L+2 \leq m \leq 2+L+1+M$ în fișierul de intrare.

Pentru cerința de la bonus, se vor produce N grupuri de linii, fiecare având următoarea componență: o linie cu numărul sol_n de soluții, urmate de sol_n cu frazele respective.

Vezi exemplu în Secțiunea5.3.

5 Exemplu

5.1 Graful de cuvinte

Fie fișierul `text1`:

```
Fisier de test...
...cu cinci linii de test,
...linii scurte,
...linii foarte scurte,
...linii SCURTE de test!
```

Cele 8 cuvinte din graf vor fi: *fisier*, *de*, *test*, *cu*, *cinci*, *linii*, *foarte*, *scurte*.

Sunt în total 16 bi-grame (sintagme formate din două cuvinte). Pentru a calcula costul arcului $linii \rightarrow scurte$, calculăm întâi matricea de contingență pentru acea bigramă.

	$V = scurte$	$V \neq scurte$
$U = linii$	$O_{11} = 2$	$O_{12} = 2$
$U \neq linii$	$O_{21} = 1$	$O_{22} = 11$

Scorul *OddRatio* va fi:

$$\begin{aligned} OddsRatio^*(word_A, word_B) &= \log(O_{11} + 0.5) + \log(O_{22} + 0.5) \\ &\quad - \log(O_{12} + 0.5) - \log(O_{21} + 0.5) \\ &= \log(2.5) + \log(11.5) - \log(2.5) - \log(1.5) = 2.06388 \end{aligned}$$

Pentru celelalte sintagme

Sintamga	$OddRatio^*$
fisier de	-2.78501
de test	5.24175
test linii	1.18958
test cu	3.3673
cu cinci	4.5326
cinci linii	2.37158
linii foarte	2.37158
linii scurte	2.03688
linii de	0.587787
scurte de	1.01523
scurte linii	2.03688
foarte scurte	2.78501

Tabela 1: Măsura $OddRatio^*$ calculată pentru toate bigramele din `text1`

Cea mai mare valoare corespunde sintagmei *detest* și aceasta va fi folosită pentru calculul costului arcelor:

$Cost(linii, scurte) = 5.24175 - 2.03688 + 1 = 4.20487$

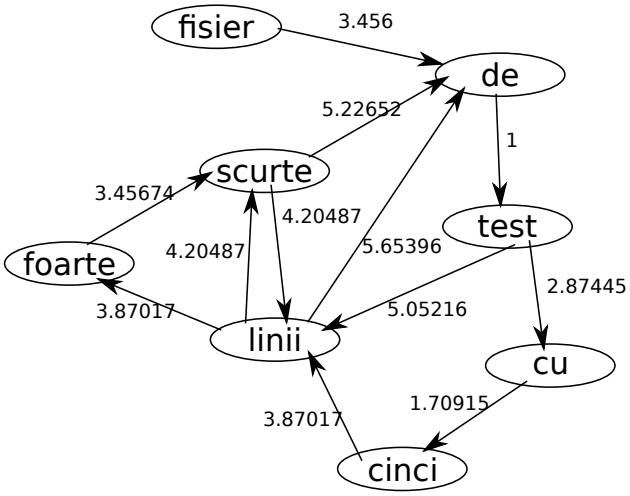

Figura 1: Graful corespunzător fișierului `text1`

5.2 Fișierul de intrare

```
data/text1
6
fisier de
linii scurte
foarte scurte
test linii
de test
scurte de
2
fisier scurte
scurte test
1
5 test
```

5.3 Fișierul de ieșire

```
3.45674
4.20487
3.45674
5.05216
```

```
1
5.22652
fisier de test linii scurte
scurte de test
1
cu cinci linii de test
```