MTA SZTAKI DSD

Department of Distributed Systems

Többnyelvű dokumentum nyelvének megállapítása

MSZNY 2011

Vajna Miklós
Pataki Máté

Probléma

- Természetes nyelvű dokumentum nyelvének a megállapítása
- Megoldott probléma egynyelvű dokumentumokra
- KOPI Plágiumkereső
  - megbízhatatlan eredetű
  - hibás (programkódot tartalmazó)
  - többnyelvű (szótár)

Új algoritmus - célok

- Az algoritmussal szemben az alábbi elvárásokat támasztottuk:
  1. Jelezze, ha a dokumentum több nyelven íródott, és nevezze meg a nyelveket
  2. Az algoritmus gyors legyen
  3. A szöveget csak egyszer kelljen végigolvasni
  4. Ne szótár alapú legyen (kódolási és betanítási problémák miatt)

Eredeti n-gram

- Csak egyszer kell végigolvasni a dokumentumot
- Meg lehet állapítani, hogy a dokumentum milyen nyelven íródott
- Még a kódolását is meg tudja határozni
- ~~több nyelven íródott dokumentumok~~

Eredeti n-gram

- a dokumentumban leggyakrabban használt nyelvet jól megállapítja
- második leggyakoribb nyelv már nem a második
- nyelvek hasonlítanak egymásra

→ nyelvek közötti hasonlósági metrika

Eredeti n-gram – magyar nyelv

1. _	18. _a	35. _a_	52. te	69. tt	86. _f
2. e	19. b	36. en	53. és	70. ke	87. az
3. a	20. d	37. ö	54. _s	71. _v	88. zt
4. t	21. a_	38. n_	55. al	72. ás	89. ár
5. s	22. v	39. _k	56. ta	73. ak	90. _n
6. l	23. t_	40. j	57. í	74. _é	91. ko
7. n	24. sz	41. ._	58. _h	75. ny	92. _A
8. k	25. el	42. i_	59. _t	76. tá	93. _sz
9. i	26. ,	43. eg	60. an	77. c	94. is
10. r	27. ,_	44. p	61. ze	78. re	95. ve
11. z	28. h	45. _e	62. me	79. to	96. gy_
12. o	29. k_	46. u	63. at	80. A	97. ít
13. á	30. .	47. le	64. l_	81. e_	98. _b
14. é	31. et	48. ó	65. es	82. ü	99. ra
15. g	32. gy	49. er	66. ő	83. ne	100. or
16. m	33. s_	50. f	67. y_	84. os	
17. y	34. _m	51. ek	68. z_	85. ál	

Eredeti n-gram

- 400-as lista
- $r_{max} = 400^2$
- $h_{százalékos} = (r_{max} - r) / (r_{max} / 100)$

Eredeti n-gram

- Szeged Wikipédia szócikkek

Magyar
1. magyar: 35.49
2. breton: 27.70
3. szlovák: 27.42
4. eszperantó: 26.98
5. közép-frízi: 26.79

Angol
1. angol: 44.37
2. skót: 35.67
3. romans: 35.34
4. német: 33.74
5. román: 33.73

Német
1. német: 57.13
2. holland: 38.15
3. közép-fríz: 37.71
4. dán: 37.48
5. fríz: 36.58

Olasz
1. olasz: 35.21
2. román: 33.95
3. katalán: 33.46
4. spanyol: 32.18
5. romans: 31.78

Eredeti n-gram

- kétnyelvű, 50-50 százalékban kevert dokumentumok

Magyar-angol
1. angol: 40.80
2. magyar: 39.45
3. skót: 38.41
4. afrikaans: 34.69
5. közép-fríz: 34.19

Angol-német
1. német: 53.47
2. angol: 44.14
3. fríz: 40.98
4. közép-fríz: 40.61
5. holland: 40.08

Magyar-olasz
1. olasz: 49.56
2. romans: 45.25
3. katalán: 41.60
4. latin: 41.26
5. román: 41.18

…

10. magyar: 38.02

Magyar-francia
1. francia: 38.16
2. katalán: 36.74
3. eszperantó: 34.26
4. spanyol: 34.08
5. romans: 33.71

…

7. magyar: 33.2

Új algoritmus

- Hasonlósági metrika
  - nyelvminták hasonlósága

Magyar
1. breton: 104 541
2. közép-fríz: 104 751
3. svéd: 106 068
4. eszperantó: 106 469
5. afrikaans: 106 515

Angol
1. skót: 85 793
2. francia: 88 953
3. katalán: 89 818
4. latin: 90 276
5. romans: 92 936

Olasz
1. romans: 79 461
2. román: 85 232
3. katalán: 85 621
4. spanyol: 86 138
5. latin: 86 247

Új algoritmus

- Hasonlósági metrika

$$h_i' = h_i \quad ha \quad i = 1$$

$$h_i' = h_i - \frac{\sum_{k=1}^{i-1} h_k \times h_{LiLk}}{\sum_{k=1}^{i-1} h_i} \quad ha \quad i > 1$$

Az algoritmus tulajdonképpen minden nyelv valószínűségét csökkenti az előtte megtalált nyelvek valószínűségével, így kompenzálva a nyelvek közötti hasonlóságból adódó torzulást.

Új algoritmus

- Szeged Wikipédia szócikkek

Magyar
1. magyar: 35.49
2. kínai: 2.09
3. japán (euc jp): 1.81
4. koreai: 1.70
5. japán (shift jis): 1.58

Angol
1. angol: 44.21
2. nepáli: 3.84
3. kínai: 2.53
4. vietnami: 2.08
5. japán: 1.14

Német
1. német: 57.13
2. kínai: 2.55
3. japán (shift jis): 2.19
4. japán (euc jp): 1.93
5. nepáli: 1.27

Olasz
1. olasz: 35.21
2. kínai: 1.07
3. perzsa: 0.68
4. japán: 0.57
5. jiddis: 0.55

Új algoritmus

- kétnyelvű, 50-50 százalékban kevert dokumentumok

Magyar-angol
1. angol: 40.80
2. magyar: 9.40
3. thai: 1.54
4. armeniai: 1.39
5. koreai: 1.37

Angol-német
1. német: 53.47
2. angol: 7.79
3. walesi: 2.08
4. fríz: 1.48
5. nepáli: 1.44

Magyar-olasz
1. olasz: 49.56
2. magyar: 7.44
3. walesi: 2.31
4. breton: 1.92
5. ír: 1.68

Magyar-francia
1. francia: 38.16
2. magyar: 2.11
3. thai: 1.42
4. koreai: 1.16
5. kínai: 0.70

Új algoritmus

10% angol, 90% magyar: 1. magyar: 38.01 2. koreai: 1.53 3. thai: 1.20 4. japán (euc): 1.14 5. japán (shift): 1.09	40% angol, 60% magyar: 1. angol: 37.62 2. magyar: 5.41 3. japán (euc): 1.47 4. thai: 1.46 5. japán (shift): 1.45	70% angol, 30% magyar: 1. angol: 44.92 2. vietnámi: 1.74 3. mingo: 1.67 4. kínai: 1.46 5. armén: 1.36
20% angol, 80% magyar: 1. magyar: 37.93 2. thai: 1.18 3. koreai: 1.17 4. japán: 1.16 5. armén: 1.11	50% angol, 50% magyar: 1. angol: 40.93 2. magyar: 5.30 3. thai: 1.49 4. japán (shift): 1.47 5. japán (euc): 1.37	80% angol, 20% magyar: 1. angol: 46.56 2. vietnámi: 2.07 3. mingo: 2.00 4. japán: 1.47 5. walesi: 1.43
30% angol, 70% magyar: 1. magyar: 37.47 2. angol: 4.91 3. thai: 1.22 4. armén: 1.18 5. japán: 1.16	60% angol, 40% magyar: 1. angol: 41.66 2. magyar: 3.43 3. kínai: 1.50 4. vietnámi: 1.48 5. mingo: 1.45	90% angol, 10% magyar: 1. angol: 48.1 2. vietnámi: 1.51 3. nepáli: 1.40 4. thai: 1.05 5. kínai: 1.05

Konklúzió

- Felismeri a többnyelvű dokumentumokat
- Minimum 30% kell, hogy legyen a második nyelv aránya
- Ki tudtuk szűrni vele a rosszul konvertált és többnyelvű dokumentumok több mint 90%-át
- Beépítettük a KOPI Plágiumkereső rendszerbe

KOPI Portal

http://kopi.sztaki.hu

Köszönöm a figyelmet!

Web: *http://dsd.sztaki.hu*

Email: *vajna.miklos@sztaki.hu*