Jakub Janicki
14.05.2021

Sprawozdanie z laboratorium 5 – Klasteryzacja

Doświadczenie zostało przeprowadzone na 100 pierwszych liniach pliku lines.txt

1. Zostały zaimplementowane 4 metryki
    a. cosinusowa
    b. lcs
    c. levenstein
    d. dice
2. Zostały zaimplementowane 2 sposoby oceny jakości klasteryzacji
    a. Indeks Daciesa-Bouldina
    b. Indeks Dunna
3. Została zaimplementowana funkcja wyznaczająca stopliste i usuwa z tekstu wyznaczone elementy

```
[',', "'", ':', '.', ')', '(', 'TEL', 'LTD', '+7', '``', '812', '495', 'LLC', 'STR.', '1', 'A', '<', '>', 'FAX', '@', '6',
'5', 'INN']
```

4. Został zaimplementowany algorytm klasteryzacji

    Pierwsze kilkadziesiąt linii obliczone za pomocą algorytmu z użyciem metryki lcs

```
''PA INTERIOR BOLSHAYA LUBYANKA STREET 16/4 MOSCOW 101000 RUSSIA INN/KPP 7704550148//770801001 495-984-8611

''SSONTEX Sp.ZO.O.IMPORT-EXPORTUL PRZECLAWSKA 5 03-879 WARSZAWA POLAND NIP 113-01-17-669
''SSONTEX SP.ZO.O.IMPORT-EXPORT UL PRZECLAWSKA 5 03-879 WARSZAWA POLAND NIP 113-01-17-669 TEL./FAX :0048 022 217 6532--
SSONTEX SP.ZO.O IMPORT-EXPORT 03-879 WARSZAWA UL PRZECLAWSKA 5 NIP:113-01-17-669

''TOPEX SP Z O.O SPOLKA KOMANDYTOWA UL POGRANICZNA 2/4 02-285 WARSZAWA POLAND
TOPEX SP.Z.O.O. SP.K UL.POGRANICZNA 2/4 02-285 WARSZAWA.POLAND
TOPEX SP.Z.O.O. SP.K UL.POGRANICZNA 2/4,02-285 WARSZAWA POLAND
TOPEX SP.Z O.O. SP.K UL POGRANICZNA 2/4 02-285 WARSZAWA
TOPEX SP.Z O.O. SP.K UL.POGRANICZNA 2/4,02--285 WARSZAWA
TOPEX SP Z O O SP K. UL.POGRANICZNA 2/4,02-285 WARSZAWA POLAND
TOPEX SP.Z O.O. SP.K UL.POGRANICZNA 2/4,02-285 WARSZAWA T:0048 225730397 F:0048 2257 30400
TOPEX SP.Z O.O. SP.K UL.POGRANICZNA 2/4 02-285 WARSZAWA POLAND
TOPEX SP Z O.O SP.K UL.POGRANICZNA 2/4 02-285 WARSZAWA POLAND
TOPEX SP.Z O.O SPOLKA KOMANDYTOWA UL POGRANICZNA 2/4,02-285 WARSZAWA +48 22 57 30 300 +4822 57 30 400
TOPEX SP.ZO.O SPOLKA KOMANDYTOWAUL.POGRANICZNA 2/4 02-285 WARSZAWA POLAND

'MASTER PLUS CO.' 143000 RUSSIA MO ODINSOVO MOJAISKOE SHOSSE,153G +7495 7273939

2TIGERS GROUP LIMITED ROOM 504 JINSHAZHOU SHANGSHUI ROAD GUANGZHOU 510160

ALDETRANS 105066 MOSCOW RUSSIA TOKMAKOV LANE 11 495 641-03-89

A-LIFT JSC 1 PROSPEKT MARSHALA ZHUKOVA MOSCOW 123308 RUSSIA T 495 784-7961

ALISA 1/5 Derbenevskaya str. Moscow Russia Tel./Fax 495 987-13-07 postal code 115114

ALLIANCE-TRADE INN 7816391055 / KPP 784601001 190020 Saint Petersburg quay of the Obvodny channel 138 bulk 1 liter.B

ALTAIR LIMITED COMPANY 199004 SAINT-PETERSBURG 1 LINE H.20 LIT A OF.8-H

ARIVIST 198035 RUSSIA SAINT-PETERSBURG GAPSALSKAYA STR.,5 OFFICE 1-3; +78123277732 +781 23277729.VOLOKNO @ YAHOO.COM
ARIVIST 198035 RUSSIA SAINT-PETERSBURG GAPSALSKAYA STR.,5 OFFICE 1-3; +78123277732 FAX+ 78123277729
ARIVIST 198035 RUSSIA SAINT-PETERSBURG GAPSALSKAYA STR.,5 OFFICE1-3; +78123277732 FAX+ 78123277729
ARIVIST 198035 RUSSIA SAINT-PETERSBURG GAPSALSKAYA STR.,5 OFFICE1-3; +78123277732 + 78123277729
ARIVIST 198035 RUSSIA SAINT-PETERSBURG GAPSALSKAYA STR.,5 OFFICE1-3; TEL.+78123277732 +78123277729
ARIVIST 198035 RUSSIA SAINT-PETERSBURG GAPSALSKAYA STR.,5 OFFICE1-3; TEL.+78123277732 +78123277729 VOLOKNO @ YAHOO.COM
ARIVIST 198035 RUSSIA SAINT-PETERSBURG GAPSALSKAYA STR.,5 OFFICE1-3; TEL.+78123277732 FAX+78123277729 VOLOKNO @ YAHOO.COM

AVANPORT INN 7839413675 KPP 783901001 190020 SAINT PETERSBURG QUAY OF THE OBVODNY CHANNEL 134-136-138 BUILD 101 LIT A
AVANPORT INN 7839413675 KPP 783901001 190020 SAINT PETERSBURG QUAY OF THE OBVODNY CHANNEL,134-136-138 BUILD 101 LIT A
```

5. Obliczam jakość klasteryzacji korzystając z indeksów dla każdej z metryk

metric	davies_bouldin	dunn
lcs	0.854475	0.0942249
cosine	2.31604	0.000129199
dice	0.45015	0.0437318
levenstein	0.810033	0.0047619

6. Został zmierzony czas poszczególnych funkcji

metric	clustering	davies_bouldin	dunn
lcs	36.9807	32.1882	15.5649
cosine	0.699987	0.334777	0.0601475
dice	0.65501	0.515242	0.264692
levenstein	72.6199	11.245	2.05416

7. Komentarz
   - Analizując stoplistę która była wyznaczana na podstawie częstości występowania słowa należało dodać warunek który wykluczy słowa o długości większej niż 5, gdyż często były to części adresu. Dzieje się tak ponieważ przetwarzaliśmy pierwsze 200 linii tekstu. Przy całym tekście problem prawdopodobnie byłby znikomy ale wciąż występował.
   - Algorytm klasteryzacji działa prawidłowo dla większości przypadków ale zdarzają cię pojedyncze linie które zostały przyporządkowane nie prawidłowo. Dzieje się tak głównie dla linii które zawierają kilka niekompletnych informacji
   - Współczynnik sigma w algorytmie klasteryzacji został dobrany na podstawie oceny pierwszych 200 linii tekstu osobno dla każdej metryki. Dobieranie w taki sposób współczynnika prawdopodobnie wpływa negatywnie na jakość klasteryzacji
     - cosinusowa 0.3
     - lcs 0.65
     - dice 0.3
     - levenstein 0.03
   - W każdym przypadku jako liczba ngramow została przyjęta liczba 2
   - Dla metryki wartości indeksów są niespodziewane, ciężko stwierdzić czym jest to spowodowane
   - Analizując tabele czasów zauważamy, że metryka kosinusowa ora dice działają zdecydowanie szybciej niż lcs i levenstein. Levenstein okazał się najwolniejszy

7. Poprawienie jakości klasteryzacji

- Aby mieć pewność że żadna istotna dana często powtarzająca się np. RUSSIA nie została dołączona do stoplisty, należy ręcznie ją przygotować
- Lepsze wyznaczenie współczynnika sigma w jakości klasteryzacji np. na podstawie indeksów