Naloga 2 - Podobnost jezikov

Melanija Kraljevska

27 October 2019

1 Izbrani jeziki

Pri tej nalogi so obravnavani naslednji jeziki:

1. Slovanski jeziki:
   - slovenščina
   - srbščina
   - makedonščina
   - bolgarščina
   - ruščina
   - beloruščina
2. Romanski jeziki:
   - španščina
   - portugalščina
   - galicijiščina
   - francoščina
   - italijanščina
3. Germanski jeziki:
   - angleščina
   - nemščina
   - nizozemščina
   - norveščina
   - finščina
   - švedščina
   - danščina
4. Ostali:
   - albanščina
   - grščina

1.1 Preobdelava datotek

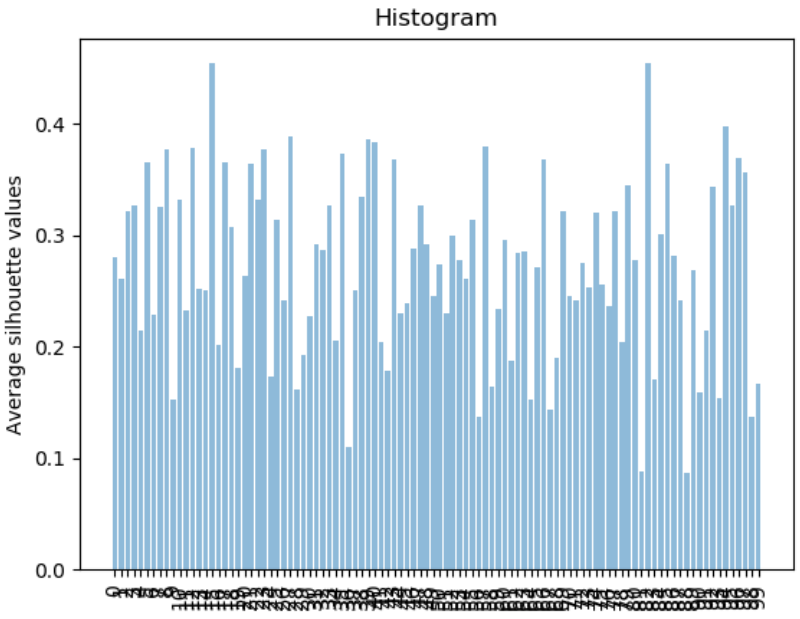
Vsaka datoteka oziroma jezik je predstavljen kot slovar, ki vsebuje vse trojke sosednjih črk v besedilu (interpukcijske znake in številke se ne upoštevajo). Pri besedah ki imajo manj kot 3 črk, se manjkajoče črke nadopolnijo s presledki, primer: 'of ', 'la ', 's '. Pri besedah ki imajo število črk ki ni deljivo z 3, ne nadomestimo s presledki, primer: cats - 'cat', 'ats'.

Primer kako izgleda slovar angleščine:

eng: 'uni': 16, 'niv': 5, 'ive': 15, 'ver': 44, 'ers': 22, 'rsa': 5, 'sal': 5, 'dec': 7, 'ecl': 7, 'cla': 9, 'lar': 8, 'ara': 8, 'rat': 12, 'ati': 65, 'tio': 89, 'ion': 103, 'of': 90, 'hum': 14, 'uma': 14, 'man': 17, 'rig': 56, 'igh': 57, 'ght': 56, 'hts': 22, 'pre': 7, 'rea': 16, 'eam': 1, 'amb': 1, 'mbl': 3, 'ble': 10, 'whe': 11, 'her': 29, 'ere': 18, 'eas': 12, 'rec': 8, 'eco': 6, 'cog': 4, 'ogn': 4, 'gni': 9, 'nit': 18, 'iti': 9, 'the': 151, 'inh': 2, 'nhe': 1, 'ren': 8, 'ent': 51, 'dig': 5, 'ign': 6, 'ity': 23, 'and': 111, 'equ': 16, 'qua': 14, 'ual': 14, 'ina': 7, 'nal': 28, 'ali': 13, 'lie': 4, 'ien': 8,..

Vsi slovarji jezikov pripadajo skupnemu slovarju, ki ima kot keys ime (skrajšano ime) jezika in kot values ima slovar kot v primeru zgoraj.

2 Rezultati razvrščanja



]

Figure 1: Histogram ki vsebuje povprečne vrednosti 100 silhuet pri 5 medoidov

Najboljša združitev algoritma k-medoids izmed 100, ima povprečno silhueto 0.45457056808840973. Dobili smo naslednje clustre:

- slv', 'src5', 'mkj', 'blg', 'rus', 'ruw'
- 'aln'
- 'spn', 'por', 'gln', 'frn', 'itn', 'eng'
- 'nrn', 'dns', 'swd'
- 'ger', 'dut', 'fri', 'grk'

Lahko opazimo da vsi slovanski jeziki se nahajajo v isti skupini. Albanščina je v svoji skupini. V tretjem clustru najdemo vse romanske jezike ter angleščina zraven. To nas ne preseneča, ker glede na različne vire ima 45% vseh angleških besed francoski izvor. Naslednji cluster vsebuje samo skandinavske jezike. V zadnji skupini najdemo ostale germanske jezike in grščina zraven. Algoritam je takšen da mora jezike razvrstit v točno 5 clustrov, tako da je pričakovano da se lahko najde tudi jezik v nekem clustru ki ni najbolj ustrezni, tukaj je recimo grščina.

Najslabša združitev ima povprečno silhueto 0.08722727068614688. Tukaj dobimo naslednje clustre:

- 'gln', 'ger', 'eng', 'dns', 'swd'
- 'mkj', 'blg', 'rus'
- 'slv', 'src5', 'ruw', 'aln'
- 'spn', 'frn', 'dut', 'fri', 'grk'
- 'por', 'itn', 'nrn'

Tukaj je rasporeditev precej slabša. Vsi romanski jeziki so razdeljeni v prvem, čertem in zadnjem clustru, skupaj z germanskimi jeziki. Slovanski jeziki so razdeljeni med sabo v drugem pa tretjem, kjer je zraven albanščina ki je precej različna od ostalih.

3 Napovedovanje jezika

V datoteko text.txt se nahaja del besedila iz španske novice. Funkcija recognize_language('text.txt') vrne 3 najbolj podobne jezike izmed 20 izbranih. Poleg tega vrne verjetnostjo zadetka. Funkcija na enak način predstavi besedilo kot slovar ki vsebuje frekvence trojk sosednjih črk. Potem izračuna razdaljo med besedilom in ostalimi jeziki. Podobnost dobimo kot 1-d. Verjetnost dobimo po formuli: (1-d)/sum(1-d), oziroma če je podobnost jezika i besedila s, verjetnost dobimo tako da to število delimo z vsoto podobnosti do vseh jezikov.

El independentismo ofrece síntomas de agotamiento. La movilización indefinida alentada desde el Govern de la Generalitat para responder a la sentencia del Tribunal Supremo sobre el 1-O pierde fuerza en las calles. Este sábado, 350.000 personas se manifestaron en Barcelona, según la Guardia Urbana, para defender la culminación del proyecto separatista y pedir la excarcelación de los líderes del procés, lo que supone un menos que en la huelga general del pasado 18 de octubre, cuando 525.000 secesionistas se congregaron en la capital catalana con el mismo fin. ...

Funkcija vrne naslednje rezultate:
spn 7.43 %
frn 7.11 %
gln 6.66 %

4 Dodatno

Na istih podatkih in z isto tehniko računanja razdalj med besedili uporabite tudi hierarhično razvrščanje. Na kratko (v največ treh stavkih) komentirajte rezultate.

V hc.py je implementirana metoda hierarhičnega ravrščanja. Za threshold smo se odločili za 0.43 max razdaljo med clustri. Dobimo naslednje clustre:

- 'grk'
- 'aln'
- 'ruw', 'slv', 'rus', 'src', 'mkj', 'blg'
- 'eng', 'itn', 'frn', 'spn', 'por', 'gln'
- 'swd', 'nrn', 'dns', 'ger', 'dut', 'fri'

Metoda nam vrne zelo dobre rezultate, do zdaj najboljše. Grščina in albaščina sta posebna skupina, tretja skupina so samo slovanski jeziki, četrta skupina vsebuje vse romanske jezike in angleščina. Zadnja skupina pa vse germanske jezike.

Na spletu najdite novičarske strani v prej izbranih dvajsetih jezikih in ponovite razvrščanje na novicah. Komentirajte rezultate.

Datoteka news vsebuje 20 tesktovnih datotek o različnih novicah v ustrezne jezike. Dobimo naslednje clustre:

- 'nrn', 'swd'
- 'aln'
- 'slv', 'src5', 'rus', 'ruw', 'itn', 'ger', 'dut', 'fri', 'dns'
- 'spn', 'por', 'gln', 'frn', 'eng'

- 'mkj', 'blg', 'grk'

Lahko rečemo na sta prva in druga skupina v redu. Četrta vsebuje romanske jezike in angleščino. Zadnja skupina ima dva zelo podobna jezika, grščina pa odstopa. Tretja skupina pa je zelo slabo sestavljena. Pri različnih poiskusih dobimo zelo različne rasporeditve.