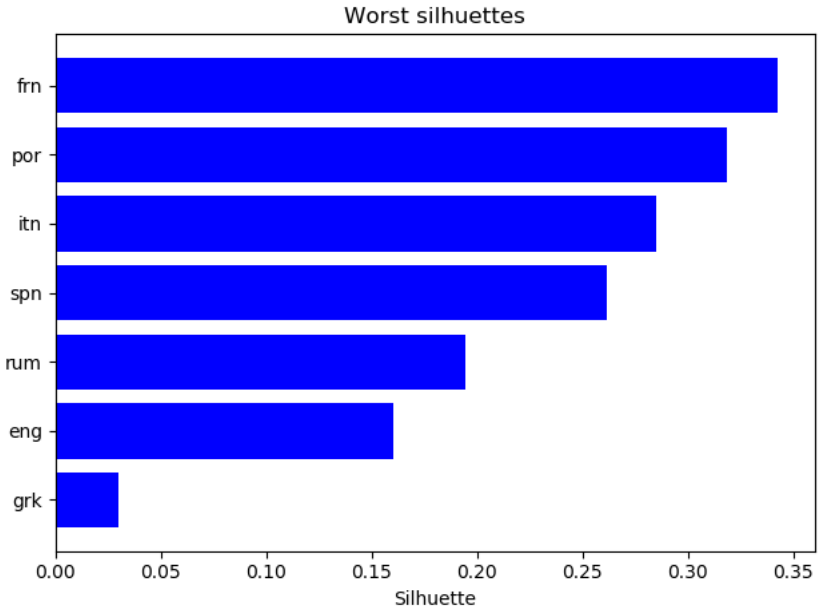
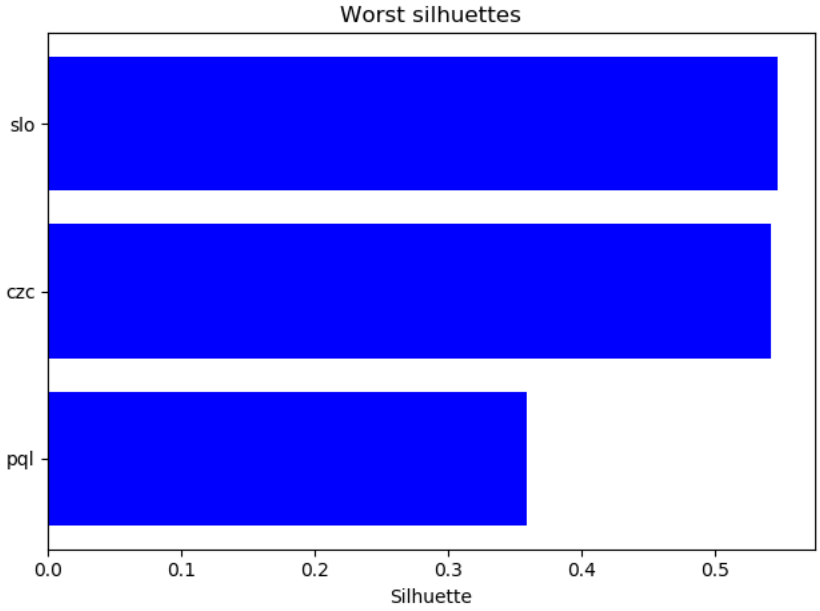
Izbrani jeziki Jeziki so bili izbrani tako, da so pokrivali zahtevane jezikovne skupine in da so bili čim bolj raznoliki in pisani v različnih abecedah. Vsi jeziki so evropski. Jeziki z abecedami, ki vsebujejo znake, ki niso kontekstno neodvisni (japonščina, kitajščina, korejščina) niso bili uporabljeni, saj jih Pythonova knjižnica *unidecode* slabo "prevede"v osnovne ASCII znake. Izbrani romanski jeziki so italijanščina, francoščina, portugalščina, španščina in romunščina. Germanski jeziki so nemščina, danščina, švedščina, nizozemščina, angleščina, norveščina in islandščina. Slovanski jeziki so slovenščina, slovaščina, srbščina, ruščina, bulgarščina, češčina in polščina. Ostali jeziki so grščina, litvanščina, estonščina in turščina. Besedila so bila predobdelana tako, da je bil celoten tekst spremenjen v male znake abecede in pretvorjen iz posameznih abeced v osnovne ASCII znake z uporabo knjižnice *unidecode*.

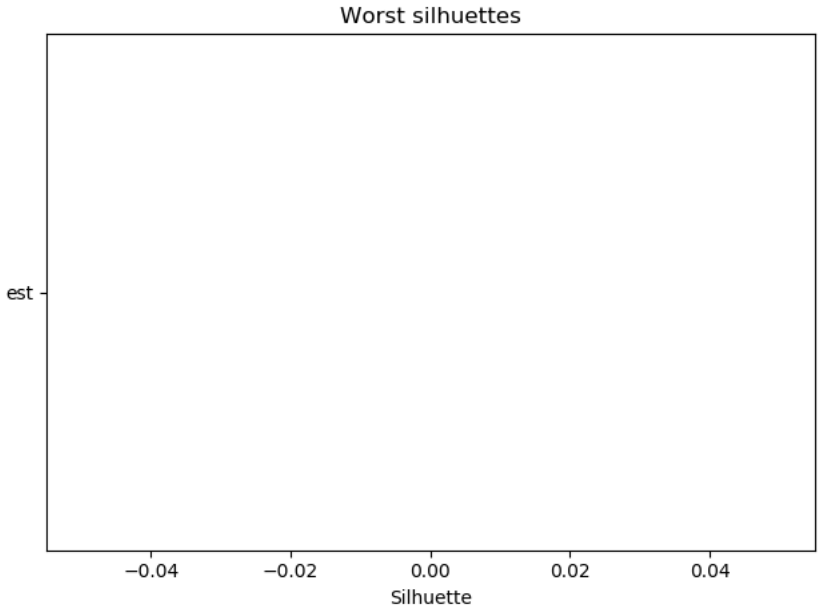
Rezultati razvrščanja Pri razvrščanju sta bili najslabša in najboljša porazdelitev izbrani na podlagi povprečne silhuete vseh držav. V primeru da je bila v gruči samo ena država, je le ta imela silhueto z vrednostjo 0. Pri najslabši porazdelitvi je imela povprečna silhueta vrednost 0.153, pri najboljši pa 0.287. Najboljše gruče so med seboj ločile slovanske, germanske in romanske jezike. V določenih gručah pride do vsiljicev, na primer angleščina med romanski jeziki. Jeziki, ki izrazito iztopajo pa se lahko celo pojavijo same v svoji gruči, na primer malteščina. Slabe gruče se oblikujejo kadar je bil medoid eden izmed iztopajočih jezikov.



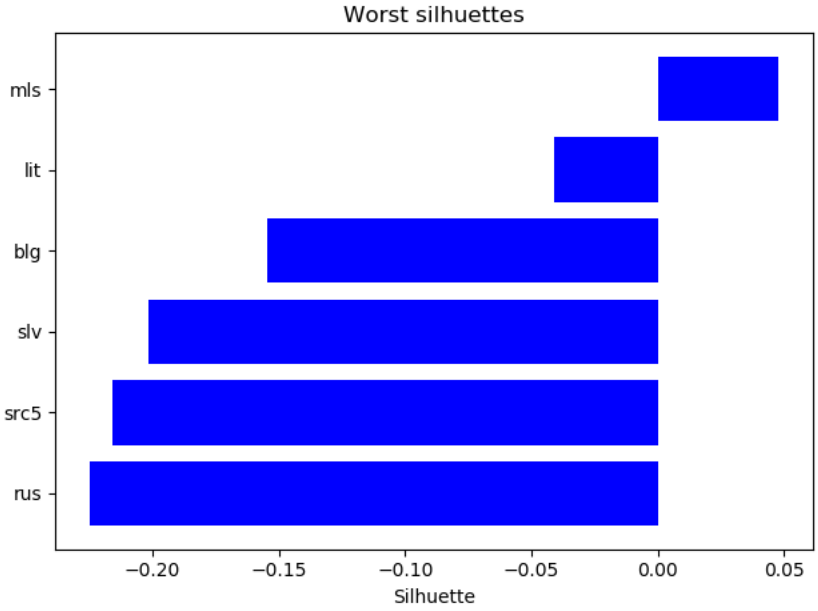
Slika 1: Silhuete slabe gruče 1.



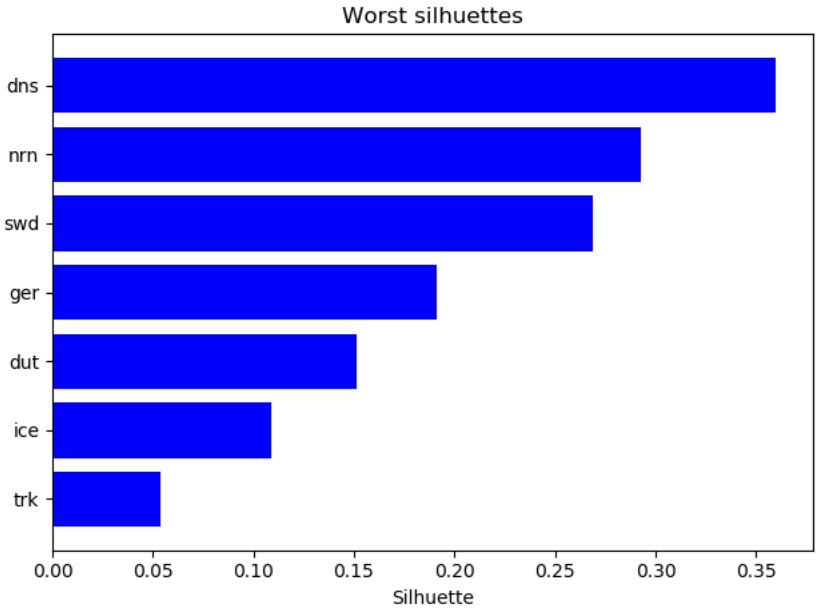
Slika 2: Silhuete slabe gruče 2.



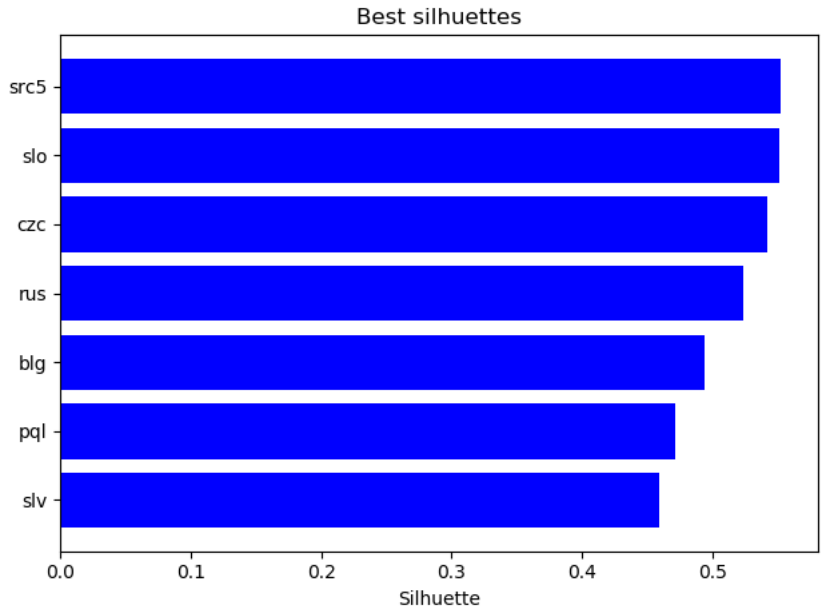
Slika 3: Silhuete slabe gruče 3.



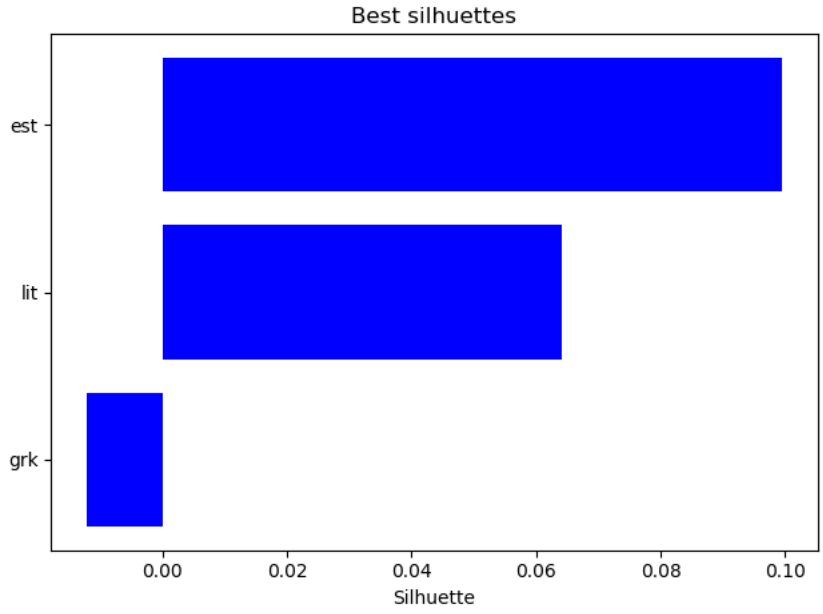
Slika 4: Silhuete slabe gruče 4.



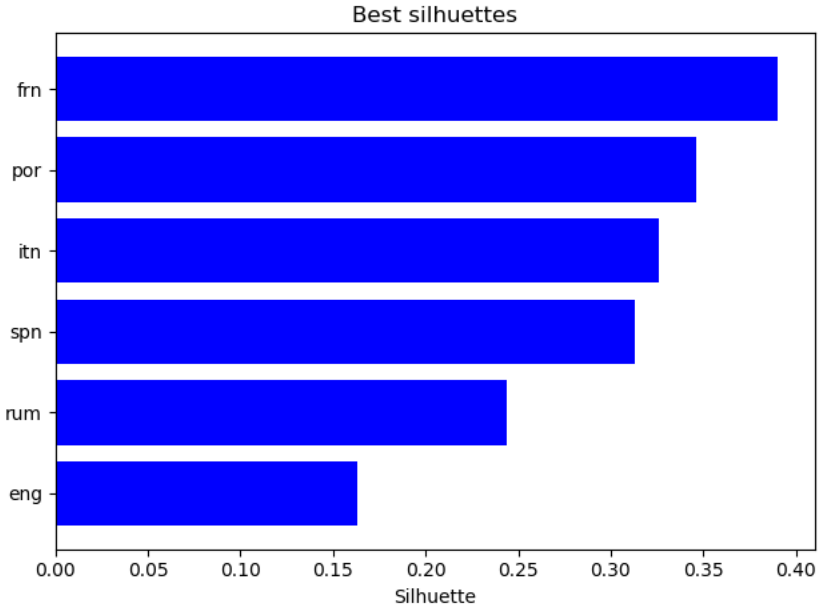
Slika 5: Silhuete slabe gruče 5.



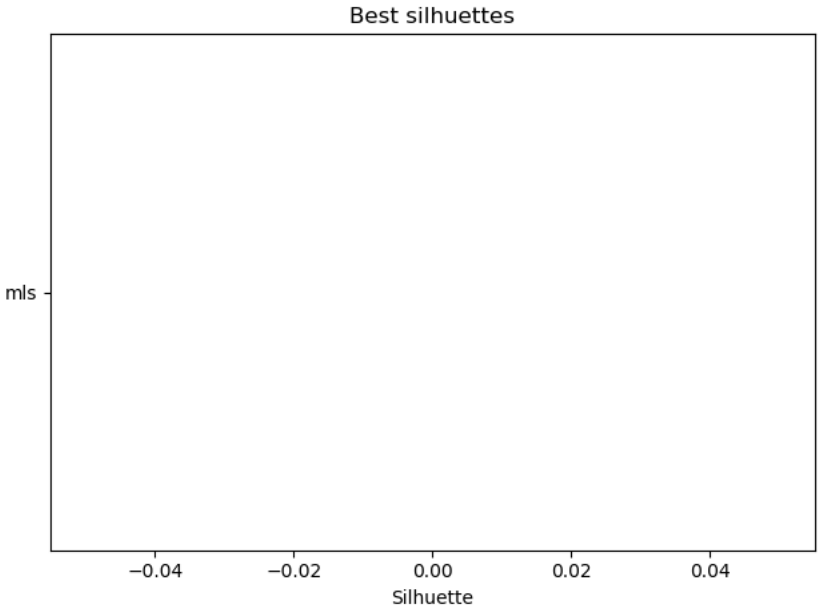
Slika 6: Silhuete dobre gruče 1 - slovanski jeziki.



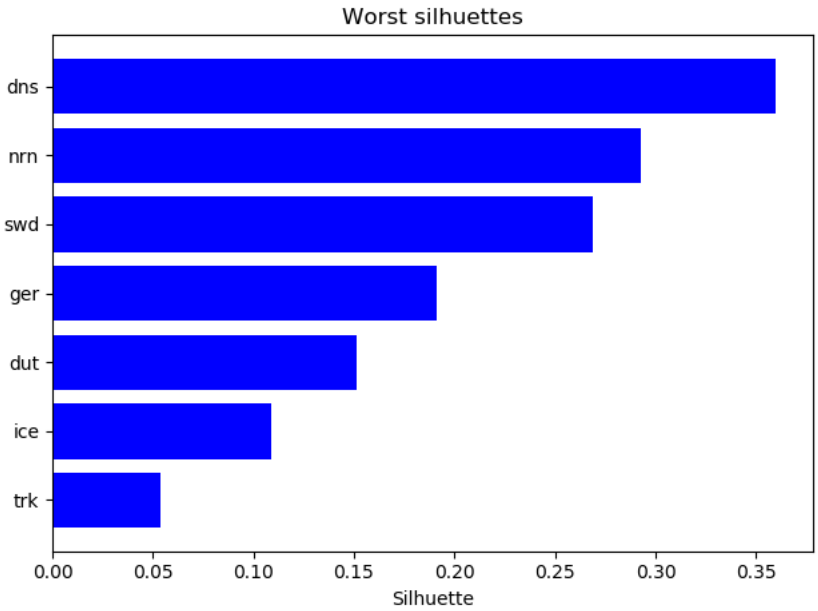
Slika 7: Silhuete dobre gruče 2 - baltska jezika z grščino, ki slabo pripada gruči.



Slika 8: Silhuete dobre gruče 3 - romanski jeziki z izstopajočo angleščino.



Slika 9: Silhuete dobre gruče 4.



Slika 10: Silhuete dobre gruče 5 - germanski jeziki z turščino.

Napovedovanje jezika Pri napovedi jezika se najprej izračuna frekvenca trojk za vse podane jezike, nato pa prav tako za podan odlomek. Sledi izračun kosinusne razdalje med frekvencami podanega teksta in frekvencami ostalih jezikov. Verjetnost se izračuna po formuli $p = 1 - razdalja$, nato pa se vzame tri jezike z največjimi verjetnostmi.

Tabela 1: Tabela jezikov z odlomki in s kakšnimi verjetnostmi pripadajo najbolj verjetnim jezikom.

jezik	tekst	najdene države in verjetnosti
švedščina	Kristerssons regering består av allt att döma av hans eget parti och Kristdemokraterna. Men de två andra alliansparti**er**na, Centern och Liberalerna, har tidigare sagt att man skulle rösta nej till M-ledarens så kallade ”3–2–1”-regering, där inte alla partierna ingår.	švedščina - 0.67 islandščina - 0.61 norveščina - 0.60
nemščina	Wie wird es nach den Midterm-Wahlen aussehen? Viele Meinungsforscher und Medien versuchen, als Reaktion auf die unzureichende Stimmungserfassung vor November 2016, neue Formen der Umfragen zu etablieren.	nemščina - 0.73 norveščina - 0.69 svedščina - 0.67
portugalščina	As novas orientações são bastante mais rígidas do que as publicadas em 1998, quando os pais eram aconselhados a desenvolver métodos alternativos à palmada em resposta a um comportamento indesejado.	portugalščina - 0.74 litvanščina - 0.73 ruščina - 0.63
češčina	To je vedle dalších podmínek nutné k tomu, aby si mohli podat žádost o belgické občanství. Belgické orgány před rozsudkem žádost několika britských úředníků o udělení občanství odmítly. Zdůvodňovaly to tím, že úředníci unijních institucí mají diplomatický status, což znamená, že v zemi mimo jiné neplatí daň z příjmu, která je v Belgii jedna z nejvyšších v Evropě.	češčina - 0.63 srbščina - 0.62 polščina 0.62

Analiza izbranih novic Iz spleta so bile pobrane različne novice v 20 jezikih. Če nad njimi izvedemo razvrščanje na podlagi medoidov, so silhuete slabše, torej so tudi nastale gruče slabše. Najslabša povprečna silhueta je 0.103, najboljša pa 0.183, kar je za 0.104 manj od najboljše povprečne silhuete pri gručenju na podlagi deklaracij v različnih jezikih. Razlogov za to je več. Novice so krajše, med seboj se razlikujejo v dolžini in vsebina ni enaka. Vse to drži pri deklaraciji človekovih pravic.

Izjava o izdelavi domače naloge. Domačo nalogo in pripadajoče programe sem izdelal sam.