Izbrani jeziki. Izbrani jeziki: danski, norveški, nemški, češki, slovaški, nizozemski, bosanski (cirilica), bosanski (latin), portugalski, španski, ruski, beloruski, slovenski, francoski, italijanski, estonski, latvijski, litovski, sumerski, angleški ter albanski.

Predobdelava datotek. Najprej sem prebral ime datoteke, ter ga pretvoril v ime jezika s pomočjo index.txt. Vse unicode znake v datoteki sem nato pretvoril z unidecode knjižnico, zamenjal vsa ločila ter dvojne presledke za enojni presledek, spremenil vse znake v male tiskane ter jo bral po tri znake naenkrat. Pojavitve posameznih trojk sem za vsak jezik hranil v objektu Point, ki ima slovar s trojkami kot ključi in število pojavitev kot vrednosti. Te objekte pa sem prav tako hranil v slovarju z jezikom kot ključ.

Porazdelitev vrednosti silhuet. Slika 1 prikazuje vsoto vrednosti silhuet za vsak jezik s 100 naključno izbranimi inicializacijami. Vrednost silhuete nam pove kako dobro smo jeziku določili skupino. Iz porazdelitve lahko vidimo, katerim državam algoritem dobro določi skupino in katerim ne. Iz porazdelitve lahko tudi vidimo, da se albanski jezik še posebej slabo grupira, saj ima negativno vrednost. To je verjetno zato, ker je zelo različen od vseh ostalih in zanj ni nobene dobre skupine.

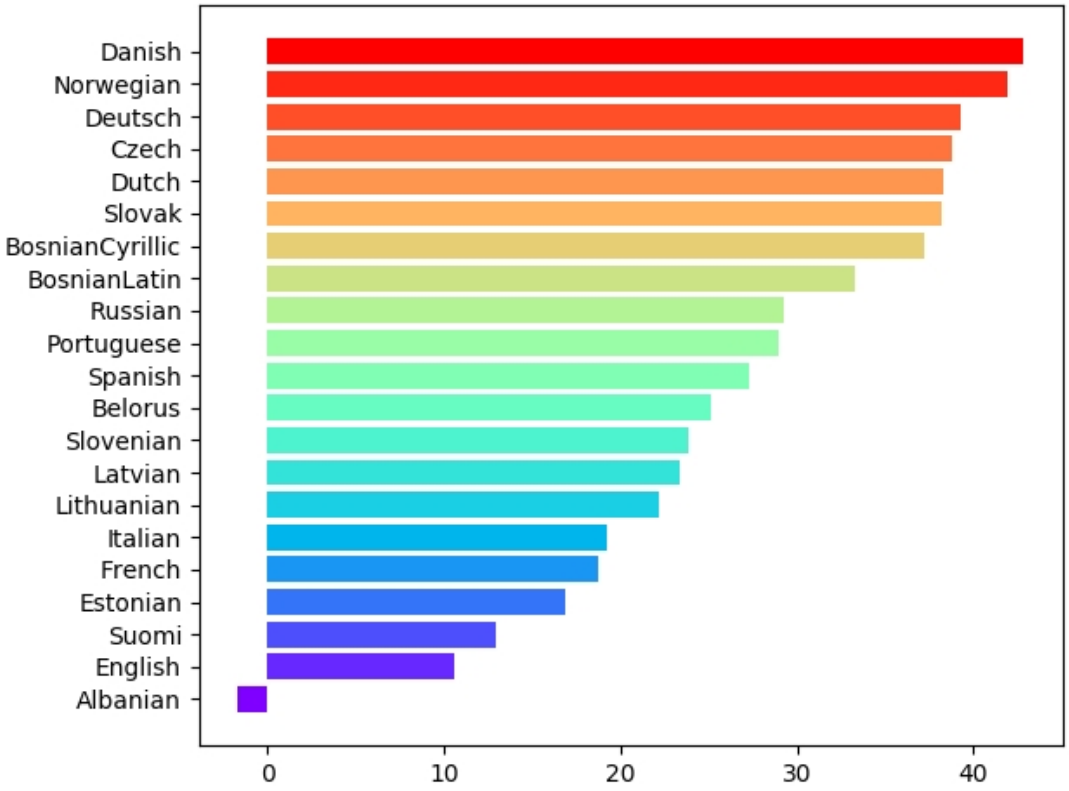

Рис. 1: Porazdelitev vrednosti silhuet.

Najboljša silhueta. Slika 2 prikazuje najboljšo silhueto ter skupine jezikov. V rdeči barvi lahko vidimo slovanske jezike, v oranžni imamo dva baltska jezika, v zeleni nekaj romanskih ter angleščino, v modri dva uralska jezika ter v vijolični germanske ter albanščino. Silhueta je dobra, saj imajo jeziki visoke silhuetne vrednosti.

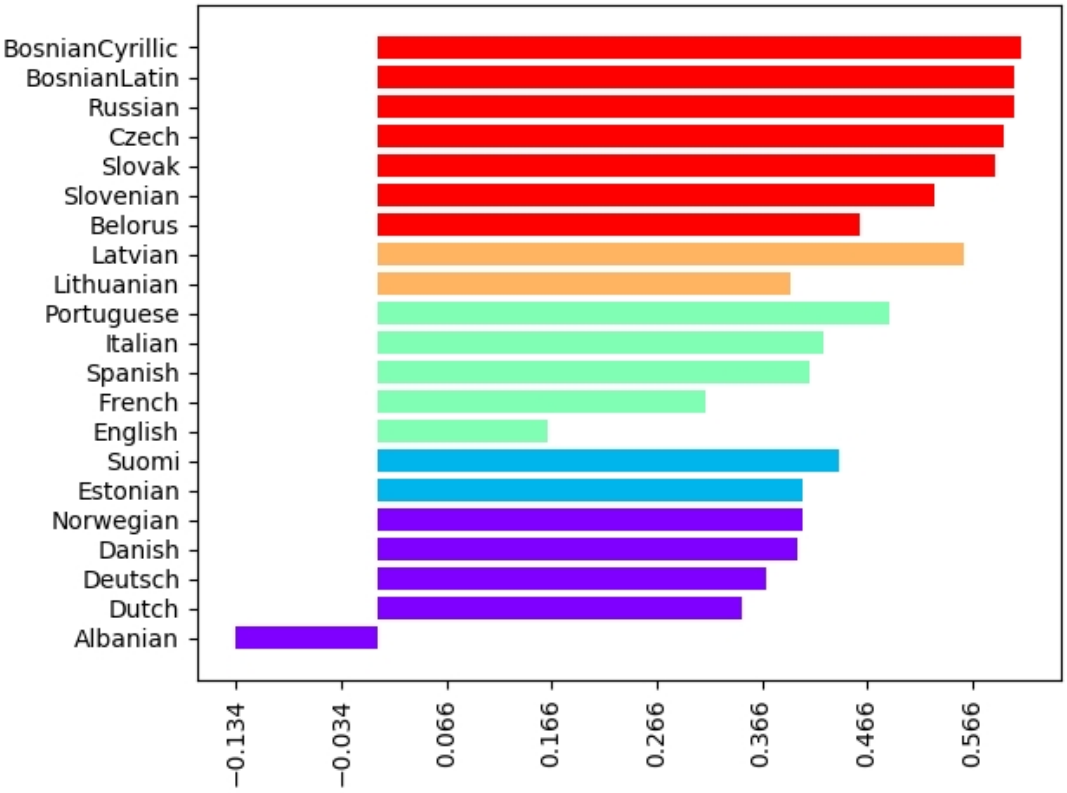

Рис. 2: Rezultat razvrščanja z najboljšo silhueto.

Najslabša silhueta. Slika 3 prikazuje najslabšo silhueto ter skupine jezikov. Že takoj lahko opazimo večje število negativnih vrednosti silhuet ter manj smiselne skupine kot prej.

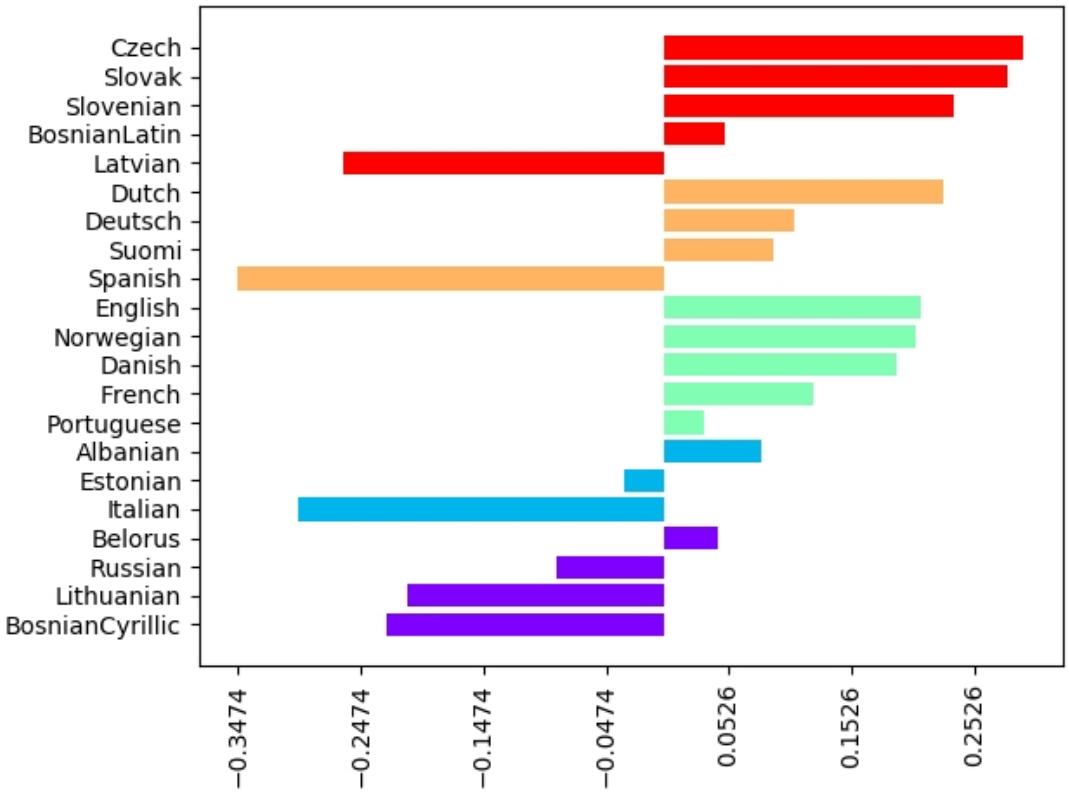

Рис. 3: Rezultat razvrščanja z najboljšo silhueto.

Napovedovanje jezika. Ugotavljanje jezika iz poljubnega besedila je bilo precej preprosto. Vse kar sem moral narediti je prebrati besedilo ter zanj narediti nov Point. Nato sem izračunal razdaljo le tega do vseh ostalih "pointov". Te razdalje sem delil z vsoto vseh razdalj, da sem dobil verjetnost, s katero lahko trdimo, da je besedilo v nekem jeziku.

```
    def compare_file(self, file_path: str, name="unknown") -> None:
        # preberemo datoteko in naredimo nov Point
        point = read_file(file_path, name)

        # izracunamo razdalje ter jih sortiramo
        comparisons = sorted([(point.similarity_with_point(p), p.lang)
                              for p in self.points], reverse=True)
        # izpis
        for comparison in comparisons[0:3]:
            print(comparison, comparison[0] /
                    sum(x for x, _ in comparisons) * 100)
```

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

Таблица 1: Besedila in prepoznani jeziki.

jezik / datoteka	verjetnost
dutch1.txt	
Dutch	14.09%
Deutsch	10.70%
Danish	7.96%
danish1.txt	
Danish	12.47%
Norwegian	11.07%
Dutch	8.88%
french1.txt	
French	13.79%
Spanish	8.75%
Portuguese	7.42%
spanish1.txt	
Spanish	12.54%
French	8.77%
Portuguese	8.72%
russian1.txt	
Russian	9.39%
Slovak	7.04%
Czech	6.89%
italian1.	
Italian	11.55%
Portuguese	6.47%
Spanish	6.35%
slo1.txt	
Slovenian	9.96%
BosnianLatin	8.83%
Czech	7.41%
english1.txt	
English	16.56%
French	7.14%
Danish	6.24%
albanian1.txt	
Albanian	16.10%
French	6.48%
Slovenian	5.4%
lithuanian1.txt	
Lithuanian	11.82%
Latvian	7.61%
Spanish	6.06%

Priloge

Uporabljena besedila.

dutch1.txt De dader sloeg toe in een zaal waarin vijfhonderd mensen passen, zo meldt de Franse krant Le Parisien. Volgens Benjamin, die in de zaal zat, riep de man het maar liefst zes keer. „Eerst lachten mensen, toen begonnen de eersten zich zorgen te maken, anderen vroegen hem zijn mond te houden. Toen hij opstond renden mensen in paniek naar de uitgang."

Een andere bezoeker laat weten: „Mensen klommen over de stoelen heen. Er waren vrouwen die in het gangpad op de grond vielen terwijl anderen over hen heen stapten. Ik zal die beelden nooit vergeten."

De man wist in eerste instantie te ontsnappen, maar werd later toch aangehouden. De politie vond niets verdachts in de bioscoop.

Het is niet de eerste keer dat er paniek uitbreekt in een bioscoop tijdens The Joker. Eerder deze maand begon een man in een bioscoop in het Britse Cheshire opeens te vloeken en andere bezoekers uit te schelden.

danish1.txt Danmarksrekorden for kæmpegræskar blev sat sidste år under Tivolis danmarksmesterskab i kæmpegræskar. 589,4 kilo lød rekorden på dengang. Men det må siges at blegne, når der sammenlignes med solen.

Halloween-solen stammer fra oktober 2014, men NASA opfordrer til, at man downloader billedet for at fejre torsdagens højtid. Billedet er taget af en rumsonde, som blev sendt op for at observere solens indflydelse på jorden.

NASA har blandet to billeder sammen taget ved forskellige bølgelængder for at skabe en sol med et halloween-lignende udseende.

french1.txt Sous la houlette du général Pinochet, dictateur exemplaire, le Chili a été, dès les années 1970, le laboratoire d'une formule qui fait aujourd'hui florès : l'alliance de l'autocratie et de l'ultralibéralisme. Passé sous régime démocratique, il serait à croire que le pays soit toujours sous cette double coupe, alors qu'une minorité de familles monopolise la terre et les biens, et que l'armée, protégée par des chars, tire contre une foule (dix-neuf morts, quatre cents blessés) qui proteste contre l'augmentation du prix du ticket du métro. Ce qui n'a pas empêché, vendredi 25 octobre, plus d'un million de personnes de manifester dans les rues de Santiago. Cette histoire, ce cynisme, cette cruauté qui ne désarment pas, le cinéaste Patricio Guzman, depuis son exil français, l'a durablement documenté depuis le coup d'Etat qui a causé la mort de Salvador Allende.

spanish1.txt En una providencia, el pleno del tribunal de garantías atiende por unanimidad la suspensión cautelar solicitada por el Ejecutivo y avalada por el Consejo de Estado contra el plan estratégico de acción exterior y relaciones con la UE 2019-2022, aprobado el pasado 25 de junio.

Lo ha hecho una vez que el Gobierno invocó el artículo 161.2 de la Constitución para que se produzca la suspensión inmediata después de que el Ejecutivo catalán contestara negativamente al requerimiento de incompetencia que se le remitió en agosto.

russian1.txt Президент России Владимир Путин отметил, что в каждый приезд любуется столицей Венгрии. С этих слов он начал переговоры в узком составе с венгерским премьер-министром Виктором Орбаном в среду, 30 октября.

Российский лидер также отметил, что Будапешт красив в любую погоду, даже несмотря на то что в эту среду в венгерской столице дождь и 7 градусов по Цельсию.

«Только сейчас мы (правящая партия Венгрии. — Ред.) потеряли его: было две столицы в Европе, где власть была не у левых, а у правых, — это были Мадрид и Будапешт. А теперь осталась только одна столица — это Мадрид», — цитирует ответ венгерского премьер-министра

italian1.txt La nuova imposta sul consumo "è dovuta dal produttore o fornitore nazionale o dal rappresentante fiscale del produttore o fornitore estero all'atto della cessione dei prodotti alle rivendite ovvero i tabaccai. La norma impone inoltre che tali prodotti siano venduti solo nei rivenditori autorizzati, escludendo la vendita online. "È vietata la vendita a distanza, anche transfrontaliera, ai consumatori che acquistano nel territorio dello Stato si legge nel testo. "L'Agenzia delle dogane e dei monopoli, fermi i poteri dell'autorità e della polizia giudiziaria ove il fatto costituisca reato, comunica ai fornitori di connettività alla rete Internet ovvero ai gestori di altre reti telematiche o di telecomunicazione o agli operatori che in relazione ad esse forniscono servizi telematici o di telecomunicazione, i siti web ai quali inibire l'accesso, attraverso le predette reti, offerenti prodotti di cui al comma 1".

slo1.txt Na odru ljubljanske Drame se drevi ustavlja prva slovenska godba na pihala, ki že več kot 15 let preigrava skoraj izključno ulični džez New Orleansa. Cerkljanski Kar Češ Brass Band, ki neworleanški džez interpretira na svojevrsten in svež način, so v gledališče povabili v sklopu cikla Drama Akustika. Tretji oktobrski konec tedna na Ravnah na Koroškem zdaj že tradicionalno poteka Festival slovenskega jazza, katerega spremljajoči dogodki se na Koroškem vrstijo že od septembra. Tridnevni festivalski vrhunec je v Kulturnem centru Ravne odprl Big Band RTV Slovenija z vsestranskim glasbenikom Boštjanom Gombačem. Murskosoboški policisti so 33-letniku z območja Murske Sobote zasegli posušeno konopljo, sadike in gotovino, za katero sumijo, da jo je dobil s preprodajo. Pri tem so v hiši osumljenega odkrili poseben prostor za gojenje konoplje pod umetno svetlobo. V tem prostoru so našli in zasegli tudi 20 sadik konoplje, visokih do 90 centimetrov, in dober kilogram posušene oz. delno posušene konoplje.

english1.txt Tuesday's motion - Prime Minister Boris Johnson's fourth attempt to secure a general election since he took office in July - will still need the approval of the House of Lords (the second chamber of the UK Parliament).

It is almost certain to pass and, if it does, it will be the country's third general election in less than five years - and the UK's first December election for nearly 100 years.

Mr Johnson said the move would help "get Brexit done allowing Britain to move forward with its withdrawal from the European Union (EU).

albanian1.txt "Ajo që Franca i propozoi Këshillit të Bashkimit Evropian ishte ideja e modernizimit të procesit të pranimit, i cili do të mundësonte që vendet kandidate që kanë përmbushur kriteret në fusha të caktuara të fillojnë të përdorin përfitimet e saj në vend që të presin fundin e tërë procesin", deklaroi Falcon në një intervistë për FoNet, cituar nga European Western Balkans

Ai shpjegoi se procesi aktual i pranimit të ngadaltë duhet të bëhet "më i shpejtë dhe më përparimtar".

Sipas tij, rezultatet e procesit aktual të pranimit nuk janë të mjaftueshme, qytetarët nuk shohin asgjë konkrete, dhe procesi shpesh është burokratik.

"Ideja është që, kur një vend të përfundojë negociatat, do të jetë në gjendje të ulet në të njëjtën tryezë me vendet anëtare dhe të marrë pjesë në hartimin e politikave, por pa të drejtë vote", shpjegoi Falcon në intervistë.

lithuanian1.txt Jis pakartojo, kad įvykusiame susitikime abi šalys patvirtino bendrą siekį mažinti įtampą dėl kariuomenės judėjimų ir apsikeitė nuomonėmis apie pastangas stiprinti pasitikėjimo ir saugumo priemones, užtikrinti panašių sprendimų priėmimo skaidrumą ateityje. „Būtina pabrėžti, kad tokia padėtis susidarė dėl pasitikėjimo ir informacijos stokos. Labai svarbu, kad partneriai amerikiečiai parodytų siekiantys konstruktyvaus dialogo pagal Vienos dokumentą dėl pasitikėjimo ir saugumo stiprinimo priemonių", – kalbėjo sekretorius. „Tačiau mūsų šalies pozicija dėl regiono militarizacijos nepriimtinumo lieka nepakitusi. Mes ir toliau laikysimės nuoseklaus užsienio politikos kurso, siekdami sumažinti karines ir politines įtampas Europoje ir palaikyti abipuse pagarba paremtą dialogą tarp šalių partnerių taikos ir saugumo užtikrinimo vardan", – patikino jis.

Skaitykite daugiau: https://www.lrytas.lt/pasaulis/ivykiai/2019/10/30/news/po-vasingtono-paaiskinimo-baltarusijos-sprendimas-keisti-plana-del-jav-kariu-12379855/