Ivan Antešić (63130003) **Naloga 2** 24. februar 2023.

Uvod. Pri drugi domači nalogi smo z analizo podobnosti indoevropske jezike razvrstili v skupine, glede na njihov izvor. Pri tem smo uporabili metodo k medoidov.

Izbrani jeziki. Iz podanih prevodov Splošne deklaracije človekovih pravic smo izbrali 31 naslednji jezikov:

- Afrikaans (afk)
- Albanian (aln)
- Breton (brt)
- Czech (czc)
- Danish (dns)
- English (eng)
- Estonian (est)
- Frisian (fri)
- French (frn)
- German (ger)
- Scottish Gaelic (gls)
- Greek (grk)
- Hungarian (hng)
- Italian (itn)
- Kurdish (kdb1)
- Latvian (lat)
- Latin (ltn)
- Macedonian (mkj)
- Norwegian (nrn)
- Portuguese (por)
- Polish (pql)
- Romanian (rum)

- Russian (rus)
- Slovakian (slo)
- Slovenian (slv)
- Spanish (spn)
- Serbian (src3)
- Swedish (swd)
- Turkish (trk)
- Ukranian (ukr)
- Sorbian (wee)

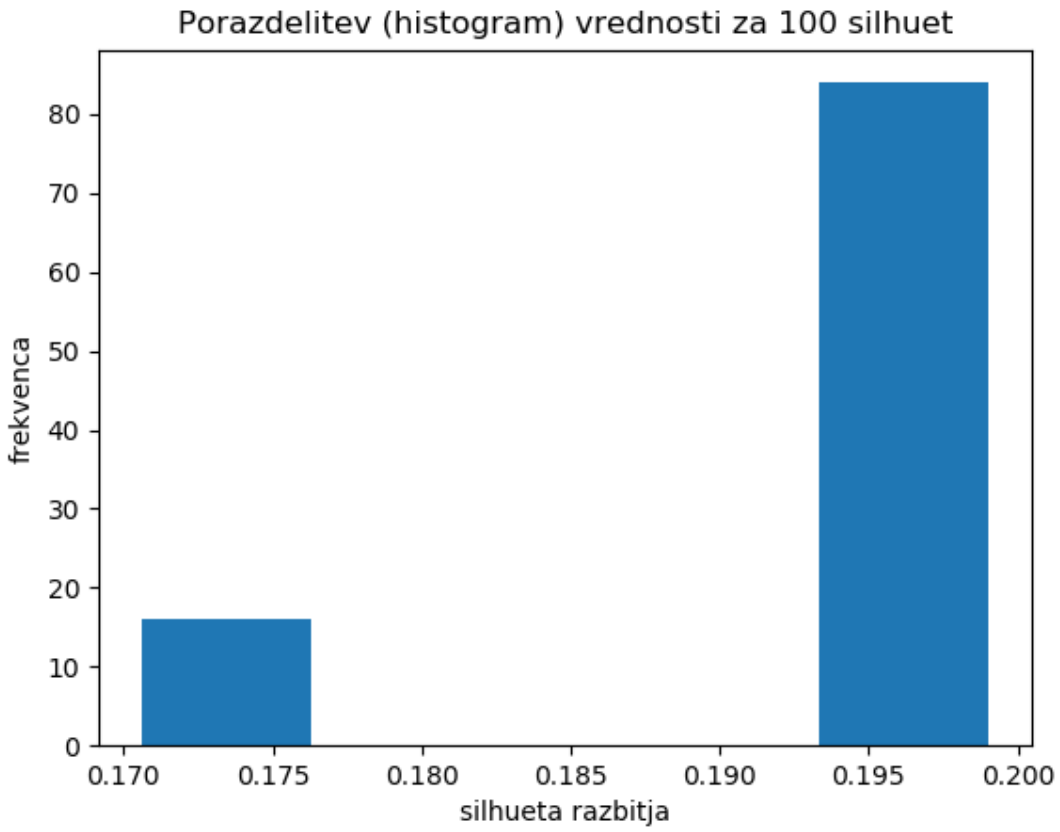
Datoteke, ki vsebujejo prevode smo najprej prebrali v niz in dekodirali posebne unicode znake s knjižnjico *unidecode*. Nato smo izločili vse znake, ki niso presledki ali velike in male črke ASCII abecede. Pri tem smo pazili, da nikjer ni skupaj več zaporednih presledkov. Vektor za posamezni primer (jezik) smo sestavili iz frekvenc pojavitev vseh trojic znakov, ki so vsebovane v obdelanem nizu. Nad takšno strukturo smo pognali PAM (angl. *partitioning around medoids*) algoritem za razvrščanje s k medoidi.

Rezultati razvrščanja Razvrščanje z naključnimi petimi medioidi smo ponovili stokrat. Za vsako razvrstitev smo izračunali silhueto in njihovo razporeditev prikazali na histogramu (slika 1).

Ne glede na začetne medoide na koncu dobimo vedno eno od dveh razvrstitev (prikazani v tabeli 1 in 2). Posledično dobimo dve skupini zelo podobnih silhuet, ki se med seboj razlikujejo le za nekaj decimalk. Razvrstitve dosežemo v povprečno petih iteracijah algoritma.

Pri najboljši razvrstitvi opazimo smiselne skupine slovanskih, germanskih in romanskih jezikov. Zanimivi so primeri angleščina, estonščina, latvijščina in grščina. Izgleda, da niso dovolj karakteristični za ostale skupine in jih je za izbrane medoide, algoritem razvrstil na obrobje romanskih skupin. Skupaj so razvrščeni tudi zelo specifični vzhodni jeziki: turščina, albanščina, madžarščina in kurdščina. Škotsko keltski jezik je edini predstavnik svoje skupine. Čeprav je bretonščina tudi keltski jezik, le ta spada v drugo vejo izvora in je zato morda v skupini germanskih jezikov, ki so jo obkrožali skozi zgodovino (npr. anglo-saška plemena).

Napovedovanje jezika Tekstovno datoteko v poljubnem jeziku smo obdelali na isti način kot prevode Splošne deklaracije človekovih pravic in kreirali isto strukturo vektorjev. Nato smo izračunali kosinusno podobnost med neznanim jezikom in vsemi prevodi ter rezultate razvrstili po podobnosti padajoče. Za verjetnost pravilne napovedi jezika smo uporabili kar kosinusno podobnost pomnoženo s 100. Če so vektorji v pozitivnem prostoru (in v našem primeru so vedno) je podobnost omejena med 0 in 1: vrednost 1 če se vektorja ujemata (napoved je 100% pravilna) in vrednost 0 če sta popolnoma različna (napoved je 0% pravilna).



Slika 1: Histogram silhuet razvrščanja (razbitja) za sto ponovitev z naključnimi 5 medoidi.

skupina 0	skupina 1	skupina 2	skupina 3	skupina 4
czc 0.15	*gls 0.0	afk 0.48	eng 0.67	brt 0.61
mkj 0.49		aln 0.74	frn 0.39	dns 0.24
pql 0.49		*fri 0.0	grk 0.7	est 0.76
rus 0.49		ger 0.48	itn 0.5	hng 0.73
*slo 0.0		kdb1 0.65	lat 0.74	*nrn 0.0
slv 0.47			ltn 0.64	swd 0.4
src3 0.41			por 0.33	trk 0.69
ukr 0.49			rum 0.54	
wee 0.5			*spn 0.0	

Tabela 1: Rezlutat razvrščanja za najslabšo silhueto 0,171. Z zvezdico so označeni medoidi skupine. Številka poleg kratice jezike predstavlja razdaljo do medoida.

skupina 0	skupina 1	skupina 2	skupina 3	skupina 4
afk 0.57	aln 0.71	eng 0.67	*gls 0.0	czc 0.15
brt 0.67	hng 0.75	est 0.78		mkj 0.49
*dns 0.0	*kdb1 0.0	frn 0.39		pql 0.49
fri 0.56	trk 0.56	grk 0.7		rus 0.49
ger 0.5		itn 0.5		*slo 0.0
nrn 0.24		lat 0.74		slv 0.47
swd 0.36		ltn 0.64		src3 0.41
		por 0.33		ukr 0.49
		rum 0.54		wee 0.5
		*spn 0.0		

Tabela 2: Rezlutat razvrščanja za najboljšo silhueto 0.199. Z zvezdico so označeni medoidi skupine. Številka poleg kratice jezike predstavlja razdaljo do medoida.

Napisali smo tudi funkcijo, ki najprej izbere najbližji medoid in nato enako kot prej izračuna podobnosti med vektorji, vendar tokrat le znotrja skupine izbranega medoida. Tako nam ni potrebno izračunati vseh razdalj med primeri in neznanim jezikom, če bi bilo primerov veliko. Vendar je to optimizacija le, ko že imamo na voljo razvrstitev v skupine, zato smo spodnje rezultate (tabela 3) raje izračunali z preprostejšo prvo metodo.

odlomek besedila 0	3 najboljše napovedi in njihove verjtnosti (%)		
czc1.txt	czc 42.145	slo 37.824	slv 29.142
ger1.txt	ger 39.641	fri 32.62	eng 30.248
itn1.txt	itn 62.552	rum 42.405	por 41.148
itn2.txt	itn 56.477	spn 38.034	por 34.286
hng1.txt	hng 44.564	grk 18.476	kdb1 16.837
por1.txt	por 58.675	spn 48.138	frn 32.769
slv1.txt	slv 51.601	mkj 49.005	src3 48.423
slv2.txt	slv 43.651	src3 39.114	mkj 36.744
slv3.txt	slv 34.253	src3 31.455	mkj 30.075
src31.txt	src3 38.755	mkj 33.141	slv 29.003

Tabela 3: Rezultati napovedovanja jezika za 10 različnih krajših odlomkov. Odlomki se nahajajo v dodatku.

Priloge

Izbrani odlomki za testiranje napovedovanja jezika:

czc1.txt Hra Age of Empires II: The Age of Kings byla vydána v roce 1999 a přinesla některé nové herní prvky, jako např. brány v opevnění a umísťování jednotek do budov. Hra je zasazena do prostředí středověku od jeho temných počátků až do začátků renesance, včetně španělských výbojů v Mexiku. Hráč si může vybrat z třinácti (a v datadisku The Conquerors ještě z dalších

pěti) civilizací.

ger1.txt Age of Empires II: HD Edition - The Forgotten ist ein Echtzeit-Strategie-Add-on zu Age of Empires II: HD Edition und wurde von Hidden Path Entertainment, Forgotten Empires und SkyBox Labs entwickelt und von den Microsoft Game Studios für Windows-basierende Computer veröffentlicht wurde. Das Spiel erschien am 7. November 2013.

itn1.txt Resta da risolvere, però, il problema di quale rapporto ci sia tra Perun e Thor, il dio germanico della folgore (nella Rus' erano presenti i Variaghi), e Perkūnas, divinità baltica raffigurata proprio come Perun. In definitiva, non è ancora chiaro se Perun fosse una divinità propria del mondo slavo, o presa "in prestitođal mondo germanico o da quello baltico (è nelle lingue baltiche che il nome di Perkūnas è ricollegabile al "tuono": ancora oggi in lettone perkůns significa 'tuono', mentre parole analoghe in ucraino o in polacco potrebbero essere entrate in tempi posteriori). Tra l'altro, non è escluso che il nome del dio Thor si trovasse scritto su qualche fonte in caratteri runici, e gli autori delle cronache slave possano aver confuso nella traslitterazione una P con una P.

itn2.txt Il 21 agosto 2017, dopo più di 10 anni dall'uscita dell'ultimo aggiornamento ufficiale (relativo ad Age of Empires III - The Asian Dynasties), in occasione del Gamescom di Colonia, Microsoft ha annunciato[5] che è in lavorazione il quarto capitolo della saga, Age of Empires IV, pubblicando anche un video che però nulla svela sulla futura ambientazione del gioco, né sulla sua possibile data di uscita. Alla fine del video compare il logo "Microsoft Windows 10", suggerendo che il gioco possa uscire solo per questa versione del sistema operativo.

hng.txt A játékhoz számos rajongói mod és kiegészítés (pl. egyéni küldetések és térképek) tölthető le az internetről, köszönhetően az olyan, mindegyik részben megtalálható eszközöknek, mint beépített térképszerkesztők és AI-szkriptnyelvek, amelyekkel ilyenek is készíthetőek.

por1.txt Age of Empires é uma série de jogos eletrônicos para computador e consoles portáteis desenvolvida pela Ensemble Studios e publicada pela Microsoft. O primeiro título da série foi Age of Empires, lançado em 1997. Depois dele, outros seis títulos da franquia principal e quatro títulos derivados foram lançados. A maior parte da série é formada pelo gênero de estratégia em tempo real, sendo que seus modos de jogo se resumem a dois estilos principais: mapa aleatório e campanha. Além do gênero, os jogos da série são caracterizados por marcarem eventos históricos e muitas outras surpresas.

slv1.txt Zgoraj omenjeno prepričanje je možno razložiti na ta način: Na začetku staroslovanske literature v Bolgarskem Kraljestvu sta bili kroniki Hamartolos in Malala prevedeni v slovanščino. Te kronike opisujejo migracije narodov iz področja Senaar po Poplavi. Po tem mnenju so Evropejci potomci Jafeta, ki je pripotoval iz Senaarja skozi Malo Azijo na Balkan; tam naj bi se razdelili v različne narode in se razšli. Slovanski bralec teh kronik bi tako lahko verjel, da je

bila začetna točka migracij Slovanov zato Balkan in območje spodnje Donave. Ker zgodovinske avtoritete postavljajo na to mesto v regiji v tistem času Ilire, je bilo potrebno, da se naredi tudi to pleme slovansko. V kasnejših bitkah Slovanov za ohranjanje njihovih jezikov v liturgiji je bila ta možnost zelo privlačna, saj bi se lahko potem sklicevali na slovansko terjatev avtoritete Svetega Jurija ali celo Svetega Pavla. (Mnenja, ki so sicer trenutno popularna, vendar ne odražajo dejstev, so velikokrat povzeta v zgodovinskih zapisih.

slv2.txt Implementirajte postopek za razvrščanje v skupine na podlagi medoidov (k-medoids clustering), ki razdalje med besedili meri s kosinusno razdaljo s primerjavo frekvenc trojk sosednjih črk in to nujno brez uporabe polnih matrik. Pri računanju razdalj morate torej upoštevati le trojke, ki jih dani besedili dejansko vsebujeta; program bi moral delovati podobno hitro, četudi bi namesto trojk primerjali enajsterke. [50Razviti postopek poženite s 100 naključno izbranimi inicializacijami oziroma začetnim izborom k=5 medoidov. Postopek razvrščanja z medoidi je lahko namreč precej odvisen od začnega izbora medoidov. Na ta način boste dobili 100 različnih razvrstitev oziroma rezultatov postopka. Vsakega od 100 razvrstitev ocenite z metodo silhuete, ki jo razvijete sami. Izrišite tudi porazdelitev (histogram) vrednosti silhuet. [30

slv3.txt Age of Empires (slovensko »Doba imperijev«) pogosto s kratico AoE, je priljubljena serija strateških videoiger z zgodovinskim ozadjem. Natančneje sodijo med realno-časovne strategije. Razvija jih je podjetje Ensemble Studios, izdaja pa Microsoft. Prvi del je izšel leta 1997, od takrat pa je izšlo še šest delov in več izpeljank.

src31.txt До средине 16. века, читава модерна Србија била је у склопу Османског царства, све док га није прекинула Хабзбуршка монархија, која је почела да се шири према Централној Србији од краја 17. века, а одржавала је упориште у модерној Војводини. Почетком 19. века, Српска револуција успоставила је националну државу као прву уставну монархију у региону, која је касније проширила своју територију