Yannick Kuhar (63160187) **Naloga 2** November 1, 2018

1 Izbrani jeziki.

L = jezik je v Latinici
C = jezik je v Cirilici
D = drugo

Slovanski jeziki:	Slv(L)	Slo(L)	Rus(C)	Blg(C)	Czc(L)	
Germanski jeziki:	Eng(L)	Ger(L)	Swd(L)	Fin(L)	Ice(L)	
Romanski jeziki:	Fra(L)	Esp(L)	Por(L)	Ita(L)		
Ostalo:	Swa(L)	Gre(D)	Trk(L)	Kkn(D)	Chn(D)	Jpn(D)

1.1 Predobdelava datotek.

Sprva smo sestavili slovar, oblike {'država': besedilo}, katerega smo na predavanjih imenovali corpus. Nato smo iz besedila sestavili vse možne trojice znakov besedila in izračunali njihovo frekvenco. Tako smo dobili točke v večdimenzionalnem prostoru, oblike {'država': tocka(vektor frekvenc)}.

2 Rezultati razvrščanja.

V tem razdelku si bomo pomagali s tremi priloženimi datotekami (histogram.txt, minSihueta.txt in maxSihueta.txt), v katerih se nahajajo rezultati. Podobnosti sihuet se nahajajo na intervalu [0.21, 0.4], kar kaže, da imajo jeiki skupnega prednika. Na podoben, čeprav manj učinkovit način, se naredijo skupine.

3 Napovedovanje jezika.

Uporabimo podatke, ki jih že imamo, torej slovar točk. Besedilo predelamo na isti način kot vhodne podatke in dobimo točko, nato iteriramo čez slovar točk in v vsaki iteraciji izračunamo kosinusno razdaljo med točkama. Jezik določimo na podlagi maksimalne kosinusne razdalje.

3.1 Tabela.

Danes je lep dan, na nebu ni bilo nobenega oblaka.	SLOVAK	Napačno
Today is a good day, there are no clouds in the sky.	SLV	Napačno
Heute ist ein guter Tag es gibt keine Wolken an dem Himmel.	GER	Pravilno
Dnes je dobrý den, na obloze nejsou žádné mraky.	ITN	Napačno
Aujourd'hui est un bon jour, il n'y a pas de nuages dans le ciel.	FIN	Napačno

Glede na rezultate metoda ni natančna.

4 Bonus naloga: Članki.

V tem razdelku bomo primerjali rezultate med naborom prevodov Deklaracije o človekovih pravicah(rezultati v histogram.txt) in nabor 20-ih člankov iz interneta(histogram_clankov.txt). Glavna razlika med histogramoma je da histogram prevodov je porazdeljen na intervalu [0.21, 0.4] medtem, ko je histogram člankov porazdeljen na intervalu [-0.85, 0.35]. Sklepan, da je razlika med intervaloma zaradi drastične razlike v vsebini dokumentov.