Podobnost jezikov

Karmen Gostiša (63130057)

31. oktober 2017

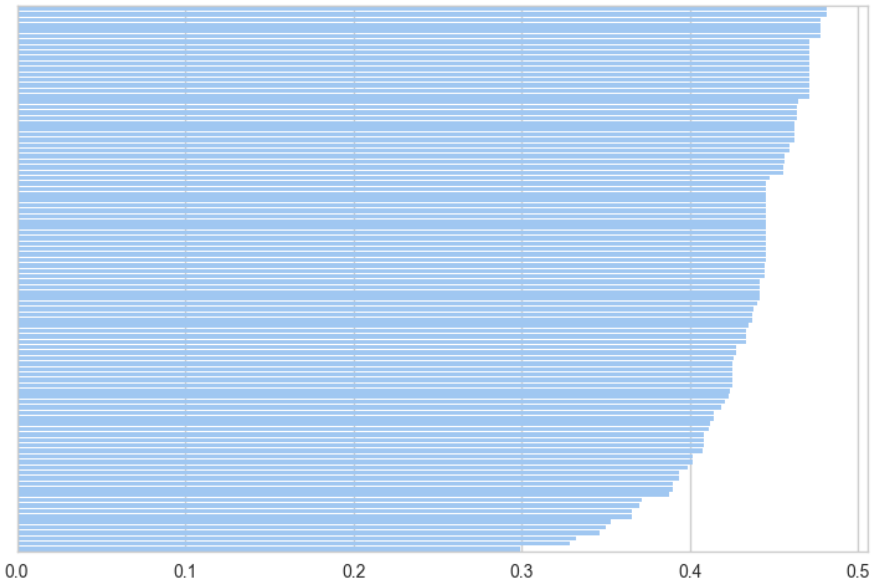
1 Izbrani jeziki

Izbrala sem naslednje jezike: bosanski, hrvaški, češki, danski, angleški, francoski, nemški, galicijski, grški (grška pisava), italijanski, nizozemski, poljski, portugalski, romunski, ruski (cirilica), škotski, slovaški, slovenski, španski, švedski, ukrajinski (cirilica).

Iz vseh besedil sem odstranila pike, vejice, vse vrste oklepajev, pomišljaje, vezaje, dvopičja, podpičja in števila. To sem storila z orodjem Notepad++ z ukazom nadomesti vse (angl. *replace all*) znotraj menija poišči v datotekah (angl. *find in files*), ki to akcijo omogoča za več datotek znotraj neke mape naenkrat. V kodi sem izvedla prečrkovanje s pomočjo knjižnice `unidecode` in besedila razdelila na besede, ki sem jih nato razbila na 3-terke znakov.

2 Rezultati razvrščanja

Slika 1 prikazuje vrednosti silhuet za 100 naključno določenih začetnih izborov $k = 5$ medoidov.



Slika 1: Vrednosti silhuet za 100 naključnih začetnih izborov medoidov.

Pri najboljši silhueti sem dobila naslednje skupine jezikov:

1. danski, nizozemski, švedski
2. francoski, galicijski, portugalski, španski, italijanski, romanski
3. angleški, grški, škotski
4. nemški
5. bosanski, hrvaški, češki, poljski, ruski, slovaški, slovenski, ukrajinski

Razvrščanje pri najslabši silhueti pa je dalo naslednje skupine jezikov:

1. nemški, ruski
2. švedski, ukrajinski
3. bosanski, hrvaški, češki, angleški, francoski, galicijski, italijanski, poljski, romanski, škotski, slovaški, španski
4. danski, nizozemski
5. grški, portugalski, slovenski

Rezultati se zdijo smiselni. Pri najboljši silhueti se opazi, da se v posamezni skupini nahajajo jeziki iz iste jezikovne družine. Nekoliko čudno je sicer, da se nemški jezik nahaja sam v svoji skupini in ne skupaj z ostalimi zahodnogermanskimi jeziki v 3. skupini, kjer se je začudoma znašel tudi grški jezik. Morda je to posledica transliteracije ali pretvorbe v male znake. Blok slovanskih jezikov je lepo viden v 5. skupini. Pri najslabši silhueti pa vidimo, da niti ena skupina ne vsebuje samo jezikov iz iste jezikovne družine.

3 Napovedovanje jezika

Odlomke iz besedil in prve tri napovedane jezike po verjetnosti prikazujeta tabeli 1 in 2.

Odlomek iz besedila	**Napovedani jeziki po verjetnosti**
Matematika od grčkog mathema znanost je egzaktna točna nedvojbena znanost koja izučava aksiomatski definirane apstraktne strukture koristeći matematičku logiku	hrvaški, bosanski, slovenski
Der er flere opfattelser af hvornår matematikken opstod Længe før matematik udviklede sig til sit eget kundskabsområde har mennesker været optaget	danski, nizozemski, švedski
The area of study known as the history of mathematics is primarily an investigation into the origin of discoveries in mathematics	angleški, škotski, grški

Tabela 1: Odlomki iz besedil in prvi trije napovedani jeziki po verjetnosti.

Odlomek iz besedila	Napovedani jeziki po verjetnosti
L'histoire des mathématiques s'étend sur plusieurs millénaires et dans de nombreuses régions du globe allant de la Chine à l'Amérique centrale	francoski, angleški, škotski
Die Geschichte der Mathematik reicht zurück bis ins Altertum und den Anfängen des Zählens in der Jungsteinzeit	nemški, danski, škotski
La storia della matematica ha origine con le scoperte matematiche e prosegue attraverso l'evoluzione nel corso dei secoli dei propri metodi e delle notazioni matematiche il cui uso si sussegue nel tempo	italijanski, francoski, romunski
Данная статья представляет собой обзор основных событий и тенденций в истории математики с древнейших времён до наших дней	ruski, ukrajinski, slovaški
Zgodovína matemátike je področje ki se prvenstveno ukvarja z izvorom novih odkritij v matematiki in v manjši meri s standardnimi matematičnimi metodami in zapisi v preteklosti	slovenski, bosanski, hrvaški
El sueco es una lengua germánica del norte de Europa hablada por entre y millones de personas	španski, portugalski, galicijski
Matematikens historia är historien om hur människan genom tiderna och i olika regioner utvecklat olika matematiska teorier	švedski, danski, nizozemski

Tabela 2: Odlomki iz besedil in prvi trije napovedani jeziki po verjetnosti.

Iz rezultatov tabel 1 in 2 lahko opazimo, da je program v vseh primerih napovedal jezik besedila na prvo mesto. Postopek napovedovanja lahko opišem z naslednjimi koraki:

Korak 1 Na Wikipediji sem poiskala članke o zgodovini matematike v različnih jezikih in vsakega posebej shranila v svojo datoteko. Vse datoteke z besedili sem ustrezno predobdelala na enak način kot je zapisano v poglavju 1.

Korak 2 Iz besedil sem zgradila vektorje - slovarje, ki hranijo 3-terke znakov posameznih besed in njihovo frekvenco.

Korak 3 Med posameznim besedilom in izbranimi jeziki sem izračunala kosinusno podobnost ter vrednosti razvrstila padajoče. Tako sem na prvem mestu dobila jezik, za katerega napovedujem, da je najbolj verjeten, da predstavlja besedilo.

4 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelala sama.