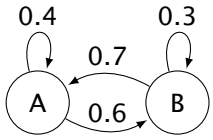
Ali lahko računalnik spiše esej?

Vid Kocijan

→ **Za domačo nalogo moramo spisati esej o Prešernovem Krstu pri Savici. Ker nam gresta matematika in programiranje bolje od slovenske književnosti, nas zanima, ali lahko spišemo program, ki bi namesto nas spisal esej o Krstu pri Savici. V tem članku si bomo ogledali markovske verige in njihovo uporabo za generiranje besedil.**

Markovska veriga

Markovska veriga je matematični model za opisovanje sistemov, kjer je verjetnost naslednjega dogodka v zaporedju odvisna zgolj od zadnjega dogodka v njem. Predstavimo ga kot množico stanj S in povezav med njimi P. Uteži na izhodnih povezavah iz vozlišča $s_i \in S$ predstavljajo verjetnosti prehoda iz stanja s_i v druga stanja. Vsota vseh uteži na izhodnih povezav iz nekega vozlišča se morajo tako sešteti v 1.



SLIKA 1.
Preprosta markovska veriga z dvema stanjema.

Preprost primer markovske verige lahko najdemo na sliki 1. Veriga je sestavljena iz dveh stanj $S = \{A, B\}$ in štirih povezav $p_{A,B} = 0{,}6$, $p_{B,A} = 0{,}7$, $p_{A,A} = 0{,}4$ in $p_{B,B} = 0{,}3$. Naš cilj je ustvariti markovsko verigo, kjer bodo vozlišča besede, povezave pa verjetnosti, da v našem eseju ena beseda sledi drugi. Nato lahko spišemo program, ki bi se po omenjeni markovski verigi sprehajal in pri tem ustvarjal esej. V vsakem stanju naključno izberemo naslednje stanje, sledeč verjetnostni porazdelitvi, ki jo podajajo uteži. Vzemimo markovsko verigo s slike 1. Če smo zadnjo izpisali črko *A*, pravimo, da smo v stanju *A*, bomo v naslednjem koraku z verjetnostjo 0,4 ponovno izpisali črko *A*, z verjetnostjo 0,6 pa bomo izpisali črko *B*. Podobno, če smo zadnjo izpisali črko *B*, pravimo, da smo v stanju *B*, bomo v naslednjem koraku z verjetnostjo 0,3 ponovno izpisali črko *B*, z verjetnostjo 0,7 pa naslednjo izpišemo črko *A*.

Če simuliramo gibanje po omenjeni markovski verigi, lahko dobimo besedilo, podobno sledečemu:
BABABBBAABABABAABAAABABABBAABBAABBAAB
ABBABABBAABAABABABABABBAABBAABBAABA.
Kot vidimo, je v generiranem besedilu malo več črk *A*, kljub temu pa se črki *A* in *B* večinoma izmenjujeta.

Pisanje besedil z markovskimi verigami

Sestavljanje take verige na roko je seveda zahtevno, verjetno dosti bolj, kot pisanje samega eseja. Na srečo lahko na spletnem portalu `dijaski.net/` najdemo šest esejev o Krstu pri Savici. Uteži v markovski verigi lahko določimo s statistično analizo omenjenih esejev. Spišemo program, ki prebere omenjene eseje in za vsako unikatno besedo izračuna, katere besede so ji sledile in kako pogosto. Če je besedi rad v 23 % sledila beseda imam, bo od stanja S_{rad} v stanje S_{imam} vodila povezava s težo 0,23.

Oglejmo si dejansko izsek verige, ustvarjene na ta način. V tabeli 1 lahko najdemo izsek iz ustvarjene verige, seznam besed, ki najpogosteje sledijo besedi *je*.

beseda	verjetnost
bil	0,0611
to	0,0556
v	0,0389
bila	0,0333
tudi	0,0333
črtomir	0,0222

TABELA 1.
Šest besed, ki najverjetneje sledijo besedi *je* in pripadajoče verjetnosti.

Da poenostavimo zahtevnost besedil, iz njih izločimo vsa ne-končna ločila, vse velike tiskane črke pa spremenimo v male. Pike, vejice in klicaje obravnavamo kot samostojna stanja, poleg tega pa dodamo še dve stanji: `<start>` in `<konec>`. Naš esej se začne s stanjem `<start>`, ko dosežemo stanje `<konec>` pa z esejem zaključimo. Ti dve »besedi« dodamo na začetek in konec vsakega besedila. Omenjeni besedi nam bosta olajšali generiranje besedila. Naš sprehod po markovski verigi bomo začeli v stanju `<start>`, ko dosežemo stanje `<konec>`, pa generiranje zaključimo. Omenjeni besedi seveda izbrišemo iz besedila, preden ga izpišemo bralcu.

Program, spisan po zgornjem kuharskem receptu, nam bo med drugim spisal sledečo umetnino:

krst pa ga je zgrajen iz končnega dialoga jasno razviden značaj in poln besa vodi svoje patriotsko mišljenje in krvavega boja pa verjetno razočara marsikaterega bralca. te zvrsti ki jih prešeren je čutil do izraza misel kot sem jaz saj izvemo da je čopova smrt matije čopa...

Za tak esej verjetno ne bi dobili pozitivne ocene. Človeški jezik je prezapleten, da bi ga lahko opisali s preprosto markovsko verigo. Naslednja beseda v eseju je namreč bolj odvisna od širšega konteksta kot zgolj od zadnje besede. Naš model zato nadgradimo in uporabimo markovske verige drugega reda. Namesto, da bi naslednjo besedo izbirali glede na to, katera beseda je bila zadnja, jo izbiramo glede na zadnji dve besedi v eseju. Če imamo skupek besed »Krst pri Savici je« bomo opazovali, kako pogosto se posamezna beseda znajde za sosledjem besed »Savici je«.

Modelu jezika oz. jezikovnemu modelu, ki ga sestavimo iz markovskih verig višjih redov, pravimo tudi n-gram jezikovni model. Definirajmo n-gram jezikovni model bolj formalno. Naj bo A zaporedje $n-1$ besed, x ena beseda, in Ax zaporedje iz n besed, ki ga dobimo, če zlepimo A in x. Naj bo $N(A)$ število pojavitev zaporedja A v analiziranih besedilih. Verjetnost, da beseda x sledi zaporedju A, $\mathbb{P}(x|A)$, definiramo kot

$$\mathbb{P}(x|A) = \frac{N(Ax)}{N(A)}.$$

Bralcu v premislek prepuščamo, ali je sledeča definicija smiselna, torej, ali se vsota verjetnosti vseh besed, ki lahko sledijo zaporedju A, vedno seštejejo v 1.

Markovska veriga prvega reda, ki smo jo ustvarili, tako ustreza ravno 2-gram, oz. bigram jezikovnemu modelu. Da pridobimo koherentnejše besedilo, postopek ponovimo s trigram jezikovnim modelom oz. z markovsko verigo drugega reda. Ponovimo analizo vseh esejev in generiramo sledeče besedilo:

krst pri savici je zgrajen iz treh delov iz posvetilnega soneta matiji čopu nato pa nekako klone in se noče podjarmiti. v uvodu je zgradba skoraj v celoti epska saj o dogodkih poroča jedrnato in poudarja le prvine. v celoti epska saj o dogodkih poroča jedrnato in poudarja le prvine...

To se zdi morda bolj podobno smiselnemu eseju, vendar bi za tak esej še vedno dobili negativno. Hkrati pa nas zmoti, da v generiranem besedilu začenjamo opažati sosledje besed iz esejev, na podlagi katerih smo zgradili markovsko verigo. To ni presenetljivo, naš nabor esejev je relativno majhen, unikatnih trojic besed, ki se v njem pojavijo, pa je veliko. Naš program bo tako pogosto generiral sosledje besed, ki so se pojavile v esejih. Tega si ne želimo, saj se plagiatorstvu želimo ogniti.

Kaj pa, če povečamo nabor besedil, ki jih analiziramo? Zavržemo eseje o Krstu pri Savici in zberemo širši nabor literarnih del. Za potrebe tega članka je bila uporabljena večina proznih del s spletne strani `lit.ijs.si/leposl.html`. Trubarja, Janeza Svetokriškega in Brižinske spomenike odstranimo, saj se jezik v teh delih zelo razlikuje od današnje slovenščine. Nato ponovimo vajo, tokrat z naborom več kot

100 strnjenih besedil. Z markovsko verigo drugega reda lahko dobimo sledeče besedilo:

že peti dan pijan prišel domov mu je samostanski vojak. svetin je začel praviti ali si pozabil kaj sem mord sam govoril ž njo pa jo strahuje da si človek oddahnil naslonil se je ozrl šepavec proti oknu skomignila z rameni. bila je mokra. ko je drugo...

Z 4-gram jezikovnim modelom oz. z markovsko verigo tretjega reda pa bolj koherentno:

že peti dan so bili zdoma. no z gradišča res ni daleč do belega dvora nemara se še nocoj vrneta na družinski pomenek vsekakor pa jutri. domačini so goste pospremili do ceste kjer so ga obvezali. nic nevarnega samo praska! in petdeset kron mu je dal oče ker mu iz gozda ni mogel ničesar prinesti...

Oglejmo si izsek iz dobljenega 4-gram jezikovnega modela, ki ga lahko najdemo v tabeli 2. Kontekst treh besed je dovolj, da je ustvarjeno besedilo relativno smiselno. Kljub temu pa je mnogo premalo, da bi program lahko pisal vsebinsko konsistentno besedilo, ki ima rdečo nit. Ali slepo povečevanje reda verige zares reši ta problem?

beseda	verjetnost
da	0,0705
zdelo	0,0590
bilo	0,0210
v	0,0210
tresel	0,0171
zdela	0,0114

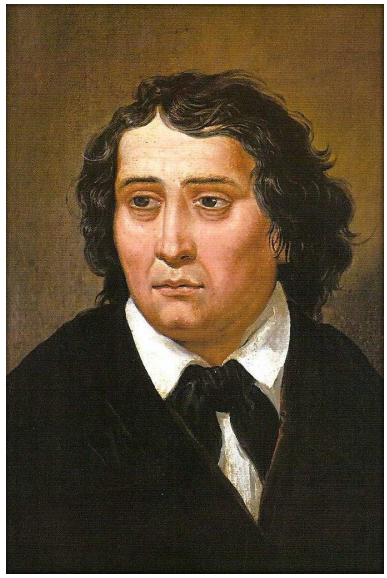
TABELA 2.
Šest besed, ki najverjetneje sledijo sosledju besed *se mu je* in pripadajoče verjetnosti.

Pri 5-gram jezikovnem modelu se ponovno zatakne, saj generirana besedila ponovno postanejo prepodobna besedilom, ki smo jih statistično analizirali.

Kje je meja?

Ker število različnih n-teric eksponentno narašča glede na n, s povečevanjem reda markovske verige eksponentno narašča tudi potreba po količini besedila za analizo. S še večjim naborom podatkov lahko ustvarimo markovsko verigo četrtega reda, potem pa se znova zatakne. 5-gram jezikovni model velja za najkompleksnejši smiselni model, ki ga je moč zgraditi s tako metodo. Tudi Google, dandanes verjetno eden od največjih zbirateljev podatkov, se ni trudil zbirati n-teric besed prek dolžine $n = 5$[1].

Za boljše modeliranje jezika dandanes uporabljamo močnejše metode, ki temeljijo na nevronskih mrežah in globokem učenju, kar pa presega obseg enega članka v Preseku. Če bralca zanima, kako se obnašajo trenutno najnaprednejši generatorji (angleškega) jezika, se z njimi lahko pozabava na naslednji spletni strani: `transformer.huggingface.co/`. Koda in gradiva, uporabljena pri pisanju tega članka, so dostopna na `github.com/vid-koci/presek\_generiranje\_besedila`. Ne glede na izjemne napredke pri avtomatskem generiranju besedila, bralcem priporočamo, da svoje eseje še naprej pišejo sami.



SLIKA 2.
France Prešeren (vir: Wikipedia)

× × ×

[1] Zbirka dostopna na `ai.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html`