

Hľadanie podobností v ľudových piesňach

Už na začiatku vyberania bakalárskej práce som bol rozhodnutý pracovať na niečom, čo bude súvisieť s hudbou, teda nejakým spôsobom uľahčí alebo spríjemní ľuďom užívanie si ich hobby – hudby. Pri hľadaní témy som sa od IoT gadgetov cez analýzu zvukových stôp až po vyhľadávanie a analýzu podobností v melódiách dostal ku textom ľudových piesní.

V rámci štúdia som začal pracovať na zbieraní, zobrazovaní a pomalom systematickom triedení textov „ľudoviek“, aby som uľahčil život sebe a ostatným, čo majú chuť z času načas spievať, respektíve nejakým spôsobom zjednotil texty ľudových piesní, ktorých je na internete mnoho, ale sú na rôznych miestach, chýba im konzistentnosť, sú niekedy nekorektne zapísané a opakujú sa. Aj vďaka tomuto vo mne skrsol nápad na nástroj alebo aplikáciu určenú na usporadúvanie, mazanie duplicít, jednoduché vyhľadávanie a inú analýzu týchto pesničiek.

Prejdime k podstate mojej práce. Ide o systém, ktorý na základe konkrétnej piesne alebo útržku textu bude vykonávať analýzu výskytu podobných textov v databáze. Predstaviť si to môžeme tak, že skrz nejaké užívateľské prostredie vložíme text, názov a voliteľne aj iné parametre piesne (napr. región). Po spustení algoritmu uvidíme piesne, ktoré majú percento podobnosti nad istou hranicou. Podobnosť v našom prípade je (zjednodušene povedané) vyhodnotená na základe množstva rovnakých/ podobných slov. V tejto časti sa budem musieť oboznámiť so základmi ukladania dát, ktoré sú určené na to, aby nad nimi bol vykonávaný data mining a tiež aj s data mining algoritmi.

Náročnejšia časť práce bude pozostávať zo zjednocovania podobných slov za pomoci rôznych metód a nástrojov ako napríklad Tvaroslovník, tvorený na našej univerzite, alebo nami generovaný Tvaroslovník určený špeciálne na ľudovú reč. Ukážka na príklade: chceme dosiahnuť, aby slová „dzivče“ a „dzifče“ boli vyhodnotené ako zhodné. Táto časť práce je netriviálna a je v oblasti, kde neexistuje najlepšie riešenie, takže sa budeme len snažiť nájsť čo najlepšie riešenie na základe analýzy efektivity, inými slovami čo najlepšie aproximovať ideálne riešenie.

V práci je nutné vytvoriť užívateľské rozhranie, ktoré nám umožní takéto hľadanie vykonávať. Ak sa neukážu problémy pri zoznamovaní sa s novými technológiami, alebo prípadne pri algoritmickej časti práce, vybudujeme „mergovací“ (na základe analógie operácie merge pri „version control systems“ ako git a pod.) systém, teda systém, ktorý nám zobrazí piesne a dá nám priestor na úpravu nekorektných zápisov textov, odstránenie duplicít, respektíve spojenie viacerých piesní s nekompletným textom do jednej s kompletným textom.

Bakalárska práca by mi mala vytvoriť prehľad v najnovších technológiách a v neposlednom rade v už predtým spomínanej oblasti informatiky dolovania dát.