Zagovor

1. Danes bom zagovarjal raziskovalno nalogo Analiza besedil s pomočjo podatkovnega rudarjenjnja.
2. Kaj je torej podatkovno rudarjenje? Podkovno rudarjenje je računalniška analiza podatkov oz. v tem primeru besedil. Podatke je potrebno pred analizo primerno pripraviti, saj je računalnik zelo strikten pri načinu zapisu besedila. Tako smo se pri pripravi podatkov poskušali znebiti pregibane besede in jih pretvorili v imenovalnik ednine ter odstranili naglase. Pretvarjanje pregibanih besed v osnovno obliko smo dosegli z lematizacijo, črke oz. besede z naglasom smo se pa znebili s pomočjo kratkega python programa. Rezultate smo predstavili z grafi, tabelami in oblaki besed.
3. Raziskovalna naloga nam priporoča, kako primerjati umetnostna besedila različnih avtorjev med sabo. Prikaže kako primerjati besedila istega avtorja med sabo in prikaže prencip analiziranja posameznega besedila. Prikaže tudi kako si pomagati analizirati besedila s pomočjo računalnika, še posebej v programskem okolju Orange.
4. Zastavili smo si raziskovalna vprašanja, saj je naloga raziskovala nekaj novega in nismo mogli tvoriti hipotez. Prvo zastavljeno vprašanje je bilo, ali je program Orange primeren za analizo slovenskih šolskih besedil. Raziskovalna naloga je bila namenjena šolskim besedilom saj sem tudi jaz šolar in sem si poskušal pomagati analizirati šolska besedila z računalnikom. Pregeledovali smo tudi ali so pridobljeni podatko zanesljivi. Ali lahko računalniški program tako rekoče bere med vrsticami besedila. Ali lahko program Orange primerja besedila med sabo. Kako uporabiti to znanje v vsakdanjem življenju.
5. V teoretičnem delu smo opisali programski jezik python in program Orange. Opisali smo tudi literarne ustvarjalce in literarna obdobja. Vzeli smo tri slovenske klasike, Prešeren, Cankar in Linhart. Pri vsakemu smo vzeli kar nekaj besedil, nekje med 30 – 50 ter jih analizirali. Zelo smo se osredotočili na Prešernovo literarno delo Krst pri Savici, saj smo ga letos obravnavali na daljavo.
6. V eksperimentalnem delu smo razložili način pridobivanja besedil. Način priprave besedil, kako smo odstranili ostrivce in kratice, ter kako smo lematizirali besedilo. Razložili smo tudi prikaz rezultatov in kako smo programirali v programskem okolju Orange. Na desni strani je prikazan eden izmed skupkov modulov, vsak modul na sliki izvaja svoj program in ko jih združimo v želenem zaporedju lahko dobimo različne reprezentacije rezultatov.
7. Prikazan modul je oblak besed in vsebuje vsa prešernova besedila. Oblak besed se bere tako da večja kot je beseda večkrat se pojavlja v besedilu oz. besedilih. Z njim lahko preprosto najdemo osrednji motiv in temo besedila. Opazi se da je lematizacija pri besedi bogomila preagresivna in jo pretvori v besedo bomogomiti.
8. Prikazan je tudi modul sentimentalna analiza kjer so prešernova besedil razporejena v podvrste, ki so označene v legendi desno spodaj. Na grafu horiznotalna os prikazje negativno vrednost besed v besedilu vertikalna os pa negativno vrednost v besedilu. Z grafa lahko razberemo, da prešernovi sonetje nesreče vsebujejo močnejše besede nekatera besedila pa sploh nimajo čustev. Do teh rezultato je potrebo biti malo kritičen, saj so nekatera besedila zelo kratka in ne vsebujejo močnih besed. Modul sentimentalna analiza podpira slovenščino ampak ne do te mere kot angleščino.
9. Programska analiza besedil je dijakom priročna pri prebiranju besedi ter pri pripravi na prebiranje besedil. Pomagajo si lahko tudi med pisanju spisov in besedil, ki se navezujejo oz. opisujejo analizirano besedilo. Posebno je priročna analiza šolskih besedil med delom na daljavo, saj imajo učenci dostop do računalnika.Zanesljivost rezultatov niha glede na modul. Nekateri moduli kot je oblak besed so relativno zanesljivi med tem ko moramo biti malo skeptični do rezultatov pridobljenih z modulom sentimentalna analiza. Program Orange nam omogoča izvrstne načine primerjave ter iskanja podobnosti med besedili.Kot že prej opisana je analiza besedil uporabna v šoli oz. učencem. Analiza besedil je v pomoč tudi pri umetni inteligenci, ki se prilagaja človeškem počutju.
10. Program orange je izvrstna rešitev za analizo besedil. Uporaben je za ljudi, ki ne znajo programiirati in tudi za ljudi ki to znajo. Omogočajo svoj razvoj modulov, seveda pa je za to potrebno zanje programiranja. Sentimentalna analiza v slovenščini malo šepa še posebej na krajših besedilih z šibkejšimi besedami. Oblak besed je s pomočjo lematizacije in pred urejanja besedil močno orodje za analizo slovenskih besedil.