PROJEKTNI ZADATAK ZA NoSQL i big data analitika Text mining

**Svaki student bira netrivijalan tekstualni izvor podataka na engleskom jeziku (Twitter, forum, HTML, kompleksniji tekstualni dokument/ tekstualne dokumente … )**

**Koristeći NTLK biblioteku potrebno je nakon učitavanja teksta:**

1. Normalizirati tekst
2. Tokenizirati tekst
3. Pročistiti tekst od suvišnih elemenata
4. Stemming teksta
5. Lematizacija
6. Maknuti stop words
7. Izračunati frekvencije riječi u listi i sa prikazom u grafičkom obliku
8. Prikazati concordance za najfrekventnije riječi
9. Prikazati kolokacije
10. Izračunati leksički diverzificitet
11. Prikazati primjer primjene korpusa teksta npr Brown
12. Tagirati riječi i prikazati lingvističko stablo na temelju vlastitog uzorka teksta

Navedeni elementi su minimum koji morate odraditi sa tehničke strane, a koriste se u svrhu kreativnog rješenja problema primjerice :

* Usporedba javnog mijenja na twitter populaciji između dva razdoblja 9mj 2020 i 10 mj 2020 vezano uz Covid 19
* Usporedba teksta dva književna djela koja obrađuju istu tematiku
* Usporedba kolekcije vijesti predsjedničkih izbora u Americi sa dva portala iz istog razdoblja koji naginju različitim političkim opcijama
* Analiza tržišta rabljenih automobila pomoću kometara sa društvenih mreža/foruma

Na ispitu je potrebno pokazati rezultate analize kroz prezentaciju koja ne smije trajati više od 5 min , a treba biti usmjerena ka prikazu najznačajnijih rezultata

Na ispitu student će aktivirati kod i pokazati rezultate , te će morati objasniti pojedine tehničke detalje realizacije projekta.

Kodove (Ipython notebook i mongo rješenje + .ppt za ovaj zadatak šaljete na mail [goran@goranklepac.com](mailto:goran@goranklepac.com) kao link sa servisa tipa dropbox tjedan dana prije službenog ispitnog roka, prvi ispitni rok je iznimka , za njega rezultate šaljete do 01.03.2021)