# 2. Seminarska naloga: poročilo

# Inteligentni sistemi

## 14. 1. 2024

## Uvod

To poročilo 2. seminarske naloge je delo Arjana Skoka in Vite Naglič. Cilj naloge je klasifikacija novic v njihove zvrsti z uporabo metod procesiranja naravnega jezika (NLP). Naloga je rešena v pythonu z uporabo več knjižnic, pretežno scikit-learn.

## Predpriprava besedil

Podatki so podani v json obliki. Vsak vnos predstavlja podatke o eni novici. Vsega skupaj imamo 148122 novic. Za vsako novico podanih več atributov, a nas najbolj zanimajo njen naslov, kratek opis in zvrst novice. Podan je tudi link do celotne novice, od koder bi lahko pobrali njeno besedilo, a je proces zelo počasen in se zaradi tega za uporabo tega nisva odločila. Manjkajočih je bilo zelo malo vnosov naslovov in opisov. V teh primerih sva se odločila, da upoštevava samo

Naravni jezik sam po sebi ni primeren za obdelavo z scikitom. Zato ga moramo na to posebej pripraviti. Proces prepriprave je sestavljen iz naslednjih korakov:

1. Delitev stavkov na posamezne besede in transformacija vseh črk v male
2. Odstranitev ločil
3. Odstranitev pogostih besed, ki ne prispevajo ničesar k analizi besedila
4. Lematizacija in korenenje besed, ki besede vrne v svojo osnovno obliko, da jih lahko obravnavamo kot eno samo (npr. running 🡪 run)

Ta proces je šele polovica potrebnega predprocesiranja. Druga polovica pride v obliki vektorizacije.