Soft Computing 2023/24 – K3 – OCR

- Skup podataka za izradu kolokvijuma se nalazi u folderu ***data***.
- Skup podataka za problem prepoznavanja mađarskih reči se nalazi u folderu ***data/pictures***.
- Potrebno je samostalno i programski definisati skup podataka za treniranje na osnovu dostupnih slika.
- ***data/res.csv*** sadrži tačno rešenje za svaku sliku.
- Kreirati rešenje koje će ostvariti najmanji **zbir rastojanja** između tačne i prepoznate reči svih slika.
- **Rastojanje** za jednu sliku se računa na sledeći način:
  - Ako su tačna i prepoznata reč jednake dužine, računa se **Hemingovo rastojanje**, koje predstavlja broj pozicija na kojima se dva stringa jednake dužine razlikuju:
    - rastojanje(“pera”, “pera”) = 0
    - rastojanje(“Pera”, “pero”) = 2
  - Ako su reči različite dužine, računa se **Hemingovo rastojanje** za podstringove tačne i prepoznate reči čija je dužina jednaka dužini kraće reči i na dobijenu vrednost se dodaje apsolutna vrednost razlike dužine dve reči:
    - rastojanje(“zika”, “zikac”) = 1
    - rastojanje(“zika”, “zivac”) = 2
- Za najveći broj bodova (**22**) potrebno je ostvariti **zbir rastojanja <= 2**.

Napomene za izradu i slanje rešenja

- Rešenje zadatka u vidu ***Python*** **skipte** slati na ftn.soft.computing@gmail.com na sledeći način:
  - *Email Subject*: **SC23-G<grupa sa vežbi>-SV-<broj indeksa>**, gde je broj indeksa u formatu **XX-YYYY** (npr. **SC23-G1-SV-07-2020**)
  - *Email Body:* prazan ili sa porukom po izboru
  - *Attachment:* Python skripta nazvana po istom šablonu kao i *Email Subject*: **SC23-G<grupa sa vežbi>-SV-<broj indeksa>.py**
- **Navedena email adresa se koristi isključivo za slanje rešenja.** Eventualna pitanja i nedoumice šaljete asistentima na njihove email adrese.
- Moguće je raditi u *Jupyter Notebook* okruženju, ali se kao rešenje **mora** poslati *Python* skripta. Generisanje skripte od Notebook-a se vrši kroz File meni na sledeći način:
  - **File > Download as > Python (.py)** ili
  - **File > Save and Export Notebook As… > Executable Script**

- Potrebno je omogućiti da se skripta izvršava pomoću sledeće komande:

  python <ime skripte>.py <putanja do foldera sa podacima>

  npr.: **python SC23-G1-SV-07-2020.py data/**

  Preporuka da se za pristupanje putanji do foldera sa podacima koriste argumenti komandne linije (**sys.argv**).

- Prilikom izvršavanja, potrebno je da skripta ispisuje rezultat za svaku ulaznu sliku i u poslednjem redu konačan rezultat rešenja (**zbir rastojanja**). Ispis za svaku sliku treba da bude u sledećem formatu:

  <ime slike>-<tačno rešenje>-<dobijeno rešenje>

  Primer za pojedinačnu sliku: **captcha_6.jpg-kis macska-kis macska**

- Sva rešenja će se evaluirati u *Python3* okruženju sa sledećim instaliranim bibliotekama. **Nije dozvoljeno koristiti druge biblioteke, kao ni pretrenirane modele** (konvolutivnih neuronskih mreža i slično).

- Vreme izvršavanja skripte **ne sme da prekorači 10 minuta** na mašini sa 8 CPU jezgara i 16 GB RAM memorije.

- **Svako nepoštovanje gorenavedenih stavki rezultuje gubitkom bodova.**

- Izvorni kodovi će se analizirati zajedno sa ostalim kodovima iz generacije. **Plagijat znači automatsku diskvalifikaciju i sankcije za plagijatore.**

- **Broj osvojenih bodova se formira na osnovu postignutog rezultata i znanja pokazanog na usmenoj odbrani.**