Euroopan unionin
osarahoittama

PaddleOCR-mallin jatkokoulutus annotoidulla ja synteettisellä aineistolla

Alkuperäinen ongelma

- Tekstintunnistus on ensimmäisiä vaiheita tiedonrikastamisessa -> muut vaiheet ovat suuresti riippuvaisia sen laadusta
- Tekstintunnistus saavuttaa hyviä tuloksia hyvälaatuisella aineistolla,...
- Mutta heikkenee helposti huonompilaatuisilla kuvilla
- Voisiko tuota parantaa?
- Tavoitteet:
  - Luoda avoimen lähdekoodin koneoppimismalli suomenkielelle
  - Joka on tarpeeksi kevyt suurien aineistomassojen tekstintunnistukseen
  - Tutkia paraneeko tulokset Tesseractiin verrattuna?

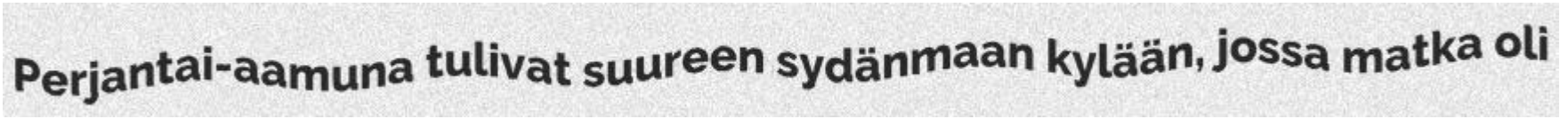
PaddleOCR

- Avoimen lähdekoodin malli
- Pieni ja nopea malli
- Koostuu kolmesta eri koneoppimismallista
  - Rivintunnistus
  - Rivinkäännöksen tunnistusmalli
  - Tekstintunnistusmalli
- Tekstintunnistus malli
  - 80+ kieltä (ml. ruotsi)
- Alkuperäisessä mallissa ongelmia ääkkösten tunnistuksessa
- Lisäksi käsinkirjoitetun tunnistuksessa oli parantamisen varaa

Synteettinen aineisto

- Luotiin lisäaineistoa TRDG-kirjastolla (https://github.com/Belval/TextRecognitionDataGenerator)
- Kirjasto luo rivikuvia tekstisyötteestä
- Voi luoda rivejä, joissa on satunnaista vaihtelua (eri fontteja, tekstinkääntö, tekstin aaltoilu yms.)
- Mahdollistaa suurien aineistojen luonnin nopeasti (10k kuvaa n. 3 min)
- Tekstit Gutenberg-projektista ja Internet Archive:sta

Esimerkkejä luodusta aineistosta



Koulutuksesta

- Jatkokoulutus ELKAn aineistolla ja synteettisellä aineistolla
- Yhteensä rivejä koulutuksessa 160 000 kuvaa
- Koulutettu sekä kone- että käsinkirjoitetulla aineistolla 1900-luvulta
- Käsinkirjoitettua aineistoa kuitenkin aika pieni määrä (8,3k vs 152k kuvaa)
  - Tulokset käsinkirjoitetulla eivät niin hyviä
- Suurin osa aineistosta suomenkielistä, mukana myös ruotsin kieltä ja todella vähän muita kieliä

Tuloksia (CER)

Malli	Elka testi (4273 rivikuvaa)	DALAI testi (3475 rivikuvaa)	KA testi (3247 rivikuvaa)	ELKA testi käsi (714 rivikuvaa)
Tesseract	0,046	0,027	0,044	0,793
Alkuperäinen	0,067	0,039	0,068	0,502
Koulutettu	0,020	0,012	0,023	0,207

- Koulutettu ELKAn aineistolla
- Testattu sekä ELKAn ja KAn aineistolla
  - Vaikuttaa yleistyvän kumpaankin hyvin
- Lisäksi pieniä testejä eri organisaatioiden aineistoilla

Entä sitten?

- Projektissa kokeiltiin myös tietojen nappaamista lomakkeista
- Kokeilussa käytettiin mm. suuria kielimalleja

ALUS[[665.0, 115.0], [913.0, 109.0], [915.0, 181.0], [667.0, 187.0]]
Nimi[[203.0, 253.0], [289.0, 259.0], [286.0, 297.0], [200.0, 290.0]]
Vesterinen[[373.0, 254.0], [599.0, 254.0], [599.0, 288.0], [373.0, 288.0]]
Kotipaikka.[[790.0, 259.0], [979.0, 259.0], [979.0, 293.0], [790.0, 293.0]]
Vaasa[[1061.0, 256.0], [1265.0, 256.0], [1265.0, 291.0], [1061.0, 291.0]]
Rekisteröity[[204.0, 330.0], [383.0, 336.0], [382.0, 371.0], [203.0, 365.0]]
p:nä[[837.0, 336.0], [907.0, 336.0], [907.0, 368.0], [837.0, 368.0]]
kuuta v. 1[[1115.0, 333.0], [1274.0, 333.0], [1274.0, 368.0], [1115.0, 368.0]]
Numerolla[[206.0, 408.0], [360.0, 408.0], [360.0, 443.0], [206.0, 443.0]]
Merkkikirjaimilla[[529.0, 408.0], [778.0, 408.0], [778.0, 440.0], [529.0, 440.0]]
Rakennuspaikka ja.-vuosi Hankilahden. konepajalla.Helsingissä. Konepa-[[201.0, 477.0], [1416.0, 475.
janjohtajan A Hellbergin ja Insinööri Frans Dahlbergin[[236.0, 510.0], [1404.0, 515.0], [1404.0, 550.0], [23
johdol1a vv. 1917-1920.[[234.0, 542.0], [701.0, 545.0], [700.0, 579.0], [233.0, 577.0]]
Taklaus- ja rakennuslaatu[[204.0, 627.0], [611.0, 629.0], [611.0, 664.0], [204.0, 661.0]]
.rautainen, limisaumainen, l-mastoinen,[[594.0, 627.0], [1401.0, 629.0], [1401.0, 671.0], [594.0, 669.0]]
kanneilinen hinaajahöyrylaiva 173 ind.h.v. höyrykoneella.[[231.0, 661.0], [1376.0, 666.0], [1376.0, 701.0],
Päällerakennuksia: Ruhvi keskikannella, ruhvin koroke, lamp[[231.0, 758.0], [1441.0, 768.0], [1441.0, 808
puhuone,"kappi perällä," pannukappi"ja pannukapin koroke.[[224.0, 793.0], [1362.0, 798.0], [1361.0, 840
Pituus keulain ulkoreunain välillä[[204.0, 925.0], [706.0, 925.0], [706.0, 957.0], [204.0, 957.0]]
22.04[[1093.0, 927.0], [1205.0, 927.0], [1205.0, 962.0], [1093.0, 962.0]]
metriä[[1247.0, 925.0], [1347.0, 925.0], [1347.0, 960.0], [1247.0, 960.0]]
5.08[[1111.0, 990.0], [1200.0, 990.0], [1200.0, 1034.0], [1111.0, 1034.0]]
Leveys laudoituksen ulkosivujen välillä[[204.0, 1000.0], [792.0, 997.0], [793.0, 1032.0], [204.0, 1034.0]]
27[[1287.0, 1012.0], [1309.0, 1012.0], [1309.0, 1037.0], [1287.0, 1037.0]]
2.25[[1111.0, 1057.0], [1202.0, 1057.0], [1202.0, 1102.0], [1111.0, 1102.0]]
Korkeus keskisessä poikkileikkauksessa[[206.0, 1074.0], [793.0, 1074.0], [793.0, 1106.0], [206.0, 1106.0]]


```
{
"shipName": "Vesterinen",
 "previousShipName": null,
 "homePort": "Vaasa",
 "previousHomePort": null,
 "registration": {
  "date": null,
  "place": null,
  "number": null,
  "callSign": null
 },
 "construction": {
  "place": "Hankilahden konepajalla Helsingissä",
  "years": "1917-1920",
  "supervisors": "Konepajanjohtajan A Hellbergin ja Insinööri Frans Dahlbergin johdolla"
 },
 "vesselType": "rautainen, limisaumainen, 1-mastoinen, kannellinen hinaajahöyrylaiva
173 ind.h.v. höyrykoneella",
 "engine": "173 ind.h.v. höyrykone",
 "speed": null,
 "dimensions": {
  "length": 22.04,
  "width": 5.08,
  "height": 2.25,
  "unit": "metriä"
 },
 "capacity": {
```

Mistä löytää?

- Koulutettumalli ja API
  - https://github.com/project-AIDA/Finnish_PaddleOCR/tree/main
- Koulutus aineisto
  - https://huggingface.co/datasets/Kansallisarkisto/AIDA_ocr_training_data
- Koulutus malli jatkokoulutusta varten
  - https://huggingface.co/Kansallisarkisto/PaddleOCR_training