# Ajapaiga tekstiandmete töötlemine ning piltide asukoha tuvastamine

Karl Taal, Brigitta Rebane, Laura Katrin Leman

Task 2

**Taust:**
Ajapaik on projekt, mille eesmärk on avastada, kuidas nägid eri kohad välja aastate eest. Ajapaik asub aadressil https://ajapaik.ee/?page=1 ning seal on võimalik tutvuda erinevate fotode ja albumitega, vaadata kaardil nii vanu kui tänapäeval tehtud pilte kaardile märgitud paikadest, märgistada ise tuttavaid kohti kaardile, pildistada ajaloolisi vaateid üle, lisada Ajapaika omalt poolt vanu fotosid jne. Ajapaiga piltidel on metaandmed, mis sisaldavad muuhulgas pildi pealkirja, kirjeldust, geograafilise asukoha infot jpm. Ajapaigas on võimalik ka pilte ning albumeid otsida, kuid need annavad üldiselt vaste vaid siis, kui otsingusõna esineb täpselt samal kujul pildi kirjelduses või pealkirjas. Väga palju fotosid Ajapaigas on ka kaardile märkimata, sest neil puudub asukohainfo.

**Eesmärgid**: (keskendume eelkõige piltidele, millel on eestikeelne pealkiri või kirjeldus)

1. Ajapaigas otsingumootor töötab hetkel nii, et otsingusse sõnu lisades saad sa vasteid ainult täpsete sisestuste korral. Näiteks kui kirjelduses on sõna “majad”, siis otsingusse sisestades “maja” ei tule vastuseks midagi. Esimeseks projekti eesmärgiks on leida iga pildi kohta lemmad, mis saaks lisada pildi välja “keywords” alla, et siis otsing annaks eeltoodud näite põhjal vastuse.
2. Teiseks projekti eesmärgiks valisime piltide kirjelduste ja pealkirjade uurimise. Täpsemalt me üritame pealkirjadest ja kirjeldustest tuvastada asukoha. Seejärel kui oleme leidnud asukoha nimed, üritame me saada sellele kohale vaste koordinaatide kujul ja siis saaks kasutada seda Ajapaiga piltide geotagide populeerimiseks.
3. Kolmanda eesmärgina võib ära märkida veebist info kogumise harjutamise ja praktiseerimise. Andmete saamiseks oleks meil kaks võimalust - kas küsida Ajapaiga asutajatelt, et nad koostaksid meile csv failid andmebaasist küsides vajalikku infot või siis ise veebist (nt APIst) koguda info kokku mingi enda koostatud programmiga. Otsustasime proovida teist varianti.

**Äriedu stsenaarium**

- Oleme leidnud piltidele keywordid otsingumootori tulemuslikumaks muutmiseks.
- Oleme leidnud piltidele koordinaatide kujul asukohad, kus kirjelduses või pealkirjas on märgitud asukoht, kuid pildil ei ole määratud asukohta.
- Oskame veebist infot koguda Pythoni programmidega.

**Olemasolevad ressursid:**

- Ajapaiga API, kus on iga pildi metainfo- pealkiri, kirjeldus, asukoht jne.

- Pythoni moodul geopy annab kohanime järgi sulle koordinaadid.
- Pythoni moodul estnltk leiab sõna lemmad ja võimaldab tuvastada sõnaliike
- Google Maps - selle abil saame käsitsi kontrollida mingit osa tööst, kas leitud koordinaadid on täpsed
- Ajapaiga inimestega ühine Slacki channel.
- Eesti asukohtade nimede loetelu
- asukohanimed- http://xgis.maaamet.ee/knravalik/ (sinna sisestades saaks asukoha nimega koordinaadid eesti formaadis ja need saab konverteerida tavalisteks koordinaatideks.
- Alevikud- https://www.riigiteataja.ee/akt/74442
- Linnad- https://et.wikipedia.org/wiki/Eesti_linnad
- Külad- https://et.wikipedia.org/wiki/Eesti_k%C3%BClade_loend

**Nõuded, eeldused ja piirangud:**
Töö edukaks lõpetamiseks oleks vaja ligipääsu piltide metainfole. Selle oleme me kooskõlastanud Ajapaiga inimestega ja meil on olemas Ajapaiga API, kust saab vajaliku info kätte. Suurimaks piiranguks projekti käigus võib esineda suur andmete hulk. Näiteks lihtsalt kõikide piltide metainfo csv faili paigutamiseks kulus programmil aega 2 tundi. Me keskendume pigem piltidele, kus on olemas eestikeelne kirjeldus ja pealkiri ja see võib vähendada piltide hulka. Piiranguks on kindlasti ka kontrollitav osa tööst, nimelt me ei jõua tõenäoliselt kõiki pilte üle käia kontrollimaks, kas leiti õiged koordinaadid. Peab valima mingi osa ja siis saab umbkaudse täpsuse teada.

**Riskid ja ettenägematud olukorrad**
Suurimaid riske on see, kui meie programm hakkab segamini ajama asukoha nimesid inimeste nimedega. Selle tagajärjel võivad saada paljud pildid vale asukoha info. Selle vältimiseks proovime me luua programmi, mis leiab asukoha ainult siis kui on kindel, et leitud on asukoha nimi. Kahtlasemate otsuste korral jätame asukoha lisamata.

**Terminoloogia:**
- API- Rakendusliides ehk programmiliides ehk rakendustarkvara liides ehk API (inglise keeles *Application Program Interface*) on arvutiprogrammides alamprogrammi määratluste, protokollide ja tööriistade komplekt rakendustarkvara ehitamiseks.
- Lemma- sõna algvorm, nt nimisõnade puhul ainsuse nimetav ja tegusõnade puhul ma-infinitiiv *(tegema, võtma jne)*.

**Kulud ja tulud:**
Kulusid ega tulusid meil ei ole.

**Andmekaeve eesmärgid:**
Oleme koostanud endale veebist infot kogudes andmestikud, kus on ainult väljad, mida meil vaja läheb. Andmeid hoiame csv failides. Kuna me ei leidnud tervikut andmestikku Eesti põhiliste asulate nimestikust, siis koostame selle ise. Me leidsime linnade nimede loetelu, alevike loetelu ja külade loetelu.

**Andmekaeve edustsenaarium:**
Oleme kogunud endale vähemalt 90% ulatuses olemasolevast piltide informatsioonist, millega tööd teha. Asukohanimede andmestiku loomisel võiks olla kaetud 80% ulatuses Eesti suurimate asulate nimed, st et talunimesid kindlasti ei käsitleta jms.

Task 3

**Andmete kogumine**

***Vajalikud andmed***: Ülesannete täitmiseks on meil vaja Ajapaiga piltide pealkirju ja piltide (eestikeelseid) kirjeldusi, hiljem ka märksõnade ja asukoha veerge. Pealkirjade kirjelduste infot kasutame lemmatiseerimiseks ning leitud lemmade seast valime sisukad sõnad, näiteks noomenid, sh pärisnimed, mida lisada pildi märksõnade veergu. Lisaks vaatame, kas pealkirjast ja kirjeldustest on võimalik tuvastada kohanimesid, ning kui see kohanimi veel ei sisaldu pildi asukohainfos, lisame selle sinna. Kohanimede analüüsiks on meil lisaks vaja ka nimekirja Eesti kohanimedest, millega pildi kohta käivat kohanime võrrelda ning tuvastada, kas selline paik päriselt Eestis eksisteerib.

***Andmete kättesaadavus*:** piltide metaandmed, sh pealkirjad, kirjeldused, märksõna- ja asukohaveerud on vabalt kättesaadavad Ajapaiga API kaudu. Need korjasime sealt ja saime üle 160 000 pildi andmed. Asukoha nimed saame ka avalikust andmestikust maa ameti leheküljelt. Eesti asukohtade nimed saab ka internetist leida. Lingid on ülapool välja toodud. Koordinaadid on kättesaadavad pythoni mooduliga geopy.

**Andmete valikukriteerium:** andmed on konkreetselt seotud Ajapaiga fotodega ja sisaldavad infot, mida saame töötluseks ja uue väärtuse loomiseks ära kasutada. Asukoha nimestikku lisame kõik veebist leitavad linnade, alevike ja külade nimed.

**Andmete kirjeldus**

Andmete allikaks on Ajapaiga andmete API (https://opendata.ajapaik.ee/). Pilte on umbes 160 000 ning andmed on CSV-failis, kus igale pildile vastab üks rida. Andmetes on järgnevad väljad:
*id, rephotos, similar_photos, geotags, image, image_unscaled, image_no_watermark, height, width, aspect_ratio, flip, invert, stereo, rotated, date, date_text, title, title_et, title_en, title_ru, title_fi, title_sv, title_nl, title_de, title_no, description, description_et, description_en,*

*description_ru, description_fi, description_sv, description_nl, description_de, description_no, author, uploader_is_author, types, keywords, level, guess_level, lat, lon, geography, bounding_circle_radius, address, azimuth, confidence, azimuth_confidence, source_key, external_id, external_sub_id, source_url, first_rephoto, latest_rephoto, fb_object_id, comment_count, first_comment, latest_comment, view_count, first_view, latest_view, like_count, first_like, latest_like, geotag_count, first_geotag, latest_geotag, dating_count, first_dating, latest_dating, created, modified, gps_accuracy, gps_fix_age, cam_scale_factor, cam_yaw, cam_pitch, cam_roll, video_timestamp, face_detection_attempted_at, perceptual_hash, hasSimilar, licence, user, source, device, area, rephoto_of, video*

Nendest kasutame me nelja: pildi pealkiri (title), pildi kirjeldus (description), märksõnad (keywords) ja asukoht . Pildi pealkiri ja kirjeldus on eestikeelsed tekstiandmed. Mõnel pildil võivad kirjeldused ka puududa. Märksõnade väljas peaksid samuti olema eestikeelsed tekstiandmed, mis on saadud pildi kirjeldustest ja pealkirjadest estnltk-ga lemmatiseerimise ja sisukate noomenite eraldamise teel, kuid need piltidel üldjuhul puuduvad, ning meie töö eesmärk ongi neid lisada. Asukohainfo tuvastamiseks valime nendest sisukatest noomenitest omakorda välja pärisnimed, mida tuvastame estnltk teegiga ja võrdleme Eesti kohanimede andmestikega.

Piltide pealkirjad ja kirjeldused on sobivad tekstianalüüsiks estnltk-ga ning kui need sisadavad pärisnimede seas asukohanimesid, mida kontrollime eraldi kohaloendiga, saame neid kasutada ka pildile asukohainfo lisamiseks.

**Andmete uurimine**

Piltide pealkirjade ja kirjelduste näol on tegemist väga varieeruvate tekstiandmetega. Mõnel juhul on pildi kirjeldus või pealkiri tühjad, pealkirjade ja kirjelduste infoväärtus on väga erinev, kirjeldused võivad olla ka mitme rea pikkused. Kuna tegemist on aga üldiselt korrektse kirjakeelse eesti keelega, sobib see estnltk'ga töötlemiseks.

Andmete puhul võib esineda probleeme nii sellega, et pealkirja või kirjelduste andmed puuduvad, kui ka sellega, et need võivad olla mõningatel juhtudel esitatud mittestandardses kirjakeeles, mida on raskem korrektselt lemmatiseerida. Andmete eeltöötluseks tuleb saada õigel kujul kätte vajalikud väljad ning nendest omakorda võtta eestikeelne tekst välja kujul, millega lemmatiseerija töötada saaks. Lisaks on vajalik lemmatiseerimise tulemused andmetesse märksõnade tulpa lisada. Probleeme võib esineda ka pärisnimede seast kohanimede tuvastamisega, kuna meie võrdlevas kohanimede listis ei pruugi olla kõik

paigad olemas või siis määratakse inimese vms nimi kohanimeks. Muidugi ei proogi ka estnltk kõiki pärisnimesid edukalt tuvastada.

**Andmete kvaliteedi kindlakstegemine**

Esialgsel vaatlemisel tundub, et andmetest piisab kindlasti märksõnade tekitamiseks, kuna pealkirju ja kirjeldusi peaks saama estnltk-ga edukalt lemmatiseerida. Estnltk täpsus lemmatiseerimisel on orienteeruvalt 96%, niisiis peaks programm suutma tuvastada lemmasid ja määrata ka sõnaliike, millest vajame eelkõige noomeneid, sh pärisnimesid. Probleeme võib esineda aga pärisnimede seast kohanimede tuvastamisega, sest selleks oleks vajalik estnltk tuvastatud pärisnimesid võrrelda mingi Eesti kohanimede loendiga, vastasel juhul jääks sõelale ka nimesid, mis ei ole seotud kohtadega. Lisaks soovime leida asukohale vastavad geograafilised koordinaadid, millega pildi saaks kaardile kanda, see aga annab tõenäoliselt tulemuseks laiema ala kui see, mis pildil kujutatud: näiteks kohanimele “Tallinn” vastavad koordinaadid Google Mapsi või Regio järgi valivad tõenäoliselt kaardil Tallinna linna keskpunkti, mitte aga pildil kujutatud koha. Siiski võiks ka selline laiema haardega asukoha märkimine pidada väärtuslikuks, sest praegusel hetkel pole mõningatel juhtudel piltide kohta üldse asukohainfot isegi siis, kui pildi pealkirjas või kirjelduses on tegelikult mõnda asukohta mainitud.
Igal juhul annab Ajapaigale lisaväärtuse pealkirjade ja kirjelduste lemmatiseerimise põhjal märksõnade tekitamine, sest see võimaldab otsingut tehes saada tulemuseks mitte ainult otsitud sõnavormi täpse vaste, vaid kõik selle lemmaga seonduvad vormid.

Task 4. Planning your project (0.5 points)

- Andmete korjamine ja vormistamine
    - andmed tuleb Ajapaiga APIst ja mujalt veebist koguda ja need töötluseks sobivasse formaati üle viia
    - ajakulu: 10 tundi
    - tegija: Karl Taal
- Lemmatiseerimine ja märksõnade loomine
    - andmetest tuleb võtta pealkirjad ja kirjeldused ning need lemmatiseerida estnltk’d kasutades ja sobivad lemmad märksõnadeks lisada
    - ajakulu: 10 tundi
    - tegija: Laura Leman
- Märksõnadest asukohtade tuvastamine
    - lemmatiseeritud märksõnadest (mis on nimisõnad, sh pärisnimed) tuleb tuvastada kohanimed, kasutades kohanimede loendit
    - ajakulu: 10 tundi
    - tegija: Brigitta Rebane
- Projekti ettevalmistamine
    - teema valimine, esialgse idee väljakäimine, ettekanne

    - Ajapaiga tiimiga suhtlemine
    - projekti idee täielik muutus võrreldes esialgu praktikumis esitatud ideega
    - Ajapaiga, selle võimaluste ja puudujääkidega tutvumine jne
    - Projekti raporti esitamine
    - tegijad: kõik
    - ajakulu: 8 tundi igaühe kohta
- Projekti vormistamine ja kokkuvõtmine
    - Githubi vormistamine, kirjelduste-kommentaaride lisamine
    - Plakati tegemine
    - Plakati esitluse ettevalmistamine
    - Plakati esitlemine posterisessioonil
    - tegijad: kõik
    - ajakulu: 8 tundi igaühele
- Ajapaigaga kooskõlastamine
    - kuna teeme koostööd Ajapaigaga ja kasutame nende andmeid, aga neil puudub korralik dokumentatsioon nii andmete kui ka enda kasutatud mudelite/asukoha märkimise jms kohta, peame kõike koordineerima Ajapaiga tegijatega, mistõttu kulub palju aega suhtlemiseks, küsimuste esitamiseks, vastuste ootamiseks, segaduste klaarimiseks jne
    - tegijad: kõik
    - ajakulu: 5 tundi igaühele
- Lisaülesannete täitmine
    - võimalusel soovime kas eeltoodud ülesandeid täiendada või teha uusi asju, näiteks lisaks asukohtadele tuvastada kirjeldusest ka nimesid ja lisada need pildi metaandmete hulka
    - tegijad: kõik
    - ajakulu 5+ tundi igaühele