JodaTuni :D Päiväkirja

1. Kerro osallistuitko viikon opetukseen ja kuvaa päiväkirjan laatimisessa käyttämäsi aineisto.
2. Tiivistä jokaisen luentoviikon keskeiset asiat tekstiksi.
3. Kirjaa lisäksi ylös viisi oivallusta tai tärkeintä oppimaasi asiaa erilliseksi listaksi.
4. Listaa yhdestä kolmeen kehityskohdetta tai flippausvinkkiä luentoviikolle.
5. Kokoa noin kymmenen koodirivin demo viikon aikana oppimastasi.

Viikko 1:

1)

Ensimmäiselle luennolle en päässyt paikanpäälle. Sen sijasta katsoin Panoptosta kevään 2021 aloitusluennon. Katsoin myös vanhan koodiklinikan panoptosta.

2)

Ensimmäisellä luennolla kerrottiin aluksi kurssin historiasta ja tarkoituksesta. Seuraavaksi esiteltiin kurssin suoritusvaatimukset, eli harjoitustyö ja luentopäiväkirja. Luennolla käytiin läpi datatieteen perusteita, kuten mitä se edes tarkoittaa.

Luennolla esiteltiin kurssilla opeteltavia python pohjaisia sovelluksia, kuten Jupyter, ja Pandas. Meille näytettiin myös mitä kannattaa ladata omalle koneelle, että tarvittavat sovellukset saadaan toimimaan. Itse latasin Anaconda Navigatorin luennon aikana. Kurssilla kuulemma koodataan ja sen osaaminen hyvin auttaa paljon. Toivon että muistan vielä jotain Johdatuksesta ohjelmointiin, jonka kävin viisi vuotta sitten.

Perehdyimme siihen mitä tarkoittaa olla ”data scientist”, sekä datatieteen eroihin ja samanlaisuuksiin perinteisen tilastotieteen kanssa. Katsoimme useita kuvia, jotka havainnollistivat datatiedettä eri näkökulmista.

Perjantaina katsoin vanhan koodiklinikka demon. Kirjoittelin itse Jupyterilla mukana ja testailin asioita. Opin paljon ja sovellus vaikutti hyödylliseltä ja helpolta käyttää. Opetus meni nopeasti tosin, minkä takia tallenne oli minulle hyvä.

3)

Viisi oivallusta luennolta:

1. Koodaus on kurssilla huomattavasti tärkeämpää kuin oletin.

2. Datatiede eroaa tilastotieteestä enemmän kuin kuvittelin. Luulin niiden olevan lähes sama asia.

3. Kurssi suoritetaan pelkästään harjoitustyöllä ja oppimispäiväkirjalla. Ei tenttiä.

4. Jupyter vaikuttaa erittäin monikäyttöiseltä ja hyödyllisestä

5. CSV tiedon muokkaaminen ja esittäminen on helppoa Jupyterilla

4)

Katsoin Panoptosta vanhan koodi demon. Siinä mentiin hyvin nopeasti eteenpäin omasta mielestä, mikä hankaloittaa oppimista, jos ei katso tallenteita. Hitaampi eteneminen auttaisi siis oppimista.

5)

import pandas as pd

import matplotlib.pyplot as plt

from matplotlib import style

style.use('ggplot')

# Käytettävien kirjastojen importtaamista. Kirjastoilla plotataan (matplotlib) ja luetaan/muokataan dataa (pandas)

sivu = ’https://jokudatatiedosto.csv’

originaali **=** pd**.**read\_csv(sivu)

tiedosto **=** originaali**.**copy()

print('List of attributes:', tiedosto**.**columns**.**values**.**tolist())

# Lukee halutun nettisivun csv muotoista tietoa ja tallentaa sen muuttujaan. Printtaa listan kolumnien attribuuteista

tiedosto\_plot **=** pd**.**DataFrame([[numero1, numero2, numero3, numero4]])

tiedosto \_plot**.**columns **=** ['Kategoria1', ' Kategoria2', ' Kategoria3', ' Kategoria4']

tiedosto \_plot**.**plot**.**bar(figsize**=**(10,7))

plt**.**legend(loc**=**2)

plt**.**ylabel('Sum $')

plt**.**show()

# Plottaa mitkä vaan numerot siististi.

Viikko 2:

1)

Tein samalla tavalla kuin ensimmäisellä viikolla, eli katsoin vanhan luennon ja koodiklinikan panoptosta. Jatkan varmaan tähän malliin, sillä kotona päiväkirjan kirjoittaminen samalla kun katsoo luentoa, tuntuu mukavammalta. Voin mennä taaksepäin, jos jotain menee ohi, kun kirjoitan.

2)

Luennon aiheisiin kuuluu erilaiset datalähteet, datan kerääminen, laadukkaan datan määritelmät ja ohjelmoinnin rajapinnat sekä ruudunraavinta.

Luento alkaa datatieteen työprosessin kuvauksella. Neljä päävaihetta prosessissa ovat esikäsittely, analyysi, reflektio ja lopussa viestintä. Esikäsittelyyn eli datan keräämiseen ja siivoamiseen kuluu usein suurin osa ajasta. Reflektio ja analyysi toimivat vuorovaikutuksessa. Reflektion aikana voidaan muun muassa huomata tarve erilaiselle analyysille. Viestintä voi olla raportti tai vaikka PowerPoint-esitelmä yrityksen johdolle.

ETL (Extract/Load/Transform) on datainsinöörien prosessi ja DAD (Discover/Access/Distill) datatieteen työprosessi. Erona on arvonluonnin tärkeys datalle. Insinööri keskittyy enemmän datan sujuvaan kulkemiseen.

Analytiikan nelikenttä auttaa ymmärtämään data-analytiikan liiketoimintarelevanssia. Neljä kenttää ovat kuvaileva analytiikka, diagnosoiva analytiikka, ennakoiva analytiikka ja ohjaava analytiikka.

Ryömijöitä ja raapijoita voi käyttää apuna datan käsittelyssä. Ryömijä käy läpi osoitteita verkossa. Ryömijöitä käytetään yleensä sivujen indeksointiin. Prosessi alkaa listalla verkkosivuista, joita ryömijä käy läpi ja etsii lisää hyperlinkkejä jatkaakseen prosessia. Raapija kerää tietoa sivuilta.

Datatyyppejä on useita. Ohjelma käsittelee muuttajan tietoa datatyypin mukaan. Esimerkkejä datatyypeistä on kokonaisluvut, totuusarvot, listat, setit ja tuplet.

3)

Viisi oivallusta:

1. Ryömijät ja raapijat ovat erittäin hyödyllisiä työkaluja datatieteessä.

2. Ryömijöiden ja raapijoiden laillisuus ei ole täysin mustavalkoinen asia.

3. Dataa voidaan tallentaa ihmisen luettavaksi, kuten ASCII, tai koneen luettavaksi binaaridataksi.

4. Datan jalostaminen tarkoittaa datan siivoamista malliin, jossa sitä voidaan hyödyntää. Prosessi riippuu siis käyttötarkoituksesta.

5. Pandas on hyvä työkalu datan jalostamisessa.

4) Luennon Jupyter tiedostossa on mielestäni aivan liikaa dataa, mikä tekee luennon loppupuolesta raskaan ja hankalasti seurattavan.

5)

# kuinka testata datatyyppiä

nro =input(”Numero välillä 0-99?: ”)

try:

print('Tulos:' + str(nro))

except ValueError:

print(”Tämä ei ollut numero.”)

# Scrapyn lataus ja käyttö amazonissa

!pip install scrapy

!scrapy genspider amazon\_scraper amazon.com

import scrapy

class amazonscraper (scrapy.Spider):

name = 'amazon\_scraper'

domains = ['amazon.com']

url = [https://www.amazon.com/mikavaanamazonsivu]

#seuraavaksi scrapy tarvitsee ohjeet mitä tietoa kerätä ja mitä tehdä tiedolla.

Viikko 3:

1)

Jatkan samaan malliin, eli katson panoptosta etäluentoja. Käytän tosin nykyään enemmän aikaa Jupyter työkirjan lukemiseen ja katson luentoa, jos en ymmärrä jotain tai jokin asia jää epäselväksi.

2)

Tämän viikon luennon aiheina ovat koneoppiminen, tekoäly ja niiden sovellukset datatieteessä. Koneoppiminen voidaan jakaa ohjattuun, ohjaamattomaan, osittain ohjattuun ja vahvistettuun oppimiseen. Luennolla keskitytään erityisesti ohjattuun koneoppimiseen.

Ohjatussa oppimisessa koneelle syötetään harjoitusdataa, josta kone muodostaa jonkinlaisin funktion. Data on input/output pareissa. Seuraavaksi kun koneelle syötetään pelkkää input dataa, se osaa ennustaa outputit. Ohjatussa oppimisessa siis oletetaan, että input ja output käyttäytyvät jonkin säännön mukaisesti toisiinsa nähden.

Inputteja kutsutaan myös piirteiksi, ja ne ovat tärkeitä laskennallisessa analytiikassa. Piirre voi olla lähes mikä vaan muuttuja esim. sydämen iskutilavuus tai iskujen määrä minuutissa. Näistä piirteistä voitaisiin koneoppimisen avulla tehdä yksinkertaisia diagnooseja tai seuloja. Datatieteessä kuluu paljon aikaa datan siistimiseen, ja tämä on totta myös koneoppimisen parissa. Ihmisen tarvitsee aina siivota raakadata formattiin, jonka kone ymmärtää. Datalle tarvitsee myös usein tehdä prosesseja kuten standardisointi tai ulottuvuuksien lisääminen/vähentäminen, että tulokset olisivat luotettavia ja hyödyllisiä.

Seuraavaksi menimme läpi esimerkin, jossa otettiin puhelinarvio ja sen sisältämää dataa tutkittiin. Esimerkissä eroteltiin arvion sanat toisistaan ja hukkasanat poistettiin. Plotattuna nähtiin selvästi, mitkä sanat olivat eniten käytössä. Erilaisilla operaatioilla voitiin analysoida arvion lauseenrakenteen ominaisuuksia ja sanojen tunnesävyä.

3)

Viisi oivallusta:

1. Koneoppimista sovelletaan monella tapaa datatieteessä

2. Koneoppiminen on ollut erittäin hyödyllinen työkalu lainojen riskin analysoinnissa.

3. Ennusteiden tekeminen on joskus kiistanalainen asia esimerkiksi, jos rotu otetaan merkittäväksi tekijäksi.

4. Koodissa käytetään erittäin paljon kirjastoja, ja ne kannattaa oppia hyvin.

5. Datatieteessä on useita yleisiä virheitä, joita tulee välttää. Esimerkiksi cherry picking ja gambler’s fallacy.

4) Omasta mielestäni luennot menevät eteenpäin liian nopeasti ja työkirjoissa on erittäin paljon ylimääräisiä linkkejä, mikä vie motivaatiota opiskelulta. En koe mielenkiintoa klikata yhtäkään linkkiä, joka syventäisi mitään aihetta, kun joutuisin äkkiä kuluttamaan 5 tuntia pelkästään yhdellä luennolla.

5)

import nltk

nltk.download('punkt')

import matplotlib.pyplot as plt

from nltk.probability import FreqDist

teksti = (”Mitä vaan tekstiä, jota halutaan analysoida”)

from nltk.tokenize import word\_tokenize

tokenit=word\_tokenize(teksti)

print(tokenit)

# Tästä tulisi ulos ['Mitä', 'vaan', 'tekstiä,', 'jota', 'halutaan', 'analysoida']

# Seuraavaksi plotti

jakauma = FreqDist(tokenit)

jakauma.plot(20, cumulative=False)

plt.show()

# Alle ilmestyisi kuvaaja sanojen yleisyydestä tekstissä