DATA.STAT.510 Matemaattiset ja tilastolliset ohjelmistot, Python, kevät 2021
Harjoitustyö: Verkkokeskustelujen analysointi tilastollisin keinoin

Verkkokeskustelujen analysointi tilastollisin keinoin on tärkeä tapa tutkia yhteiskunnallisia trendejä. Tässä harjoitustyössä tutkitaan erästä laajaa tekstiaineistoa Python-kieltä ja sen kirjastoja käyttäen. Kaikki tässä harjoitustyössä pyydetyt analyysit voidaan tehdä kurssilla läpikäydyillä Python-kielen tarjoamilla keinoilla, mutta oppilaan tulee keksiä, miten keinoja voidaan soveltaa analyyseihin.
Työ palautetaan PDF-tiedostona, jossa on vastaukset, ja kooditiedostona, jossa on käytetty Python-koodi. Palauta työ Moodleen Python-osiossa olevaan harjoitustyön palautusalueeseen.

Käytettävä aineisto on Twenty Newsgroups (20NG); se on kuuluisa tutkimusaineisto, jota on käytettu lukuisissa mm. tilastotieteen ja koneoppimisen tutkimuksissa testiaineistona erilaisten menetelmien vertaamiseen. Aineisto on englanninkielinen, mutta työn tekeminen ei vaadi viestien sisällön ymmärtämistä; tämä englanninkielinen aineisto valittiin sen kuuluisuuden vuoksi ja koska englanninkieliselle tekstille oli saatavissa laadukkaita esikäsittelymenetelmiä.

20NG-aineisto sisältää uutisryhmäviestejä kahdestakymmenestä 1990-luvun Internet-uutisryhmästä. Uutisryhmät ovat seuraavat:

Uutisryhmän numero	Uutisryhmän nimi	Uutisryhmän aihe
1	alt.atheism	keskustelua ateismista
2	comp.graphics	keskustelua tietokonegrafiikasta
3	comp.os.ms-windows.misc	keskustelua Microsoft Windows -käyttöjärjestelmästä
4	comp.sys.ibm.pc.hardware	keskustelua IBM PC -tietokoneiden laitteistosta
5	comp.sys.mac.hardware	keskustelua Apple MacIntosh -tietokoneiden laitteistosta
6	comp.windows.x	keskustelua Unix/Linux -tietokoneiden X-Windows työpöytäympäristöstä
7	misc.forsale	myynti-ilmoituksia
8	rec.autos	keskustelua autoista
9	rec.motorcycles	keskustelua moottoripyöristä
10	rec.sport.baseball	keskustelua pesäpallosta
11	rec.sport.hockey	keskustelua jääkiekosta
12	sci.crypt	keskustelua kryptografiasta
13	sci.electronics	keskustelua elektroniikasta
14	sci.med	keskustelua lääketieteestä
15	sci.space	keskustelua avaruudesta
16	soc.religion.christian	keskustelua kristinuskosta
17	talk.politics.guns	keskustelua asepolitiikasta
18	talk.politics.mideast	keskustelua lähi-idän politiikasta
19	talk.politics.misc	sekalaista keskustelua politiikasta
20	talk.religion.misc	sekalaista keskustelua uskonnosta

20NG-uutisryhmistä on alunperin kerätty yhteensä lähes 20000 viestiä, noin 1000 per uutisryhmä. Esikäsittelyn jälkeen niitä on jäljellä 18622. Viesteille on tehty monenlaista esikäsittelyä, jotta niitä voisi helposti analysoida tilastollisilla ohjelmistoilla. Esikäsitelty data on tallennettu yhdeksi CSV-tiedostoksi **harjoitustyodata.csv**. Sen kukin rivi (ensimmäisen otsikkorivin jälkeen) on yksi uutisryhmäviesti. Kunkin rivin ensimmäinen sarake on viestin numero **messageID**, toinen sarake on uutisryhmän numero **groupID**, tämän jälkeen on viestin kirjoitusaikaan liittyvät muuttujat **secsfrommidnight** ja **secsfrom8am** sekä viestin positiiviseen / negatiiviseen sävyyn liittyvä muuttuja **meanvalences**. Lopuksi on 962 eri sanan esiintymämäärät viestissä. Näitä muuttujia tutkitaan allaolevissa työn osissa.

Osa 1. Sanojen esiintymämäärät. Jokainen sanakirjan sana esiintyy kussakin viestissä nolla tai enemmän kertaa; esiintymämäärä on numeroarvo. Esikäsittelyssä on valittu 962 aineistossa paljon esiintyvää sanaa, ja näiden esiintymismäärät on laskettu jokaisesta viestistä. Jokaisessa uutisryhmässä on monta viestiä, ja näin saadaan kunkin sanan esiintymämäärälle jakauma yli uutisryhmän. Laske allaoleville sanoille niiden esiintymien jakauma (keskiarvo, mediaani, keskihajonta, 0.1% ja 99.9% kvantiilit) eri uutisryhmissä, ja piirrä histogrammi esiintymämääristä.

- Tarkastele sanaa 'freedom' uutisryhmissä sci.crypt, talk.politics.guns talk.politics.mideast ja talk.politics.misc. Missä sanaa käytetään eniten? Entä vähiten?
- Tarkastele sanaa 'nation' uutisryhmissä talk.politics.guns, talk.politics.mideast ja talk.politics.misc. Missä sanaa käytetään eniten? Entä vähiten?
- Tarkastele sanaa 'logic' uutisryhmissä alt.atheism, sci.electronics, talk.politics.misc ja talk.religion.misc. Missä sanaa käytetään eniten? Entä vähiten?
- Tarkastele sanaa 'normal' uutisryhmissä comp.graphics, comp.windows.x, sci.electronics ja sci.med. Missä sanaa käytetään eniten? Entä vähiten?
- Tarkastele sanaa 'program' uutisryhmissä comp.graphics, comp.windows.x, talk.politics.misc, ja comp.sys.mac.hardware. Missä sanaa käytetään eniten? Entä vähiten?

Osa 2. Viestien pituudet. Kunkin viestin pituus on summa kaikkien sanojen esiintymismääristä. Kirjoitetaanko toisissa uutisryhmissä pidempiä viestejä kuin toisissa?

- Piirrä histogrammi viestien pituuksista uutisryhmissä rec.sport.baseball ja rec.sport.hockey. Piirrä sitten histogrammi pituuksien logaritmeistä - näyttääkö logaritmien histogrammi olevan lähempänä normaalijakautunutta?
- Tee tilastollinen testi (odotusarvojen t-testi), poikkeaako viestien logaritmisten pituuksien jakauma näiden kahden uutisryhmän välillä.
- Entä uutisryhmissä rec.autos ja rec.motorcycles, poikkeaako logaritmisten pituuksien jakauma näiden välillä?

Huom. viestien pituudet, edes logaritmimuunnoksen jälkeen, eivät ole aivan tarkalleen normaalijakautuneita joten testien tuloksia kannattaa tässä pitää lähinnä suuntaa antavina tuloksina.

Osa 3. Kirjoitusajat. Tarkastellaan viestien kirjoitusaikoja.

- Viestien kirjoitusajat (sekunteina keskiyöstä) on tallennettu **secsfrommidnight** -muuttujaan. Piirrä histogrammi viestien kirjoitusajoista.

- Tarkastellaan myös muunnettua muuttujaa **secsfrom8am**, jossa viestejä on siirretty 8 tuntia varhaisemmaksi mm. aikavyökkeiden takia. Piirrä histogrammi viestien muunnetuista kirjoitusajoista. Näyttääkö tämä enemmän normaalijakautuneelta? Laske jakauman keskiarvo, mediaani ja keskihajonta.
- Vertaile viestien muunnettuja kirjoitusaikoja ryhmissä comp.graphics ja soc.religion.christian, onko keskimääräisissä muunnetuissa kirjoitusajoissa merkitsevä ero?

Huom. kirjoitusajat, edes muunnoksen jälkeen, eivät ole tarkalleen normaalijakautuneita koska mm. niillä on rajattu arvoväli keskiyöstä seuraavaan, joten testien tuloksia kannattaa tässä pitää lähinnä suuntaa antavina tuloksina.

Osa 4. Korrelaatiot. Lasketaan sanojen esiintymismäärien korrelaatioita.

- Laske korrelaatio sanan 'jpeg' esiintymämäärän ja sanan 'gif' esiintymämäärän välillä yli kaikkien uutisryhmien viestien.
- Tee sama sanoille 'write' ja 'sale'.
- Laske korrelaatio sanan 'jpeg' esiintymämäärän ja sanan 'gif' esiintymämäärän välillä yli uutisryhmän comp.graphics viestien. Onko korrelaatio tässä ryhmässä suurempi kuin kaikkiaan?

Osa 5. Sentimentin analysointi. Sentimentin analysointi tarkoittaa, käytetäänkö viestissä enemmän positiivisia sanoja (`hyvä' = `good', `hieno' = `great' ym.) vai negatiivisia sanoja (`huono`= `bad', `harmillinen' = `unfortunate' ym.). Sentimentin analysointi on tärkeää sosiaalimedian aineistoissa kun tutkitaan, onko keskustelijoiden mielipide jostain aiheesta positiivinen (kuinka positiivinen) vain negatiivinen (kuinka negatiivinen). Viesteille on laskettu valmiiksi laajempaa sanakirjaa käyttäen muuttuja **meanvalences**, joka on "keskimääräinen sentimentti", ts. numeroarvo, kuinka positiivisesti vs. negatiivisesti viestissä puhutaan: viestin sanojen keskimääräinen sentimenttiarvo, kun positiivisen sanan sentimentti on +1 ja negatiivisen -1.

- Sentimenttiarvo on laskettu useiden satunnaismuuttujien keskiarvona; tee normaalisuuden testi, onko sentimenttiarvo normaalijakautunut yli koko aineiston.
- Laske sentimentille jakauma (keskiarvo, mediaani, keskihajonta, 25% ja 75% kvantiilit) eri uutisryhmissä. Piirrä sentimenttiarvoille histogrammi eri uutisryhmille. Etsi kolme positiivisinta uutisryhmää ja kolme negatiivisinta uutisryhmää.
- Tee tilastollinen testi, poikkeaako sentimentin jakauma uutisryhmien comp.sys.ibm.pc.hardware ja comp.sys.mac.hardware välillä. Entä uutisryhmien rec.sport.baseball ja rec.sport.hockey välillä? Entä uutiryhmien rec.autos ja rec.motorcycles välillä?

Osa 6. Uutisryhmän ennustaminen. Yritetään ennustaa viestin sanamäärien perusteella, mihin ryhmään viesti kuuluu. Verrataan kahta uutisryhmää: comp.graphics ja sci.space ja niiden dokumentteja. Merkitään kullekin dokumentille tavoitemuuttujan arvoksi 1 jos se kuuluu edelliseen uutisryhmään ja -1 jos se kuuluu jälkimmäiseen.

Tarkastellaan seuraavia sanaryhmiä, ja sitä, kuinka paljon kussakin dokumentissa on sanoja yhteensä kustakin ryhmästä.

a) Käytetään ensin syötemuuttujana pelkästään sanaa 'jpeg'. Yritä ennustaa lineaariregressiolla

tavoitemuuttujan arvoa (sitä, kumpaan kahdesta ryhmästä dokumentti kuuluu) sanan 'jpeg' esiintymämäärästä. Laske ennusteen keskimääräinen neliövirhe verrattuna tavoitemuuttujan oikeaan arvoon.

b) Käytetään nyt syötemuuttujina sanoja 'jpeg' ja 'earth'. Yritä ennustaa lineaariregressiolla tavoitemuuttujan arvoa (sitä, kumpaan kahdesta ryhmästä dokumentti kuuluu) sanojen esiintymämääristä. Laske ennusteen keskimääräinen neliövirhe verrattuna tavoitemuuttujan oikeaan arvoon.

c) Muodostetaan kahdeksan sanaryhmää, ja lasketaan kullekin ryhmälle sanojen kokonaismäärä kussakin dokumentissa.

- Ryhmä 1: atheism atheist religion christian bibl god church believ faith christ belief religi moral evil islam jesus hell muslim heaven jewish mormon angel scriptur
- Ryhmä 2: version technolog system termin wire charg radio build display code ibm sgi electr electron comput function manual softwar technic print keyboard unix internet machin phone interfac screen network modem mac ram chip batteri hardwar monitor rom printer microsoft font upgrad comp floppi disk cpu widget
- Ryhmä 3: usa freedom american nation societi congress crime punish legal world america opinion polit law countri soviet citizen govern court leader polici civil war attack judg bill vote genocid peac militari tax constitut presid soldier communiti clinton palestinian israel washington russian fbi suspect agenc canada polic turkey armenian armenia
- Ryhmä 4: car road drive driver engin mechan bus auto vehicl
- Ryhmä 5: game record play basebal sport leagu pitch hockey playoff nhl
- Ryhmä 6: organ doctor abort blood diseas heart medic treatment
- Ryhmä 7: earth univers scienc logic evid prove scientif physic caus data experi natur knowledg theori research predict space analysi sun orbit algorithm satellit compress encrypt
- Ryhmä 8: price deal cheap sale

Sanojen määrät eri sanaryhmissä on laskettu valmiiksi muuttujiin **group1** - **group8**. Käytä kaikkia 8 ryhmää ennustamaan lineaariregressiossa, kuuluuko dokumentti comp.graphics vai sci.space -dokumenttiryhmiin. Mitkä sanaryhmät auttoivat eniten ennustuksessa? Laske ennusteen keskimääräinen neliövirhe verrattuna tavoitemuuttujan oikeaan arvoon: kuinka hyvä ennustustulos saatiin?