

Tehtävä 1: Koneoppiminen

Lue oheinen koneoppimista käsittelevä teksti ja vastaa kysymyksiin tekstin ja omien tietojesi pohjalta.

Koneoppiminen on tekoälyn alue, jonka menetelmien avulla voidaan lähestyä tietojärjestelmien toteuttamiseen liittyviä tehtäviä ilman, että järjestelmään ohjelmoidaan suoraan sääntöjä tehtävän ratkaisemiseen. Koneoppimisessa pyritään tuottamaan **opetusdatan** perusteella **malli**, jota voidaan hyödyntää tehtävän suorittamiseen. Esimerkiksi kuvantunnistuksessa opetusdata voi koostua joukosta kuvia joissa tunnistettava kohde esiintyy ja joukosta jossa kohdetta ei ole. Kuvantunnistukseen kehitetyn koneoppimismenetelmän avulla tällaisesta datasta voidaan tuottaa malli, jonka avulla joukot voidaan pyrkiä erottelemaan toisistaan.

Koneoppimismenetelmiä voidaan jakaa niiden käyttämän opetusdatan perusteella. Yhden keskeisen luokan muodostavat **ohjatut** menetelmät, joiden opetusdata koostuu **syötteistä** (input) ja niitä vastaavista **tulosteista** (output). Ohjatun koneoppimisen tavoitteena on muodostaa opetusdatan perusteella malli, joka tuottaa uusille syötteille oikeita tulosteita. Esimerkiksi roskapostin suodattamista voidaan lähestyä käyttäen ohjattua koneoppimista siten, että syöte koostuu sähköpostiviesteistä ja tuloste tiedosta, mitkä viesteistä ovat roskapostia. Tällaisesta datasta opittu malli voi sisältää tietoa esimerkiksi siitä, kuinka usein kukin sana esiintyy roskaposteissa verrattuna muihin sähköposteihin. Ohjattuja menetelmiä voidaan ryhmitellä edelleen niiden tulosteen mahdollisen arvojen perusteella. **Luokittelu** on ohjattua koneoppimista, jossa arvot ennustavat äärellistä määrää ennalta määriteltäviä kategorioita (luokkia). Esimerkiksi roskapostin suodattaminen on luokittelutehtävä, jossa tulosteen arvot edustavat luokkia **roska** ja **ei-roska**. **Regressio** taas on ohjattua koneoppimista, jossa tulosteen arvot ovat jatkuvia. Esimerkiksi lämpötilan ennustaminen on regressiotehtävä: lämpötilan mahdolliset arvot ovat jatkuvalla asteikolla.

Ohjaamattomassa koneoppimisessa opetusdata ei sisällä tulosteita, ja tavoitteena on syöte-datan rakenteen mallintaminen esimerkiksi sen perusteella, mitkä syötteistä ovat keskenään samankaltaisia. **Ryvästäminen** on ohjattua luokittelua vastaava ohjaamaton oppimistehtävä, jossa tavoitteena on samankaltaisten syötteiden ryhmittely yhteen. Uutisteksteistä koostuvaa dataa ryvästämällä voi olla mahdollista löytää esimerkiksi ryhmiä, jotka käsittelevät politiikkaa, urheilua, ja säätä. Toisaalta, koska ohjaamattomassa oppimisessä toivottua tulosta ei ole määriteltä ennalta, ryvästäminen voi tuottaa myös odottamattomia ryhmiä: ryvästysmenetelmä voi esimerkiksi ryhmitellä uutisia aihepiiriin sijasta sen mukaan, ovatko uutistekstit pitkiä vai lyhyitä.

Koska koneoppimismenetelmät eivät pysty suoraan käsittelemään kaikkia mahdollisia syötteiden muotoja, datan esitystapa koneoppimista varten on yksi keskeinen kysymys näiden menetelmien soveltamisessa. Sen sijaan, että koneoppimismenetelmälle annettaisiin suoraan syötteeksi esimerkiksi suomen kielellä kirjoitettuja tekstidokumentteja, syötedata esitetään tyypillisesti **piirteiden** muodossa: kukin piirre edustaa jotain syötteen ominaisuutta, ja piirteen arvo sitä, missä määrin tämä ominaisuus kuvaa syötettä. Esimerkiksi tekstin luokittelua varten on mahdollista määritellä jokaiselle kielen sanalle piirre, ja asettaa näiden arvot sen perusteella, kuinka monta kertaa kukin sana esiintyy kussakin dokumentissa. Sen sijaan, että piirteet määriteltäisiin tällä tavalla etukäteen, voidaan myös syötteiden esitystapa oppia datasta **piirreoppimisen** avulla.

Toisin kuin ihmisten oppimisprosessissa, opitun tiedon unohtaminen ei ole haaste koneoppimisessa: koska kone voi opetella opetusdatan ulkoa, on helppoa kirjoittaa ohjelma, joka palauttaa aina oikean tulosteen kullekin opetusdatan syötteelle. Tällainen ohjelma ei kuitenkaan ole hyö-

dyllinen, ellei se pysty tuottamaan toivottuja tulosteita myös sellaisille syötteille, jotka eivät esiinny opetusdatassa. Esimerkiksi roskapostin suodatin, joka tunnistaa roskaksi ainoastaan sellaiset viestit, jotka on jo aikaisemmin luokiteltu roskapostiksi erehtyisi jokaisen uuden roskapostiviestin kohdalla, joka poikkeaa edes yhdellä merkillä opetusdatan viesteistä. Koneoppimisen keskeisenä tavoitteena onkin sellaisten mallien tuottaminen, jotka ennustavat toivottuja tuloksia opetusdatan lisäksi uusille syötteille, eli pystyvät **yleistämään** opetusdatan esimerkeistä. Mallin arviointiin käytetään **testidataa**, joka on erillinen opetusdatasta ja jota ei ole hyödynnetty menetelmän kehittämisessä. Jotta tuotettu malli olisi hyödyllinen käytännön tehtävissä ja sen arviointi antaisi realistisen tuloksen sen tarkkuudesta, on tärkeää että niin opetusdata kuin testidatakin edustavat sitä dataa johon valmista mallia tullaan soveltamaan.

Koneoppiminen ei aina tuota toivottuja tuloksia, ja sen menetelmien soveltamisessa on useita haasteita. Monet näistä haasteista liittyvät dataan. Jos opetusdatan määrä on riittämätön, sen perusteella voi olla mahdotonta oppia mallia, joka tuottaa toivottuja tuloksia: esimerkiksi tekstin luokittelumallin kehittämiseen tarvitaan usein satoja tai tuhansia esimerkkejä kustakin luokasta. Jos data taas ei edusta tehtävää johon menetelmää kehitetään, malliin voi syntyä vääristymiä ja se voi epäonnistua käytännössä. Jos esimerkiksi puheentunnistusmallin kehittämisessä käytetään vain selkeästi äänitettyä yleiskieltä, malli voi epäonnistua hälyisessä ympäristössä ja murteiden tunnistamisessa.

Koneoppimistehtäviin on mahdollista soveltaa monia eri menetelmiä, ja niiden toteuttamiseen on olemassa työkaluja monille eri ohjelmointikielille. Useat koneoppimismenetelmät perustuvat tilastollisiin menetelmiin. Esimerkiksi yksinkertaisen roskapostin luokittelumenetelmän voi muodostaa laskemalla kuinka usein kukin sana esiintyy kunkin luokan (**roska/ei-roska**) viesteissä opetusdatassa ja laskemalla sanakohtaisista todennäköisyyksistä arvion sille, että uusi viesti on roskaa. Monet viime vuosina kehitetyt koneoppimismenetelmät perustuvat **neuroverkkoihin**, jotka mallintavat toisiinsa kytkettyjä yksinkertaisia laskenta-yksiköjä. Näistä mm. kuvantunnistamisessa hyviä tuloksia saavuttaneet **syväoppimismenetelmät** mallintavat suuria, jopa miljoonista laskenta-yksiköistä koostuvia neuroverkkoja. Nämä menetelmät voivat oppia erittäin monimutkaisia kuvauksia syötteistä tulosteisiin, mutta niiden kouluttamiseen voidaan toisaalta tarvita erittäin suuri määrä dataa.

Kysymykset

Mitä koneoppimismenetelmiä seuraavissa järjestelmissä käytetään? *Valitse kullekin kysymyksissä 1.1–1.4 kuvatulle järjestelmälle oikeat vaihtoehdot. Valitse vain suoraan kuvauksen pohjalta perusteltavat vaihtoehdot, ja huomaa että oikeita valintoja voi olla useampia.*

Pisteytys: Täsmälleen oikeasta vastauksesta maksimipisteet. Puuttuvista tai ylimääräisistä vaihtoehdoista vähennetään pisteitä. Kunkin kysymyksen minimipistemäärä on 0.

Kysymys 1.1. (2 p.)

Järjestelmän A tarkoitus on tunnistaa tekstin kirjoittaja sen sisällön perusteella. Järjestelmä perustuu tilastolliseen malliin, johon on koostettu laajasta tekstiaineistosta tieto siitä, kuinka usein kukin kirjoittaja käyttää tiettyä sanaa tai sanojen yhdistelmää. Kun järjestelmään syötetään uusi teksti, se arvioi mallin perusteella kuka aineistossa edustettuna olevista henkilöistä on todennäköisimmin kirjoittanut tekstin. Järjestelmässä ja aineistossa on rajoituksena, että kullakin tekstillä on täsmälleen yksi kirjoittaja.

- ☐ Ohjattu koneoppiminen
- ☐ Ohjaamaton koneoppiminen
- ☐ Luokittelu
- ☐ Regressio
- ☐ Ryvästäminen
- ☐ Ei mitään yllä olevista

Kysymys 1.2. (2 p.)

Järjestelmä B on kehitetty osakekaupan tueksi. Järjestelmä sisältää mallin, joka pyrkii ennustamaan kunkin markkinoilla olevan osakkeen tulevan arvon sen perusteella, miten osakkeiden arvot ovat muuttuneet viimeisen vuoden aikana. Malli perustuu matemaattiseen funktioon, jonka parametrit on sovitettu käyttäen osakkeiden arvojen historiallista kehitystä edustavan tietokannan dataa.

- ☐ Ohjattu koneoppiminen
- ☐ Ohjaamaton koneoppiminen
- ☐ Luokittelu
- ☐ Regressio
- ☐ Ryvästäminen
- ☐ Ei mitään yllä olevista

Kysymys 1.3. (2 p.)

Järjestelmän C avulla voidaan hakea sähköpostikokoelmasta kaikki viestit joissa esiintyy käyttäjän määrittelemä sana. Järjestelmä perustuu tietokantaan, johon on koostettu kaikki kokoelmassa esiintyvät sanat ja kullekin sanalle viestit, joissa sana esiintyy. Järjestelmä tukee myös hakumuotoa, jossa tietokannasta haetaan syötteenä saadun sanan lisäksi myös sanoja, joiden järjestelmän kehittäjä on digitaalisen sanakirjan perusteella määritellyt tarkoittavan samaa asiaa.

- ☐ Ohjattu koneoppiminen
- ☐ Ohjaamaton koneoppiminen
- ☐ Luokittelu
- ☐ Regressio
- ☐ Ryvästäminen
- ☐ Ei mitään yllä olevista

Kysymys 1.4. (2 p.)

Järjestelmä D on kehitetty tekstiaineistojen organisointiin. Järjestelmä esittää kunkin sen tietokannassa olevan tekstin siinä esiintyvien sanojen jakaumana, ja muodostaa teksteistä ryhmiä, joiden tekstien jakaumat ovat keskenään samankaltaisia. Järjestelmä nimeää kunkin ryhmän sellaisten sanojen mukaan, jotka esiintyvät sen teksteissä useammin kuin muiden ryhmien teksteissä.

- ☐ Ohjattu koneoppiminen
- ☐ Ohjaamaton koneoppiminen
- ☐ Luokittelu
- ☐ Regressio
- ☐ Ryvästäminen
- ☐ Ei mitään yllä olevista

Mistä koneoppimiseen liittyvistä syistä seuraavat järjestelmät eivät tuota toivottuja tuloksia? Valitse kullekin kysymyksissä 1.5–1.7 kuvatulle järjestelmälle oikeat vaihtoehdot, määrittele ongelma tai ongelmat tarkemmin omin sanoin, ja kerro miten ne voitaisiin ratkaista. Vastaa kuhunkin kysymykseen lyhyesti enintään 40 sanalla. Valitse vain suoraan kuvauksen pohjalta perusteltavat vaihtoehdot, ja huomaa että oikeita valintoja voi olla useampia.

Esimerkkivastaus: Kysymys 1.x: Opetusdatan valinta – Sähköpostin luokittelumenetelmän opetusdata koostuu pelkistä roskapostiviesteistä. Opetusdataa tulisi laajentaa siten, että se sisältää myös esimerkkejä viesteistä, jotka eivät ole roskapostia.

Pisteytys: Täsmälleen oikeasta vastauksesta maksimipisteet. Puuttuvista tai ylimääräisistä vaihtoehdoista vähennetään pisteitä. Sanallinen perustelu on pakollinen. Puuttuva sanallinen perustelu tarkoittaa automaattisesti 0 pistettä. Kunkin kysymyksen minimipistemäärä on 0.

Kysymys 1.5. (2 p.)

E kehittää järjestelmää eläinlajien tunnistamiseen valokuvista Python-kieltä ja uusimpia syväoppimisen menetelmiä käyttäen. Järjestelmän kehitykseen koostetaan opetusdata, jossa on yksi valokuva kustakin eläinlajista jonka järjestelmä pyrkii tunnistamaan, sekä vastaavasti koostettu testidata. Järjestelmä koulutetaan luokittelemaan syötteenä oleva kuva siinä esiintyvää eläinlajia vastaavaan luokkaan. Järjestelmä epäonnistuu kuitenkin tuottamaan toivottuja tuloksia riippumatta siitä, miten se toteutetaan ja koulutetaan.

- ☐ Opetusdatan valinta
- ☐ Testidatan valinta
- ☐ Ohjelmointikielen valinta
- ☐ Koneoppimismenetelmän valinta
- ☐ Piirteiden valinta

Kysymys 1.6. (2 p.)

F omistaa laajan valokuva-aineiston, joka on järjestetty aiheittain siten, että kukin kuva esittää täsmälleen yhtä aihetta. F on laajentamassa aineistoa samoista aiheista otetuilla kuvilla, ja toivoo koneoppimispohjaista järjestelmää avuksi uusien kuvien järjestämiseen aiheiden perusteella. Järjestelmän kehittämisen tueksi F koostaa kustakin aiheesta sitä esittävien valokuvien kokoelman, joka lähetetään ohjelmistokehitykseen erikoistuneelle konsulttifirmalle. Tämä tuottaa järjestelmän käyttäen Lua-kielillä toteutetun koneoppimiskirjaston ryvästysmenetelmiä. Kun järjestelmä otetaan käyttöön, todetaan kuitenkin että sen tuottama ryhmittely poikkeaa täysin etukäteen toivotusta.

- ☐ Opetusdatan valinta
- ☐ Testidatan valinta
- ☐ Ohjelmointikielen valinta
- ☐ Koneoppimismenetelmän valinta
- ☐ Piirteiden valinta

Kysymys 1.7. (3 p.)

G kehittää kännykkä-applikaation tueksi kasvontunnistusmenetelmää käyttäen Java-kieltä ja neuroverkkoihin perustuvaa koneoppimiskirjastoa. Datan keräystä varten otetaan studioolosuhteissa kattava kokoelma kuvia, joissa esiintyy monia eri kasvoja, ja vastaava kokoelma, jossa kasvoja ei esiinny. Aineisto jaetaan satunnaisesti opetus- ja testidataan, ja opetusdatan perusteella koulutetun piirreoppimispohjaisen järjestelmän todetaan luokittelevan testidatan materiaalin korkealla tarkkuudella. Tästä huolimatta kännykkä-applikaation käyttäjäpalautteen mukaan järjestelmän kasvontunnistustoiminto on erittäin epäluotettava.

-
- ☐ Opetusdatan valinta
 - ☐ Testidatan valinta
 - ☐ Ohjelmointikielen valinta
 - ☐ Koneoppimismenetelmän valinta
 - ☐ Piirteiden valinta
-

Tehtävä 2: Ongelmanratkaisu

Kysymys 2.1. (1 p.)

Viljelijällä on kolme juuressäkkiä. Yhdessä lukee ”porkkanoita”, toisessa ”lanttuja” ja kolmannessa ”porkkanoita ja lanttuja”. Nimilaput ovat menneet sekaisin, ja jokainen säkki on nyt merkitty väärin. Saat valita yhden säkin, josta nostat umpimähkään yhden juureksen. Mikä säkki tulee valita voidaksesi nimetä oikein kaikki kolme säkkiä?

Pisteytys: Täsmälleen oikea ja perusteltu vastaus antaa maksimipisteet. Ilman perustelua oikeastakaan valinnasta ei saa pisteitä. Minimipistemäärä on 0.

- ☐ Porkkanoita
- ☐ Lanttuja
- ☐ Porkkanoita ja lanttuja

Nimeä kaikki säkit nostamasi juureksen perusteella, ja perustele vastauksesi lyhyesti. Vastauksesi kokonaispituus tulee olla **enintään** 100 sanalla, vähempikin sanamäärä riittää.

Kysymys 2.2. (1 p.)

Aapo rakastaa Iidaa, ja Iida rakastaa Sakua. Aapo on opiskelija, mutta Saku ei ole. Kysymys kuuluu: rakastaako opiskelija sellaista henkilöä, joka ei ole opiskelija?

Pisteytys: Täsmälleen oikea ja perusteltu vastaus antaa maksimipisteet. Ilman perustelua oikeastakaan valinnasta ei saa pisteitä. Minimipistemäärä on 0.

- ☐ Kyllä
- ☐ Ei
- ☐ Ei voi päätellä

Perustele vastauksesi lyhyesti **enintään** 100 sanalla, vähempikin sanamäärä riittää.

Kysymys 2.3. (2 p.)

Loogikot A, B ja C saapuvat kahvilaan. He käyvät siellä keskustelun erittäin älykkään kahvilanpitäjän kanssa.

- Kahvilanpitäjä: “Otatteko kaikki kahvia?”
- Loogikko A: “En tiedä.”
- Loogikko B: “En tiedä.”
- Loogikko C: “Emme.”

Kenelle kahvilanpitäjä kaataa kahvia ja kenelle ei? Vastaa jokaisen loogikon osalta erikseen.

Pisteytys: Täsmälleen oikea ja perusteltu vastaus antaa maksimipisteet. Vääristä tai puuttuvista valinnoista vähennetään pisteitä. Ilman perusteluja oikeistakaan valinnoista ei saa pisteitä. Minimipistemäärä on 0.

	Kyllä	Ei	Ei voi päätellä
Loogikko A	☐	☐	☐
Loogikko B	☐	☐	☐
Loogikko C	☐	☐	☐

Perustele vastauksesi lyhyesti **enintään** 100 sanalla, vähempikin sanamäärä riittää.

Kysymys 2.4. (1 p.)

Eräaseen tilaisuuteen on kokoontunut yhtä monta miestä ja naista. Paikalla on kaksi naista, jotka eivät välirikon vuoksi kätele toisiaan ja yksi mies, joka ei vakaumuksen vuoksi voi kätellä muita naisia kuin paikalla olevaa omaa vaimoaan. Muuten kaikki kätelevät toisiaan. Kaiken kaikkiaan tapahtuu 84 kättelyä. Kuinka monta osallistujaa tilaisuudessa on? Kukaan ei kätele itseään.

Anna vastauksesi kokonaislukuna. Mikäli et anna kokonaislukua, vastauksesi ei tallennu.

Kysymys 2.5. (3 p.)

Rusakko ja kenguru kinastelevat, kumpi on nopeampi. Ne päättävät kilpailla. Kilpamatka on meno-paluu kahden kiven välillä, jotka ovat sadan metrin etäisyydellä toisistaan. Matkaan lähdetään ensimmäisen kiven kohdalta ja kääntyä saa aikaisintaan toisen kiven kohdalla. (Luonnollisesti ilmassa ei voi kääntyä.) Kengurun loikan pituus on aina kolme metriä ja rusakon loikka on aina kaksi metriä. Rusakko loikkii kolme loikkaa siinä ajassa kuin kenguru kaksi. Jos oletetaan, että kääntymiseen eläimet käyttävät yhtä paljon aikaa, niin kumpi voittaa kilpailun, ja kuinka monen oman loikkansa päässä hävinnyt eläin on?

voittaa. Hävinnyt oli jäljessä

+

loikkaa, kun voittaja ohitti maalikiven.

Kysymys 2.6. (1 p.)

Viljelijällä on kaksi säkkiä, toisessa on porkkanoita ja toisessa on perunoita. Viljelijän ja hänen säkkiensä yhteismassa on 170 kg. Mikä on porkkanasäkin massa kilogrammoina, jos viljelijä painaa sata kilogrammaa enemmän kuin säkit yhteensä, ja perunasäkki painaa 60 % vähemmän kuin porkkanasäkki?

Anna vastaus desimaalilukuna ilman yksiköitä. Mikäli annat vastauksen jossain muussa muodossa, vastauksesi ei tallennu.

Kysymys 2.7. (6 p.)

Lentokentällä on viisi rahtikonetta. Ne ovat rinnakkain keula terminaalia kohden. Jokaisella rahtikoneella on

- eriväriset siivet (mustat, oranssit, valkoiset, vihreät tai violetit),
 - eri kohdekaupunki (Amsterdam, Ateena, Hong Kong, New York tai Tukholma),
 - eri lasti (desinfointiaine, hengityssuojaimet, kertakäyttökäsineet, virustestit tai WC-paperi),
 - eri lähtöaika (4, 5, 6, 7 tai 8) ja
 - eri kotimaa (Espanja, Italia, Kiina, Saksa tai Yhdysvallat).
1. Kiinalainen rahtikone lähtee kello viisi ja kuljettaa hengityssuojaimia.
 2. Valkoisilla siivillä oleva italialainen rahtikone on hengityssuojaimia kuljettavan rahtikoneen vieressä vasemmalla puolella. (Rahtikoneita katsotaan terminaalista päin.)
 3. Yhdysvaltalainen rahtikone lähtee kello kahdeksan.
 4. Keskimmaisella rahtikoneella on vihreät siivet.
 5. Espanjalainen rahtikone on menossa Hong Kongiin. Se ei kuljeta virustestejä.
 6. Amsterdamiin suuntaava rahtikone lähtee viideltä.
 7. Desinfointiainetta kuljettavan rahtikoneen oikealla puolella on rahtikone, joka menee Tukholmaan.
 8. New Yorkiin suuntautuva rahtikone lähtee kello neljä.
 9. WC-paperia kuljettavan rahtikoneen vieressä olevalla rahtikoneella on violetit siivet.
 10. Oransseilla siivillä varustettu rahtikone menee Amsterdamiin.
 11. Saksalainen rahtikone lähtee kello kuusi ja se on Tukholmaan menossa olevan rahtikoneen vieressä oikealla puolella.
 12. Reunimmaisena on rahtikone, joka kuljettaa kertakäyttökäsineitä.
 13. Kello kuusi lähtevän rahtikoneen vieressä olevalla rahtikoneella on mustat siivet
 14. Kertakäyttökäsineillä lastatun rahtikoneen vieressä on rahtikone, joka kuljettaa WC-paperia
 15. Kello seitsemän lähtee rahtikone, jolla on vihreät siivet

Minkämaalaisella rahtikoneella on oranssit siivet? (Max 1p.)

- ☐ kiinalainen
- ☐ italialainen
- ☐ yhdysvaltalainen
- ☐ espanjalainen
- ☐ saksalainen

Milloin italialainen rahtikone lähtee? (Max 1 p.)

- ☐ Kello 4:00
- ☐ Kello 5:00
- ☐ Kello 6:00
- ☐ Kello 7:00
- ☐ Kello 8:00

Minkämaalainen rahtikone menee Ateenaan? (Max 2 p.)

- ☐ kiinalainen
- ☐ italialainen
- ☐ yhdysvaltalainen
- ☐ espanjalainen
- ☐ saksalainen

Minkämaalainen rahtikone kuljettaa virustestejä? (Max 2 p.)

- ☐ kiinalainen
- ☐ italialainen
- ☐ yhdysvaltalainen
- ☐ espanjalainen
- ☐ saksalainen

Tehtävä 3: Merkkijonohaku

- Käyttöohje ruudukolle – Tämän sivun lopussa on lyhyt käyttöohje vastausruudukon käyttämiseen. Sama teksti on myös PDF-version lopussa.

Merkintöjä

Merkkijonohaku on yleinen toimenpide tietojenkäsittelyssä. Kun on annettu hahmomerkkijono p sekä tekstimerkkijono t , tehtävänä on etsiä tekstistä t kaikki sellaiset indeksit j , joiden kohdasta alkaen teksti sisältää hahmon p . Esimerkiksi jos $p = \text{"uubu"}$ ja $t = \text{"ubwubuuubuu"}$, tulisi löytää esiintymä kohdasta $j = 7$, koska $t[7..10] = \text{"uubu"} = p$ (katso kuvaa 1). Edellä käytetyt merkinnät kuvataan seuraavaksi tarkemmin.

Viittaamme merkkijonon merkkeihin antamalla merkin indeksin hakasuluissa merkkijonomuutujan perässä. Esimerkiksi $x[1]$ tarkoittaa merkkijonon x ensimmäistä merkkiä. Lisäksi merkintä $x[i..j]$ tarkoittaa merkkijonon x alimerkkijonoa, joka alkaa indeksistä i ja päättyy indeksiin j . Merkintä $|x|$ tarkoittaa merkkijonon x pituutta. Muotoa $x[1..j]$ olevaa alimerkkijonoja sanotaan merkkijonon x alkuosaksi ja vastaavasti muotoa $x[i..|x|]$ olevaa alimerkkijonoa sanotaan loppuosaksi. Käytämme operaattoreita $=$ ja $\neq$ luonnolliseen tapaan tarkoittamaan merkkien tai merkkijonojen samanlaisuutta ja erilaisuutta.

Esimerkki 1. Jos merkkijono x on "university", pätee mm. että:

- $|x| = 10$.
- $x[1] = \text{"u"}$, $x[3] = x[8] = \text{"i"}$ ja $x[6] = \text{"r"}$. Toisaalta esim. $x[6] \neq x[8]$ ja $x[3] \neq x[1]$.
- $x[1..3] = \text{"uni"}$, $x[3..3] = \text{"i"}$ ja $x[4..10] = \text{"versity"}$. Toisaalta $x[3..4] = \text{"iv"} \neq x[8..9] = \text{"it"}$.
- $x[1..6] = \text{"univer"}$ on yksi merkkijonon x alkuosa ja $x[7..10] = \text{"sity"}$ on yksi sen loppuosa.

Naiivi haku

Kuva 1 havainnollistaa yksinkertaista ns. naiivia hakumenetelmää, joka tutkii ehdon $p = t[j..j + |p| - 1]$ yksitellen jokaiselle mahdolliselle hahmon esiintymiskohdalle $j = 1, \dots, |t| - |p| + 1$. Kuva käyttää tehtävän alussa mainittuja hahmoa $p = \text{"uubu"}$ ja tekstiä $t = \text{"ubwubuuubuu"}$. Hahmo asetetaan aluksi kohdikkain tekstin alun kanssa eli kohtaan $j = 1$, seuraavaksi kohtaan $j = 2$, ja niin edelleen, siirtäen hahmoa askel kerrallaan tekstin loppua kohti. Kunkin tällaisen kohdistusaskelen kohdalla tutkitaan, esiintyykö hahmo nykyisessä tekstin kohdassa j . Tämä ehdon $p = t[j..j + |p| - 1]$ tutkiminen tapahtuu vertaamalla merkkejä yksitellen hahmon alusta loppuun (vasemmalta oikealle): ensin vertaillaan merkkejä $p[1]$ ja $t[j]$, ja jos ne täsmäävät, seuraavaksi merkkejä $p[2]$ ja $t[j + 1]$, ja niin edelleen. Formaalisemmin ilmaistuna merkkejä $p[i]$ ja $t[j + i - 1]$ vertaillaan järjestyksessä $i = 1, \dots, |p|$. Merkkivertailut lopetetaan kesken, jos löytyy epätäsmävyä merkkipari $p[i]$ ja $t[j + i - 1]$, koska tällöin $p \neq t[j..j + |p| - 1]$. Jos taas merkit saadaan täsmätyä viimeiseen merkkipariin $p[|p|]$ ja $t[j + |p| - 1]$ asti, pätee $p = t[j..j + |p| - 1]$ ja tämä indeksin j kohdalta löydetty hahmon esiintymä kirjataan muistiin.

Kuvassa 1 hahmon täsmätyt merkit on merkitty alleviivauksella ja epätäsmätyt merkit ylivii-vauksella. Esimerkiksi kohdassa $j = 1$ ensimmäinen merkkivertailu täsmää $p[1] = t[1] = \text{"u"}$, ja seuraava vertailu epätäsmää $p[2] = \text{"u"} \neq t[2] = \text{"b"}$. Merkkivertailut lopetetaan epätäsmäyksen jälkeen ja haku siirtyy seuraavaan kohtaan $j = 2$. Kyseisessä kohdassa tapahtuu heti alussa

	1	2	3	4	5	6	7	8	9	10	11
	u	b	w	u	b	u	u	u	b	u	u
j=1	<u>u</u>	⊕	b	u							
j=2		⊕	u	b	u						
j=3			⊕	u	b	u					
j=4				<u>u</u>	⊕	b	u				
j=5					⊕	u	b	u			
j=6						<u>u</u>	<u>u</u>	⊕	u		
j=7							<u>u</u>	<u>u</u>	<u>b</u>	<u>u</u>	
j=8								<u>u</u>	⊕	b	u

Kuva 1: Naiivi haku

epätäsmäys merkkien $p[1] = \text{"u"}$ ja $t[2] = \text{"b"}$ välillä. Huomaa, että myöhemmin kohdassa $j = 7$ kaikki merkkivertailut täsmäävät ja löydetään hahmon esiintymä tekstissä.

Täsmätyn loppuosan heuristiikka

Naiivi menetelmä siirtää hahmoa yhden askeleen kerrallaan suhteessa tekstiin (eli indeksiä j kasvatetaan aina yhdellä). Täsmätyn loppuosan heuristiikka pyrkii tekemään isompia kuin yhden askeleen pituisia siirtymiä. Yksi oleellinen muutos on kääntää ehdon $p = t[j..j + |p| - 1]$ tutkimisen yhteydessä tehdyt merkkivertailut tapahtumaan lopusta alkuun päin (oikealta vasemmalle): ensin vertaillaan merkkejä $p[|p|]$ ja $t[j + |p| - 1]$, ja jos ne täsmäävät, seuraavaksi merkkejä $p[|p| - 1]$ ja $t[j + |p| - 2]$, ja niin edelleen mahdollisesti merkkeihin $p[1]$ ja $t[j]$ asti. Formaalisemmin ilmaistuna merkkejä $p[i]$ ja $t[j + i - 1]$ vertaillaan siis järjestyksessä $i = |p|, \dots, 1$. Vertailut lopetetaan kesken, jos löytyy epätäsmävyvä merkkipari $p[i]$ ja $t[j + i - 1]$.

Jos ensimmäinen epätäsmävyvä merkkipari $p[i] \neq t[j + i - 1]$ löytyy kohdasta $i < |p|$, tiedetään tätä välittömästi edeltäneiden täsmäysten pohjalta, että $p[i + 1..|p|] = t[j + i..j + |p| - 1]$. Nyt oleellinen havainto on, että jos hahmo siirretään nykykohdasta j sellaiseen kohtaan $j + k$, että $k < |p|$, sisältyy edellämainittu täsmätty kohta $t[j + i..j + |p| - 1] = p[i + 1..|p|]$ osittain tai kokonaan myös seuraavaan tekstistä tutkittavaan alimerkkijonoon $t[j + k..j + k + |p| - 1]$. Tämä tarkoittaa, että kohdassa $j + k$ voi päteä $p = t[j + k..j + |p| - 1]$ vain jos kyseinen aiemmin täsmätty hahmon loppuosa $p[i + 1..|p|]$ täsmää hahmon sen osan kanssa, joka on kohdikkain vastaavan tekstin osan $t[j + i..j + |p| - 1] = p[i + 1..|p|]$ kanssa. Tämä ehto voidaan tutkia helposti, jos ennen varsinaisen merkkijonohaun aloittamista on muodostettu avuksi siirtymätaulukko S , jonka kukin arvo $S[h]$, missä $h = 1, \dots, |p|$, kertoo, mikä on pienin sellainen hahmon eteenpäin siirtämisen pituus, että täsmätyn loppuosan $p[h..|p|]$ ei tiedetä johtavan epätäsmäykseen uudessa kohdassa. Kun ensimmäinen epätäsmävyvä merkkipari $p[i] \neq t[j + i - 1]$ löytyy kohdasta $i < |p|$, voidaan hahmo siirtää seuraavaksi kohtaan $j + S[i + 1]$. Jos epätäsmäys löytyi heti kohdasta $i = |p|$, siirretään hahmoa vain yksi askel kohtaan $j + 1$. Jos taas löytyy hahmon esiintymä (eli ei tapahdu epätäsmäystä), on täsmätty loppuosa $p[1..|p|]$ ja hahmo voidaan siirtää kohtaan $j + S[1]$.

Kuva 2: Siirtymätaulukko S hahmolle $p = \text{"uubu"}$.

h	1	2	3	4
$S[h]$	3	3	3	2

S

	1	2	3	4	5	6	7	8	9	10	11
	u	b	w	u	b	u	u	u	b	u	u
$j=1$	u	u	w	<u>u</u>							
		u	u	w	u						
$j=3$			w	<u>u</u>	<u>b</u>	<u>u</u>					
				u	w	b	u				
					w	u	b	u			
$j=6$						u	u	b	w		
$j=7$							<u>u</u>	<u>u</u>	<u>b</u>	<u>u</u>	
								u	w	b	u

Kuva 3: Täsmätyn loppuosan heuristiikka.

Kuva 3 havainnollistaa täsmätyn loppuosan heuristiikkaa käyttäen samoja merkkijonoja $p = \text{"uubu"}$ ja $t = \text{"ubwubuuubuu"}$ kuin edellä. Hahmolle $p = \text{"uubu"}$ etukäteen laskettu siirtymätaulukko S on esitetty kuvassa 2. Siirtymätaulukon S arvojen määräytyminen kuvataan myöhemmin tarkemmin.

Kuvassa 3 esimerkiksi alkukohdassa $j = 1$ tehdään ensin täsmäys $p[4] = t[4] = \text{"u"}$ ja sitten epätäsmäys $p[3] = \text{"b"} \neq t[3] = \text{"w"}$. Täsmätyt merkit on merkitty alleviivauksella ja epätäsmätyt yliviivauksella. Täsmätty loppuosa oli siis $p[4] = t[4] = \text{"u"}$ ja epätäsmäys tapahtui kohdassa $i = 3$. Hahmo siirretään seuraavaksi $S[i + 1] = S[4] = 2$ askelta eteenpäin kohtaan $j = 1 + S[4] = 1 + 2 = 3$. Kuva 3 havainnollistaa siirtymän määräytymisen periaatetta esittämällä täsmätyn loppuosan ympäröivän suorakulmion, jonka sisällä hahmon merkkien tulee täsmätä myös seuraavassa siirtymäkohdassa. Kuva 3 esittää yli hypätyt epäkelvot hahmon siirtymäkohdat harmaalla värillä, ja niissä epätäsmäviksi tiedetyt merkit on yliviivattu. Esimerkiksi kohta $j = 2$ hypätään yli, koska siinä täsmätyn loppuosan kanssa tehtäisiin epätäsmäys $p[4] = t[4] = \text{"u"} \neq p[3] = \text{"b"}$. Sen sijaan kohtaa $j = 3$ ei hypätä yli, koska siinä täsmätty loppuosa täsmää: $p[4] = t[4] = \text{"u"} = p[2]$.

Edellämainitussa kohdassa $j = 3$ täsmätään loppuosa $p[2..4] = t[4..6] = \text{"ubu"}$, jonka jälkeen kohdassa $i = 1$ tulee epätäsmäys $p[1] = \text{"u"} \neq t[3] = \text{"w"}$. Nyt siirrytään seuraavaksi kohtaan $j = 3 + S[i + 1] = 3 + S[2] = 3 + 3 = 6$. Kohdat $j = 4$ ja $j = 5$ voidaan hypätä yli, koska kohdassa $j = 4$ olisi edessä epätäsmäys $p[2..4] = t[4..6] = \text{"ubu"} \neq p[1..3] = \text{"uub"}$ ja kohdassa $j = 5$ epätäsmäys $p[3..4] = t[5..6] = \text{"bu"} \neq p[1..2] = \text{"uu"}$. Sen sijaan kohtaa $j = 6$ ei hypätä yli, koska siinä täsmää $p[4] = t[6] = \text{"u"} = p[1]$.

Kohdassa $j = 7$ kaikki merkit täsmäävät ja täsmäävä loppuosa on $p[1..4]$. Tämän perusteella hahmo siirrettäisiin seuraavaksi kohtaan $j = 7 + S[1] = 7 + 3 = 10$, joka asettaisi hahmon jo osittain tekstin ulkopuolelle. Näin ollen esimerkin haku lopetetaan kohdan $j = 7$ jälkeen.

Siirtymätaulukon S arvojen määrittäminen

Siirtymätaulukon S arvot voidaan määrittää suoraviivaisesti tutkimalla, miten hahmon kukin loppuosa täsmää hahmon eri siirtymien kanssa suhteessa itseensä. Kuva 4 esittää esimerkkinä aiemmin kuvassa 2 esitettyjen arvojen $S[4]$, $S[3]$, $S[2]$ ja $S[1]$ määrittymisen esimerkkiahmolle $p = \text{"uubu"}$. Yleisesti ottaen arvon $S[h]$ voi määrittää etsimällä askel kerrallaan kokeillen ensimmäisen sellaisen hahmon siirtymän itseensä nähden, jossa kaikki merkit loppuosan $p[h..|p|]$ ja hahmon siirtymän välillä täsmäävät.

Kuva 4: Siirtymätaulukon S arvojen määrittäminen hahmolle $p = \text{"uubu"}$

1	2	3	4			
u	u	b	u			
	u	u	b	u		
		u	u	b	u	

$$S[4] = 2$$

1	2	3	4			
u	u	b	u			
	u	u	b	u		
		u	u	b	u	
			u	u	b	u

$$S[3] = 3$$

1	2	3	4			
u	u	b	u			
	u	u	b	u		
		u	u	b	u	
			u	u	b	u

$$S[2] = 3$$

Esimerkiksi kuvassa 4 arvon $S[4]$ kohdalla on esitetty suorakulmio, joka rajaa loppuosan $p[4..4] = \text{"u"}$. Hahmon alle on asetettu ensin yhden askeleen verran eteenpäin siirretty hahmo, ja tämän kohdalla on tutkittu, täsmääkö siirretyn hahmon suorakulmioon sijoittuva osa

1	2	3	4			
u	u	b	u			
	u	u	b	u		
		u	u	b	u	
			u	u	b	u

$$S[1] = 3$$

loppuosan $p[4..4]$ kanssa. Tässä tapauksessa ei täsmännyt: ensimmäisessä siirtokohdassa suorakulmion sisällä on epätäsmäys $p[4] = \text{"u"} \neq \text{"b"} = p[3]$. Kyseinen rivi on tämän vuoksi esitetty harmaalla: kyseessä olisi epätäsmäykseen johtava siirtymä. Sen alle on lisätty askeleen enemmän eli kaksi askelta siirretty hahmo, ja jälleen on tutkittu täsmäys suorakulmion alueella. Nyt siinä on täsmäys $p[4] = \text{"u"} = p[2]$, jonka perusteella $S[4] = 2$, koska 2 oli pienin siirtymä, joka ei johtanut epätäsmäykseen.

Vastaavasti esimerkiksi arvon $S[1]$ kohdalla on esitetty suorakulmio, joka rajaa loppuosan $p[1..4] = \text{"uubu"}$. Sen alla on esitettynä ensin yhden askeleen siirtymä, joka johtaa suorakulmion alueella epätäsmäykseen $p[2..4] = \text{"ubu"} \neq \text{"uub"} = p[1..3]$. Seuraavaksi on kahden askeleen siirtymä, joka johtaa epätäsmäykseen $p[3..4] = \text{"bu"} \neq \text{"uu"} = p[1..2]$. Lopuksi on kolmen askeleen siirtymä, jonka kohdalla suorakulmiossa täsmää $p[4] = \text{"u"} = p[1]$. Näin ollen $S[1] = 3$, koska 3 oli pienin siirtymä, joka ei johtanut epätäsmäykseen.

On hyvä huomata, että siirtymätaulukon arvot riippuvat huomattavasti hahmon rakenteesta. Yhtenä ääritapauksena voi pitää hahmoa, jonka kaikki merkit ovat keskenään samanlaisia: tällöin $S[h] = 1$ kaikilla indekseillä h . Kuva 5 näyttää esimerkkinä arvon $S[2]$ määräytymisen tällaiselle hahmolle $p = \text{"uuu"}$. Toisena ääritapauksena voi pitää hahmoa, jonka viimeinen merkki ei esiinny missään muualla hahmossa: tällöin $S[h] = |p|$ kaikilla indekseillä h . Kuva 6 näyttää arvon $S[2]$ määräytymisen tällaiselle hahmolle $p = \text{"uub"}$.

Kuva 5: $S[2]$ hahmolle $p = \text{"uuu"}$.

1	2	3				
u	u	u				
	u	u	u			

$$S[2] = 1$$

Kuva 6: $S[2]$ hahmolle $p = \text{"uub"}$.

Kysymykset

Kysymyksissä 3.1 – 3.3 käytetään merkkijonoa $x = \text{"mississippi"}$.

Kysymys 3.1. (1 p.)

Määritä arvo $|x|$.

1	2	3				
u	u	b				
	u	u	b			
		u	u	b		
			u	u	b	

$$S[2] = 3$$

Kysymys 3.2. (1 p.)

Esitä merkkijonot:

a) $x[2..5]$ (0.5 p.)

b) $x[7..9]$ (0.5 p.)

Kysymys 3.3. (1 p.)

Esitä kaikki indeksit j , joille pätee $x[j] = x[j + 3]$.

Oikeita vaihtoehtoja voi olla nolla tai useampia. Valitsemalla oikeita vaihtoehtoja saa pisteitä. Valitsemalla vääriä vaihtoehtoja vähennetään pisteitä. Kysymyksen minimipistemäärä on 0.

Tähän tulee Kysymys 3.3 vastaus

Kysymys 3.4. (3 p.)

Tässä kysymyksessä tarkastellaan täsmätyn loppuosan heuristiikan siirtymätaulukon S arvoja hahmolle $p = \text{"bbububu"}$. Kussakin alikohdassa pitää esittää yksittäisen arvon määräytyminen samassa muodossa kuin kuvassa 4. Vastausruudukkoihin on jo valmiiksi täytetty hahmon p esittävä alkurivi. Huomioi seuraavat asiat vastauksissasi:

- Vastausruudukkoon tulee merkitä loppuosaa rajaava suorakulmio.
- Epätäsmäykseen johtavia siirtymiä vastaavilla riveillä hahmo esitetään harmaalla ja sen epätäsmäviksi tiedetyt merkit yliviivataan.
- Ruudukkoon ei tule tehdä ylimääräisiä merkintöjä: pisteytys pohjautuu siihen, kuinka täsmällisesti vastausruudukkosi on oikein.

Huomaa, että kaksoisklikkaamalla solua käyttöösi tulee pikkueditori. Pikkueditori ei näytä yliviivausta, alleviivausta eikä harmaita kirjaimia. Ne näkyvät vasta kun poistut pikkueditorista painamalla **Esc**-painiketta. Voit syöttää soluun ”b”- ja ”u”-kirjaimia myös työkalupalkin avulla ilman pikkueditoria.

a) Esitä arvon $S[7]$ määräytyminen samassa muodossa kuin kuvassa 4. (1 p.)

Tähän tulee Kysymys 3.4.a vastaus.

$S[7] =$

b) Esitä arvon $S[4]$ määräytyminen samassa muodossa kuin kuvassa 4. (1 p.)

Tähän tulee Kysymys 3.4.b vastaus.

$S[4] =$

c) Esitä arvon $S[1]$ määräytyminen samassa muodossa kuin kuvassa 4. (1 p.)

Tähän tulee Kysymys 3.4.b vastaus.

$S[1] =$

Kysymys 3.5. (3 p.)

Esitä ne askeleet, jotka täsmätyn loppuosan heuristiikka tekee etsiessään hahmoa $p = \text{”bbubu-bu”}$ tekstistä $t = \text{”ubbubbubububuubub”}$. Esitä vastauksesi samassa muodossa kuin kuvassa 3. Vastausruudukkoon on täytetty valmiiksi teksti sekä hahmo kohdassa $j = 1$. Huomioi seuraavat asiat vastauksessasi:

- Hahmo esitetään jokaisessa mahdollisessa kohdassa j (myös yli hypättävissä).
- Kohta j kirjataan ensimmäiseen sarakkeeseen niillä riveillä, joissa hahmoa verrataan tekstiin. Kyseisillä riveillä lisäksi kukin hahmon täsmäävästi verrattu merkki aliviivataan ja epätäsmäävästi verrattu merkki yliviivataan.
 - Hahmon asianmukaiset merkit tulee alle- ja yliviivata myös valmiiksi annetulla kohtaa $j = 1$ vastaavalla rivillä!
- Yli hypättäviä kohtia vastaavilla riveillä:
 - kohtaa j ei kirjata vasempaan sarakkeeseen,
 - hahmo esitetään harmaalla värillä, ja
 - epätäsmääviksi tiedetyt merkit yliviivataan.
- Kuvassa 3 esitettyjen loppuosia rajaavien suorakulmioiden piirtäminen on vapaaehtoista; niiden puuttumisesta tai virheellisyydestä ei menetä pisteitä tässä kysymyksessä.
- Ruudukkoon ei tule tehdä ylimääräisiä merkintöjä: pisteytys pohjautuu siihen, kuinka täsmällisesti vastausruudukkosi on oikein.

Tähän tulee Kysymys 3.5 vastaus.

Kysymys 3.6. (6 p.)

Kysymyksessä on kaksi alikohtaa a) ja b), joista kummassakin on annettu taulukko S . Muodosta kummassakin alikohdassa erikseen jokin sellainen **pelkästään merkkejä** "u" ja "b" sisältävä merkkijono p , jonka täsmätyn loppuosan heuristiikan siirtymätaulukko vastaa kyseisessä alikohdassa annettua taulukkoa S .

a)

h	1	2	3	4	5	6	7	8
$S[h]$	8	8	8	8	3	3	3	1

(3 p.)

b)

h	1	2	3	4	5	6	7	8	9	10	11	12
$S[h]$	11	11	11	11	11	11	5	5	5	5	3	3

(3 p.)

Kysymys 3.7. (5 p.)

Muodosta sellaiset **pelkästään merkkejä** "u" ja "b" sisältävät 5 merkin pituinen hahmo p ja 12 merkin pituinen teksti t , että täsmäävän loppuosan heuristiikka suorittaa hahmon esiintymiä tekstistä etsiessään **täsmälleen** 16 merkkivertailua. Tässä lasketaan vain varsinaisen hakuvaiheen aikana suoritettut merkkivertailut (siirtymätaulukon S määrittämisen aikana tehtyjä vertailuja ei lasketa). Vertailuihin lasketaan sekä täsmäävät että epätäsmäävät vertailut. Esimerkiksi kuvan 3 esittämä haku tekee kaikkiaan 11 merkkivertailua: 2 kohdassa $j = 1$, 4 kohdassa $j = 3$, 1 kohdassa $j = 6$ ja 4 kohdassa $j = 7$.

Jonot

5 merkin pituinen hahmo p : ☐

12 merkin pituinen teksti t : ☐