TKO_8966 Johdatus kieliteknologiaan

18.3.2019. Vastaa kaikkiin kolmeen tehtävään.

1. Selitä lyhyesti seuraavat termit: (10 p)
   a. Morfologinen analyysi ja generointi
   b. Homonymia
   c. Säännölliset lausekkeet
   d. Crawl delay
   e. Tokenisointi
   f. Word2vec
   g. Kielimalli
   h. Nimettyjen entiteettien normalisointi
   i. Stemmaus
   j. Ylisovittaminen

2. Dokumenttien samankaltaisuutta voidaan vertailla tf.idf-painotetuilla vektoriesityksillä. (10 p)
Jokaisen sanan paino voidaan määritellä esimerkiksi seuraavasti:

$$w_{ij} = \begin{cases} (1 + \log tf_{ij}) \cdot (\log idf_i), & tf_{ij} \geq 1 \\ 0, & tf_{ij} = 0 \end{cases}$$

   a. Mitä tf- ja idf-termit tarkoittavat ja mikä niiden merkitys on sanojen painoja laskettaessa?
   b. Miksi kaavassa käytetään logaritmeja?
   c. Miten dokumenttien samankaltaisuus lasketaan näitä painoja käyttäen ja miten tätä hyödynnetään hakukoneissa?
   d. Miten bag-of-words-pohjaiset hakukoneet ottavat huomioon synonymian ja muut sanojen väliset semanttiset relaatiot?
   e. Miten PageRank-algoritmia käyttävät hakukoneet poikkeavat edellä mainitusta?

3. Konekääntäminen, vastaa jokaiseen kohtaan muutamalla virkkeellä (yhteensä 10 p)
   a) Miten määrittelisit konekääntämisen? Miten se eroaa tietokoneavusteisesta kääntämisestä? Mikä on käännösmuisti?
   b) Minkälaista koulutusdataa tarvitaan konekääntämisjärjestelmän opettamiseen? Mitä koulutusdataa on olemassa esimerkiksi Suomi-Englanti kieliparille?
   c) Millä eri tavoilla konekäännösjärjestelmän laatua voidaan arvioida, mitä hyviä ja huonoja puolia näissä arvioinneissa on, ja missä vaiheessa niitä olisi hyvä käyttää?
   d) Yksi konekäännösjärjestelmän arviointiin käytettävä mitta on BLEU. Mihin tämä mitta perustuu ja mitä (dataa) täytyy olla, jotta voimme käyttää tätä arviota? Mitä mahdollisia ongelmia BLEU-mitassa on?
   e) Neuroverkkomalleihin perustuva konekäännösjärjestelmä voi toimia esimerkiksi kirjain- tai sanatasolla (kirjaintasolla: Lukee koko virkkeen 'muistiin' yksi kirjain kerrallaan, generoi tämän perusteella käännöksen yksi kirjain kerrallaa, sanatasolla: lukee koko virkkeen 'muistiin' yksi sana kerrallaan, generoi käännöksen yksi sana kerrallaan). Pohdi, mitä mahdollisia hyviä ja huonoja puolia kummassakin esitystavassa on.