

Tekstinlouhinta...?

Tekstinlouhinnalla tarkoitetaan menetelmiä, joilla rakenteettomasta, luonnollista kieltä sisältävästä tekstistä saadaan kerättyä ja analysoitua dataa. Tekstinlouhinnan kohteena voi olla erilaisia datajoukkoja, kuten sähköposteja, uutisartikkeleita tai sosiaalisen median kirjoituksia ja sillä voidaan hakea yhteyksiä tai kaavoja tekstistä. Luonnollisen kielen koneellinen analysointi voi olla haastavaa, koska tietokoneiden toiminta perustuu rakenteelliseen dataan, mutta luonnolliset kielet ovat rakenteetonta dataa. Lisäksi luonnollisten kielten syntaksi voi olla hyvin epäjohdonmukaista verrattuna ohjelmointikielten kielioppiin.

Tekstinlouhintaprosessi on pitkä ja kattava prosessi, joka pitää sisällään monta toisiinsa liittyvää vaihetta. Yhdistämällä vaiheet saadaan luotua yhtenäinen prosessijono, jolla suoritetaan itse tekstinlouhintaa kokonaisuutena. Tekstinlouhintaprosessin ensimmäinen vaihe on itse käsiteltävän ja analysoitavan materiaalin keruu tai haku (information retrieval). Haku voidaan suorittaa esimerkiksi hakusanoilla, jolloin järjestelmä hakee tietyt dokumentit jostakin suuremmasta kokoelmasta muuta käsittelyä varten. Tämä voi nopeuttaa tekstin analysoimista huomattavasti, jos dokumentteja on paljon ja itse haluttua tietoa sisältäviä tekstejä on suhteellisen vähän.

Tekstinlouhinnan tärkeimpiä ja vaikeimpia osa-alueita on luonnollisten kielten prosessointi (natural language processing). Sen avulla tietokoneet saadaan ymmärtämään ihmisten käyttämiä luonnollisia kieliä. Tämä on mahdollista lauseenjäsennys- ja sanaluokanmerkitsemis-algoritmien (part-of-speech tagging) avulla. Algoritmit siis tunnistavat ja merkitsevät tekstistä sanaluokkia ja lauseenosia, jotta myöhemmissä vaiheissa käytetyt työkalut pystyvät toimimaan prosessin antaman datan pohjalta.

Tiedon keruuta (information extraction) käytetään muuttamaan automaattisesti rakenteetonta dataa rakenteelliseksi. Esimerkkejä datan muuttamisesta rakenteelliseksi on termien analyysi, nimellisten entiteettien tunnistaminen ja faktojen keruu. Työkalut voivat annettujen mallien perusteella löytää tekstistä suhteita tiettyjen asioiden välillä ja kerätä niitä rakenteelliseen muotoon ja tallentaa tietokantoihin.

Viimeisessä vaiheessa käytetään tiedonlouhinnan menetelmiä (data mining), jotta aikaisemmassa vaiheessa saadun tietokannan datajoukosta pystyttäisiin tunnistamaan toistuvia kaavoja ja yhteneväisyyksiä. Lopuksi tiedonlouhinnan avulla saatu uusi tieto tallennetaan omaan tietokantaansa ja se voidaan julkaista käyttöön.

Tekstinlouhinnalla on useita käyttökohteita monilla eri aloilla. Tekstiä louhitaan

pääasiallisesti analysoimista varten, josta saaduista tuloksista on erinäisiä hyötyjä alasta riippuen. Eniten hyötyä ja tarvetta tekstinlouhinnalle on suurten tekstien analysoimisen apuna, sillä manuaalisesti siinä menisi hyvin pitkä aika, voi kone tehdä sen paljon nopeammin ja säästää aikaa ja varoja. Tekstiä louhimalla voidaan sen sisältöä tilastoida, sieltä voidaan hakea yhteyksiä ja tutkia missä yhteydessä tietyt sanat esiintyvät. Esimerkiksi nimiä voidaan yhdistää tiettyihin titteleihin tai useamman merkityksen omaavan sanan tarkoitus voidaan tunnistaa kontekstin avulla.

Tekstinlouhintaa voidaan käyttää apuna muunmuassa roskapostisuodatuksessa analysoimalla viestien sisältöä ja vertaamalla niitä roskaposteiksi tunnistettujen viestien sisältöön, josta voidaan päätellä onko viesti roskapostia vai ei. Tekstinlouhintaa hyödynnetään myös turvallisuuden takaamiseksi seuraamalla Internetiin kirjoitettuja tekstejä, kuten uutisia, blogeja sekä foorumiviestejä ja tutkimalla niiden sisältöä. Nykyään sosiaalisen median merkityksen kasvaessa on yrityksille avuksi, että senkin sisältöä on mahdollista analysoida tekstinlouhintaan tarkoitetuilla työkaluilla. Etenkin markkinoinnille on hyötyä kun voidaan analysoida suuri määrä ihmisten sosiaalisessa mediassa käymiä keskusteluja.

Koska tekstinlouhintaa suoritetaan tietokoneilla on siihen luonnolisesti useita valmiita työkaluja. Osa näistä työkaluista on maksullisia ja osa ilmaisia. Työkaluilla on myös omat käyttökohteensa ja niiden ominaisuudet eroavat käyttötarkoituksesta riippuen. Yksi uusimmista ja yrityksille merkittävimmistä tekstinlouhintakohteista ovat sosiaalisessa mediassa olevat kirjoitukset.

Sysomos:in Media Analysis Platform on yksi kaupallinen sosiaalisen median tekstinlouhintatyökalu. Sen avulla käyttäjän on mahdollista analysoida miljardeja sosiaalisen median viestejä. Järjestelmä kykenee keräämään myös tietoja keskustelevista henkilöistä, kuten iän, sukupuolen ja ammatin. Sen on myös mahdollista päätellä, onko mielipiteiden sävy positiivinen, negatiivinen vai neutraali, sekä tunnistaa keskustelun aktiivisin osapuoli. Myös keskustelujen määrän ja muiden ominaisuuksien mittaaminen onnistuu ohjelmalta. Siitä voi olla valtavaa hyötyä markkinoinnille, johtuen sen toimittamasta tiedosta, joka tulee suoraan kuluttajilta ilman minkäänlaisia kuluttajakyselyitä tai lomakkeita.

Abzoobalta on saatavilla XPRESSOMeter, joka tarkkailee sosiaalista mediaa reaaliajassa. Ohjelma kykenee päättelemään kommenttien sävyn, ilmoittamaan käyttäjää koskevista keskusteluista niiden ollessa käynnissä ja myös osoittamaan kohteita, jotka vaativat käyttäjän huomiota. Ohjelma nopeuttaa huomattavasti yrityksen reagointikykyä sosiaalisessa mediassa, joka on hyödyksi, kun pyritään olemaan ajan hermolla.