Tampereen yliopisto

Sirkku Seitamäki

SAMANKALTAISTEN TEKSTIEN EHDOTTAMINEN

Informaatioteknologian ja viestinnän tiedekunta
Pro gradu -tutkielma
Toukokuu 2022

TIIVISTELMÄ

Sirkku Seitamäki: Samankaltaisten tekstien ehdottaminen
Pro gradu -tutkielma
Tampereen yliopisto
Tietojenkäsittelytieteiden tutkinto-ohjelma
Toukokuu 2022

Samankaltaisten tekstien löytäminen antaa mahdollisuuden tarjota käyttäjälle samasta aiheesta lisää luettavaa. Samankaltaisuuksia etsitään vertailemalla tekstejä, mitä varten tekstejä on käsiteltävä. Työn tarkoituksena on toteuttaa ohjelma ja testata sen avulla tekstin esikäsittelyn eri tekniikoiden vaikutusta samankaltaisten tekstien tunnistamiseen. Työssä käsitellyt tekstit ovat suomenkielisiä, mikä tarkoittaa, että suomen kielen monet taivutusmuodot on otettava huomioon. Tiedonhakua varten teksti esikäsitellään poistamalla välimerkkejä, muuntamalla pienaakkosiin, tokenisoimalla, karsimalla sulkusanat ja normalisoimalla, minkä voi tehdä stemmauksen tai lemmauksen avulla. Lemmausta pidetään suomen kielessä parempana vaihtoehtona kuin stemmausta.

Toteutin Javalla pienen fintextrec-nimisen ohjelman, joka voi joko esikäsitellä alkuperäiset tekstit ja tallentaa sitten käsitellyt tekstit MySQL-tietokantaan tai indeksoida alkuperäiset tekstit Solrin indeksiin. Ohjelmalla voi hakea ehdotuksia joko MySQL-tietokannasta tai Solrin indeksistä. Toteutuksen avulla vertailen kahdeksaa eri vaihtoehtoa, joista kaksi käyttää Solrin MoreLikeThis-kyselyä ja kuusi käyttää MySQL:n InnoDB:n kokotekstihakua. Testauksessa käyttämäni data koostuu reilusta sadasta vastaustekstistä, joiden aiheena on suomalaisten käsitykset kulttuuriperinnöstään.

Toteutusta on testattu hakemalla neljälle eri vastaustekstille ehdotuksia. Vaihtoehtojen, joissa lemmataan, ehdotuksissa on enemmän samoja tunnisteita kuin vaihtoehdoissa, joissa stemmataan tai joissa sanat ovat alkuperäisissä taivutusmuodoissaan. Taivutusmuotovaihtoehdon tai stemmausvaihtoehdon paras pari on toinen taivutusmuotovaihtoehto tai stemmausvaihtoehto. Koska lemmausvaihtoehtoja on useampia niin niiden kohdalla tuli esille, että sulkusanojen karsinnalla on jonkun verran merkitystä. Paras pari lemmausvaihtoehdolle on toinen lemmausvaihtoehto, jossa sulkusanat käsitellään samalla tavalla.

Avainsanat: tekstin esikäsittely, lemmaus, perusmuotoistaminen, stemmaus, sulkusanat, kokotekstihaku, MoreLikeThis

Sisällys

1 Johdanto

Jos tekstimassasta pystytään tunnistamaan samankaltaisia tekstejä, voidaan niitä ehdottaa käyttäjälle lisätietona samasta aiheesta. Samankaltaisuuksien löytämiseksi tekstejä on vertailtava. Kun tekstimassa on luonnollista kieltä ilman sisällön luokittelua tai avainsanoja, niin tekstiä on ensin käsiteltävä. Tekstistä on eroteltava sanat, jotta niitä voidaan analysoida. Sanojen vertailua voi tehostaa poistamalla sanoista päätteet. Kaikilla sanoilla ei ole yhtä suuri painoarvo tekstin merkityksen kannalta, joten tämän huomioiminen voi myös tehostaa samankaltaisuuksien etsimistä.

Tutkielman tarkoituksena on toteuttaa ohjelma ja testata sen avulla tekstin esikäsittelyn eri tekniikoiden vaikutusta samankaltaisten tekstien tunnistamiseen. Koska työssä käsitellyt tekstit ovat suomenkielisiä, oli alkuperäinen idea perusmuotoistaa tekstissä olevat sanat ja tallentaa ne yksittäin tietokantaan. Tähän perusmuotojen hakemistoon kohdistaisin kyselyitä samankaltaisten tekstien löytämiseksi. Jalostin ideaa pidemmälle ja otin käyttöön MySQL:n InnoDB:n kokoteksti-indeksin tietokannan yksinkertaistamiseksi ja kokotekstihaun kyselyn nopeuttamiseksi. Vertaan kokotekstihakua Solrin MoreLikeThis-erikoishakuun, joka hakee lähdedokumentille (source document) samankaltaisia dokumentteja. [Oracle Corporation. 2022, Apache Software Foundation 2021d]

Toteutin Javalla pienen fintextrec-nimisen ohjelman, joka voi joko esikäsitellä alkuperäiset tekstit ja tallentaa sitten käsitellyt tekstit MySQL-tietokantaan tai lisätä alkuperäiset tekstit Solriin. Ohjelmalla voi hakea ehdotuksia joko MySQL-tietokannasta tai Solrista.

Toteutuksen avulla vertailen kahdeksaa vaihtoehtoa, joista kuusi on MySQL:n InnoDB:n kokotekstihakuja (FT) ja kaksi on Solrin MoreLikeThis-hakuja (MLT). Kokotekstihakuvaihtoehtoja on useampia, jotta voin kokeilla eri yhdistelmiä tekstin esikäsittelyssä. Kaikissa FT-vaihtoehdoissa poistetaan tekstistä tietyt välimerkit, muutetaan pienaakkosiin ja tokenisoidaan. Muut käsittelytavat riippuvat vaihtoehdosta. Kahdessa ensimmäisessä FT-vaihtoehdossa alkuperäinen teksti perusmuotoistetaan ja vain perusmuodot otetaan talteen. Toisesta sulkusanat karsitaan esikäsittelyssä, mutta toisessa niitä ei karsita. Kolmannessa ja neljännessä FT-vaihtoehdossa alkuperäinen teksti perusmuotoistetaan ja jos sanalle ei ole löytynyt perusmuotoa, otetaan sana talteen alkuperäisessä muodossaan eli mahdollisesti taivutusmuodossa. Toisessa näistä vaihtoehdoista sulkusanat karsitaan esikäsittelyssä ja toisessa niitä ei karsita. Viidennessä ja kuudennessa FT-vaihtoehdossa tekstiä ei perusmuotoisteta, vaan se tallennetaan taivutusmuotoineen. Toisessa näistä sulkusanat karsitaan esikäsittelyssä, mutta toisessa

ei. Molemmissa MLT-vaihtoehtoissa Solr stemmaa sanat. Toisessa vaihtoehdossa sulkusanat suodatetaan pois, mutta toisessa ei.

Tarkoituksena on tutkia millaisia ehdotuksia eri vaihtoehdot saavat ja vaikuttavatko eri tekstin esikäsittelytavat ehdotuksiin. Sekä eroaako MySQL:n kokotekstihaun tulokset Solrin MoreLikeThis-kyselyn tuloksista.

Testaan toteutusta kvalitatiivisella aineistolla, jonka on kerännyt Museovirasto ja Suomen Kotiseutuliitto [2015]. Testidataksi valikoitui yhden kysymyksen vastaukset eli 119 lyhyehköä, suomenkielistä vastaustekstiä. Aiheena on suomalaisten ajatukset kulttuuriperinnöstä.

Tutkielman alussa käyn läpi muutamien kielitieteellisten käsitteiden määrittelyjä ja suomen kielen erityispiirteitä luvussa 2 sekä viiden tiedonhakuun liittyvän käsitteen määrittelyä luvussa 3. Seuraavaksi on vuorossa luonnollisen kielen käsittely tiedonhakua varten luvussa 4. Tämä sisältää erilaisia tapoja esikäsitellä tekstiä. Luvuissa 5-6 kerron MySQL:n ja Solrin ominaisuuksista, joita ohjelmani käyttää. Toteutus-osio alkaa ohjelman ja sen toimintojen esittelyllä luvussa 7. Sitten kerron aineistosta ja esittelen testattavat vaihtoehdot luvussa 8. Näiden jälkeen käyn läpi testeissä saamani tulokset luvussa 9. Lopussa on vielä yhteenveto.

2 Luonnollisesta kielestä

Karlssonin [2009] mukaan luonnollinen kieli on kieli, joka on kehittynyt hyvin pitkän ajanjakson aikana. Ihmiset käyttävät luonnollista kieltä lapsesta saakka kommunikointivälineenä. Luonnollisen kielen, kuten suomen, sanoilla voi olla useita merkityksiä sekä sanaston on tarkoitus joustaa ympärillä muuttuvan maailman mukana. [Karlsson 2009]

Seuraavaksi käyn läpi muutamien kielitieteellisten käsitteiden määrittelyjä sekä kerron suomen kielen erityispiirteistä, jotka vaikuttavat tekstiin perustuvaan tietojenhakuun.

2.1 Kielitieteellisiä käsitteitä

Esittelen neljä aihealueen kielitieteellisistä käsitettä: sana, sane, perusmuoto ja vartalo. Kielenpuhuja pystyy helposti antamaan esimerkkejä osaamansa kielen sanoista. Kielitoimiston sanakirja [2021b] antaa sanalle yhdeksi määritelmäksi seuraavan:

> ”kielen pienin itsenäinen (yhdeksi kirjainjaksoksi kirjoitettava) merkityssisältöinen rakenneosa; sen puhe- t. tekstiyhteydessä esiintyvä muoto, sane.”

Sanan määritelmässä mainitaan sane. Sane on Kielitoimiston sanakirjan [2021c] mukaan sana, joka esiintyy puheessa tai tekstissä joko perusmuodossa tai taivutusmuodossa. Tässä tutkielmassa käytän pääsiallisesti tutumpaa termiä sana.

Kielitoimiston sanakirjan [2021a] mukaan perusmuodon kielitieteellinen määritelmä on ”muoto joka on taivutuksessa lähtökohtana, esim. nomineilla yksikön nominatiivi ja verbeillä ensimmäinen infinitiivi”. Esimerkkejä nomineista perusmuodoissaan ovat *auto*, *keltainen* ja *yhdeksän* ja verbeistä perusmuodoissaan on *tehdä* ja *laulaa*.

Sanan vartalo on Kielitoimiston sanakirjan [2021d] mukaan se, mitä sanasta jää, kun päätteet ja liitteet on erotettu. Snowball [2022a] antaa esimerkkisanoille *yhdeksän* ja *laulaa* vartaloiksi *yhdeks* ja *laula*.

2.2 Suomen kielen erityispiirteitä

Suomen kielessä on runsaasti sijamuotoja verrattuna germaanisiin kieliin kuten englanti. Tämä vaikuttaa tekstin käsittelyyn, joka tehdään tiedon hakemista varten. Esimerkkinä monista sijamuodoista on substantiivin *teko* 140 taivutusmuotoa taulukossa 1 [Hakulinen et al. 2008].

	Yksikkö	**Monikko**
NOM	teko tekoni tekosi tekonsa tekomme tekonne	teot tekoni tekosi tekonsa tekomme tekonne
GEN	teon tekoni tekosi tekonsa tekomme tekonne	tekojen tekojeni tekojesi tekojensa tekojemme tekojenne
PAR	tekoa tekoani tekoasi tekoaan tekoamme tekoanne	tekoja tekojani tekojasi tekojaan tekojamme tekojanne
ESS	tekona tekonani tekonasi tekonaan tekonamme tekonanne	tekoina tekoinani tekoinasi tekoinaan tekoinamme tekoinanne
TRA	teoksi teokseni teoksesi teokseen teoksemme teoksenne	teoiksi teoikseni teoiksesi teoikseen teoiksemme teoiksenne
INE	teossa teossani teossasi teossaan teossamme teossanne	teoissa teoissani teoissasi teoissaan teoissamme teoissanne
ELA	teosta teostani teostasi teostaan teostamme teostanne	teoista teoistani teoistasi teoistaan teoistamme teoistanne
ILL	tekoon tekooni tekoosi tekoonsa tekoomme tekoonne	tekoihin tekoihini tekoihisi tekoihinsa tekoihimme tekoihinne
ADE	teolla teollani teollasi teollaan teollamme teollanne	teoilla teoillani teoillasi teoillaan teoillamme teoillanne
ABL	teolta teoltani teoltasi teoltaan teoltamme teoltanne	teoilta teoiltani teoiltasi teoiltaan teoiltamme teoiltanne
ALL	teolle teolleni teollesi teolleen teollemme teollenne	teoille teoilleni teoillesi teoilleen teoillemme teoillenne
ABE	teotta	teoitta
KOM		tekoineni tekoinesi tekoineen tekoinemme tekoinenne
INS		teoin

Taulukko 1. Substantiivin *teko* taivutusmuotoja [Hakulinen et al. 2008]

Taivutusmuotoon vaikuttaa sija, joita substantiivilla on 14. Luettelen ne seuraavaksi siinä järjestyksessä kuin ne ovat taulukossa 1: nominatiivi, genetiivi, partitiivi, essiivi, translatiivi, inessiivi, elatiivi, illatiivi, adessiivi, ablatiivi, allatiivi, komitatiivi ja instruktiivi. Kahteentoista sijaan voi liittyä possessiivisuffiksi, esim. *teko+si*. Yksikkö- ja monikkomuoto voivat olla samat, kuten nominatiivin *teko+ni* tai erilaiset, kuten translatiivin *teo+ksi* ja *teo+i+ksi*. Myös sanan vartalo voi vaihdella taivutusmuodoissa, esim. nominatiivin yksikkömuoto *teko* ja monikkomuoto *teo+t*. [Hakulinen et al. 2008]

Suomen kielen käsittely on siis hankalaa monien taivutusmuotojen vuoksi. Taulukon 1 esimerkkisana on nomini, mutta myös verbit taipuvat omien sääntöjensä.

3 Tiedonhaun käsitteitä

Esittelen tässä luvussa lyhyesti seuraavat tiedonhaun käsitteet: relevanssi, saanti, tarkkuus, käänteistiedosto ja TF-IDF (term frequency-inverse document frequency).

3.1 Relevanssi, saanti ja tarkkuus

Kun käyttäjä hakee tietoa jostain aiheesta, niin hänen pyrkimyksenänsä on löytää relevanttia tietoa kyseisestä aiheesta. Järvelin ja Sormunen [2010] esittelevät relevanssin kaksi pääsuuntaa, aiherelevanssin ja käyttäjärelevanssi. Aiherelevanssi tarkoittaa, että dokumentti käsittelee haussa määriteltyä aihetta. Sen voivat arvioida ihmiset, jotka tuntevat kyseisen aiheen. Dokumentin voi ajatella olevan relevantti aiheen kannalta, jos dokumentissa käsitellään aihetta edes jonkun verran. Käyttäjärelevanssi ottaa huomioon aiheen lisäksi käyttäjän, johon voi liittyä monia asioita, esim. käyttäjän esitiedot aiheesta, haun tavoitteet ja odotukset tai arvio tulosten hyödyllisyydestä.

Hakutulosten arviointikriteereistä yleisimmin käytettyjä ovat saanti ja tarkkuus. Saanti on osumien suhde kaikkiin relevantteihin dokumentteihin eli kuinka suuren osan relevanteista dokumenteista haku löysi. Saannin yhteydessä voi olla haastavaa saada selville kuinka paljon relevantteja dokumentteja on. Tarkkuus on vuorostaan osumien suhde kaikkiin löydettyihin dokumentteihin eli kuinka suuri osuus tuloksista on relevantteja. [Järvelin ja Sormunen 2010]

3.2 Käänteisindeksi

Käänteisindeksiä (inverted index) kutsutaan myös käänteishakemistoksi tai käänteistiedostoksi (inverted file). Alaterä ja Halttunen [2002] esittelevät peräkkäistiedoston ja käänteistiedoston eroja. Peräkkäistiedosto (tai peräkkäisrakenne) sisältää kaiken mitä tietueesta on tallennettu eli sinne on talletettu varsinaiset dokumentit. Tietueet on järjestetty tietuenumeron mukaan. Tavallinen taulu tietokannassa, kuten toteutuksessani oleva response-taulu, joka sisältää alkuperäiset vastaustekstit, on esimerkki peräkkäistiedostosta. Käänteistiedosto (tai käänteisrakenne) vuorostaan sisältää hakutermit ja osoitteet (esim. tietuenumerot), jotka viittaavat peräkkäistiedostoon tallennettuihin dokumentteihin. Käänteistiedoston etu peräkkäistiedostoon verrattuna on nopeampi haku. Tämän vuoksi tekstiä sisältäviin kenttiin kohdistuvassa tiedonhaussa käytetään käänteistiedostoja.

Zobel ja Moffat [2006] tutkivat hakukoneiden tekstihakua. He kuvailevat käänteisindeksin olevan kokoelma listoja. Jokaisella hakuavaimella on oma lista, joka sisältää vähintään dokumenttien tunnisteet, missä kyseinen hakuavain esiintyy. Dokumenttitason käänteishakemisto voi sisältää dokumenttien lukumäärän, missä hakuavain esiintyy sekä käänteisen listan, jossa on parina dokumentin tunniste ja hakuavaimen esiintymistiheys dokumentissa. Seuraava esimerkki on Zobelin ja Moffatin

[2006] artikkelista. Termin *house* kohdalla dokumenttien lukumäärä on kaksi esimerkkidatassa. Termiin liittyvän käänteisen listan sisältö koostuu pareista ⟨x, y⟩, missä x on dokumentin tunniste ja y on lukumäärä. Esimerkin *house* käänteisen listan sisältö on ⟨2, 1⟩, ⟨3, 1⟩, koska *house* esiintyy kerran dokumentissa 2 ja dokumentissa 3.

3.3 TF-IDF

TF-IDF tarkoittaa termifrekvenssiä ja käänteistä dokumenttifrekvenssiä. Termifrekvenssi kertoo, kuinka monta kertaa sana esiintyy dokumentissa ja käänteinen dokumenttifrekvenssi kertoo, kuinka harvoin sana esiintyy kokoelmassa (eli kaikissa dokumenteissa) [Oracle Corporation 2022]. Järvelinin ja Sormusen [2010] mukaan sanalle, joka esiintyy dokumentissa, voidaan laskea paino kertomalla termifrekvenssi käänteisellä dokumenttifrekvenssillä.

MySQL laskee relevanssiarvon TF-IDF:n avulla. Koodikatkelmassa 1 lasketaan käänteinen dokumenttifrekvenssi (IDF). [Oracle Corporation 2022]

```
${IDF} = log10( ${total_records} / ${matching_records} )
```

Koodikatkelma 1. IDF:n laskeminen MySQL:ssa [Oracle Corporation 2022]

Ensin jaetaan dokumenttien kokonaismäärä (total records) niiden dokumenttien määrällä, missä sana esiintyy (matching records) ja otetaan 10-kantainen logaritmi osamäärästä. Käänteinen dokumenttifrekvenssi kerrotaan seuraavaksi sanan esiintymien määrällä dokumentissa (TF), jotta saadaan paino. Koodikatkelmassa 2 lasketaan relevanssiarvo tulosten sijoitusvertailua varten [Oracle Corporation 2022].

```
${rank} = ${TF} * ${IDF} * ${IDF}
```

Koodikatkelma 2. Sijoituksen laskeminen MySQL:ssa [Oracle Corporation 2022]

Jos haetaan useilla sanoilla, niin MYSQL laskee joka sanalle koodikatkelman 2 mukainen relevanssiarvon, jotka lasketaan sitten yhteen. [Oracle Corporation 2022]

4 Tekstin esikäsittely tiedonhakua varten

Tekstin esikäsittely mainitaan usein tiedonlouhinnan yhteydessä. Petrović ja Stanković [2019] tutkivat tekstin esikäsittelyn merkitystä serbiankielisten lakitekstien louhinnassa. Heidän käyttämiin tekstin esikäsittelyn tekniikoihin kuuluvat tekstin siirtäminen PDF-, DOC- ja DOCX-formaateista TXT-formaattiin, välimerkkien poisto tekstistä, tokenisointi, pienaakkosiin muuntaminen, sulkusanojen poisto ja stemmaus. Esittelen ensimmäisessä alaluvussa tekstin esikäsittelytapoja, joita käytän toteutuksessani.

Suomen kielen monet taivutusmuodot haittaavat sanojen vertailua, joka on tehokkaampaa, jos sanat ovat samassa muodossa eli ilman päätteitä. Jos esimerkkinä on substantiivi, kuten *koira*, niin tällä sanalla voi olla päätteitä kuten monikon pääte *koira+t*, adessiivin pääte *koira+lla* tai molemmat päätteet yhdessä *koir+i+lla*. Päätteiden poistamisen voi tehdä normalisoimalla tekstin. Esittelen toisessa ja kolmannessa alaluvussa normalisointitekniikoista stemmauksen ja lemmauksen. Neljännessä ja viidennessä alaluvussa vertailen niiden soveltuvuutta suomenkieliseen tekstiin. Kuudennessa alaluvussa käsittelen sulkusanojen karsintaa.

4.1 Tokenisointi, välimerkkien poistaminen ja pienaakkosiin muuttaminen

Tekstin tokenisointi tarkoittaa tekstin pilkkomista pienempiin osiin, yleensä sanoihin. NLTK (Natural Language Toolkit) Projectin [2022] englanninkielisissä esimerkeissä teksti eli merkkijono on pääasiassa katkaistu aina tyhjän välin kohdalta. Poikkeuksena ovat lyhennetyt muodot, esim. *couldn't*, jotka NLTK on pilkkonut kahdeksi tokeniksi: *could* ja *n't*. Ahosen [2019] python-kielistä esimerkkiä mukaillen suoritan tokenisoinnin omassa toteutuksessani yksinkertaisimmalla tavalla eli teksti pilkotaan tyhjien välien kohdalta.

Teksti tokenisoidaan ja tekstistä poistetaan välimerkkejä ja erikoismerkkejä, jotta sanojen tunnistaminen on mahdollista. Petrović ja Stanković [2019] poistavat datastaan kaikki välimerkit, jopa yhdysmerkit, jotka johtavat yhdyssanojen katkeamiseen. He mainitsevat, että välimerkkien poistamiseen vaikuttaa se, mitä ollaan tutkimassa. Omassa toteutuksessani poistan tekstistä määrittelyssä 1 olevat merkit. Näistä kaikki eivät esiinny käyttämässäni testidatassa.

. , : ; ! ? " ' ( ) [ ] { } / \

Määrittely 1. Välimerkit, jotka fintextrec-ohjelma poistaa tekstistä

En poista yhdysmerkkiä (-), jotta se säilyy yhdyssanoissa (esim. *ei-materiaalinen*) Voikon analyysiä varten.

Tekstissä yleensä esiintyy niin isoja kuin pieniä kirjaimia. Tekstin esikäsittelyssä voidaan muuntaa kaikki kirjaimet pieniksi. Ohjelmani muuttaa kokotekstihakuvaihtoehtojen sanat

pienaakkosiin. MoreLikeThis-vaihtoehtojen kaavioissa on määritelty käytettäväksi suodatinta, joka muuntaa sanat pienaakkosiin.

4.2 Stemmaus

Stemmauksessa sana typistetään sen vartaloon. Koreniuksen ja muiden [2004] mukaan stemmaus on käytetyin normalisointitekniikoista tiedonhaussa. Normalisoinnissa kielellä on merkitystä. Otetaan esimerkkisanoiksi suomen kielestä *kissa* ja *kissat* ja englannin kielestä vastaavasti *cat* ja *cats.* Stemmasin esimerkkisanat Snowball-sivuston [2022a] Demo-ohjelmalla, joka käyttää Snowball-stemmausalgoritmia. *Kissa* ja *kissat* typistyvät samaan vartaloon *kis. Cat* ja *cats* typistyvät vartaloon *cat.* Suomen kielen monet taivutusmuodot ja astevaihtelu aiheuttavat kuitenkin ongelmia stemmauksessa. *Kissojen,* jossa on monikon ja genetiivin päätteet, typistyy muotoon *kiso*, joka on eri kuin *kis*, mihin *kissa* ja *kissat* typistyvät. Esimerkki astevaihtelusta on *mato* ja sen genetiivimuoto *madon*, joiden vartalot ovat *mato* ja *mado*.

Esittelen seuraavaksi lyhyesti Snowball-stemmerin. Snowball on kieli stemmausalgoritmeja varten [Snowball 2022c]. Porter [2001] toteaa, että Snowball kehitettiin, koska aiemman version eli Porter-stemmerin käytössä oli ongelmia ja aiempaa versiota ei voinut käyttää muissa kielissä kuin englannissa. Snowball-stemmereitä on luotu englannin kielen lisäksi monille eri kielille, esim. ranska, saksa, venäjä ja suomi. [Snowball 2022d].

Snowball [2022b] esittelee mitä suomen kielen stemmausalgoritmi tekee eri vaiheissa, kun se käy sanan läpi. Vaiheisiin liittyy useita sääntöjä, joiden perusteella stemmeri typistää sanaa, jos se on mahdollista. Esimerkkinä on sana *koirissammekin*, koska se sisältää useita päätteitä. Ensin stemmeri katsoo, onko sanassa partikkelia (esim. *koirissamme+kin*) Jos on, niin kyseinen pääte poistetaan. Toisessa vaiheessa poistetaan genetiivin pääte (esim. *koirissa+mme*). Kolmannessa vaiheessa poistetaan sijamuoto (esim. *koiri+ssa*). Neljännessä vaiheessa poistetaan muut päätteet, kuten *-mpi*. Viidennessä vaiheessa poistetaan monikon pääte (esim. *koir+i*). Kuudennessa vaiheessa siistitään sanan vartaloa. Snowball-stemmausalgoritmi antaa esimerkkisanan vartaloksi *koir*. [Snowball 2022b]

4.3 Lemmaus

Lemmauksessa tai perusmuotoistamisessa haetaan sanan lemma tai perusmuoto, esim. taivutusmuotojen *kissat* ja *kissoille* perusmuoto on *kissa.* Sanan kaikilla taivutusmuodoilla on yhteinen perusmuoto, mikä tekee lemmauksesta hyödyllistä kielissä, joissa sanat taipuvat paljon. Lemmauksen haittana on se, että lemmauksen suorittavan ohjelmalla tulee olla laaja sanasto, jotta suurimmalle osasta sanoista löydetään perusmuoto. Lisäksi yhdyssanat voivat olla ongelmallisia lemmauksessa. Tämä voidaan

välttää, jos yhdyssanat pilkotaan erillisiksi sanoiksi, jolloin ne voidaan tallentaa yhdyssanana sekä erillisinä sanoina. [Korenius et al. 2004]

Suomen kielessä perusmuotoistamisen voi tehdä Voikko-ohjelmiston avulla. Sitä voi käyttää morfologiseen analyysiin, oikeinkirjoituksen ja kieliopin tarkastukseen. Se on alkujaan kehitetty tekstin oikolukuun OpenOffice- ja LibreOffice-ohjelmia varten. [Voikko 2021]

Liitteessä 1 on Voikon suorittaman morfologisen analyysin tulokset esimerkkitekstille. Tekstin tietokantatunniste on 92 ja se on testauksessa käyttämästäni datasta [Museovirasto ja Suomen Kotiseutuliitto 2015]. Käyn seuraavaksi tarkemmin läpi yhden esimerkkitekstissä olevan sanan morfologisen analyysin tuloksen. Esimerkissä 1 on Voikon analyysi sanalle *käsitöitä*.

```
[{NUMBER=plural,
STRUCTURE==pppp=ppppp,
BASEFORM=käsityö,
SIJAMUOTO=osanto,
CLASS=nimisana,
FSTOUTPUT=[Ln][Xp]käsi[X]kä[Sn][Ny]si[Bh][Bc][Ln][Xp]työ[X]
tö[Sp][Nm]itä,
WORDBASES=+käsi(käsi)+työ(työ)}]
```

Esimerkki 1. Morfologisen analyysin tulos sanalle *käsitöitä*

Esimerkin 1 analyysin tulos sisältää seitsemän attribuuttia. NUMBER-attribuutti viittaa kielitieteelliseen lukuun, joka on esimerkissä 1 monikko. STRUCTURE-attribuutti kuvaa rakenteen, mistä näkee, että kyseessä on yhdyssana. Yhtäsuuruusmerkki kertoo mistä morfeemi alkaa. Morfeemi on kielen pienin merkityksen sisältävä yksikkö. Kirjain p tarkoittaa, että kyseessä on pienaakkosilla kirjoitettu kirjain. BASEFORM-attribuutin arvo on sanan perusmuoto. SIJAMUOTO-attribuutti viittaa nimensä mukaisesti sanan sijamuotoon. CLASS-attribuutti tarkoittaa sanaluokkaa. FSTOUTPUT-attribuutin nimessä oleva FST (finite state transducers) viittaa äärellistilallisiin automaatteihin [Tieteen termipankki 2013]. Kentän arvo on automaatin suorittamia analyysejä varten. WORDBASES-attribuutti sisältää sanan osien perusmuodot. [Corevoikko 2017]

4.4 Stemmaus ja lemmaus suomen kielessä

Vertailen stemmauksen ja lemmauksen tuloksia kahdella eri sanalla, jotka esiintyvät eri taivutusmuodossa. Tein stemmauksen sanoille Snowball-sivustolla [2022a], jossa on käytössä Snowball-stemmausalgoritmi. Perusmuotoistin sanat Oikofix-palvelussa [2022a]. Oikofixin [2022b] suomenkielinen osuus pohjautuu Voikkoon. Tulokset ovat taulukossa 2.

Sana ennen käsittelyä	Sanan vartalo	Perusmuoto
matkustan	matkust	matkustaa
matkustivat	matkustiv	matkustaa
matkustettiin	matkustet	matkustaa
kirjat	kirj	kirja
kirjamme	kirj	kirja
kirjojen	kirj	kirja, kirjo

Taulukko 2. Esimerkkejä stemmauksen ja lemmauksen tuloksista

Taulukon 2 tuloksista huomaa, että stemmauksen jälkeen lopputulokset ovat vaihtelevampia kuin lemmauksen jälkeen. *Matkustaa*-verbin kolme taivutusmuotoa typistyivät kolmeen eri vartaloon, mikä ei ole ideaalia tiedonhaun kannalta. Vaikka perusmuotoistamisen jälkeen lopputulokset ovat taulukossa 2 yhteneväisiä, on kuitenkin otettava huomioon, että Voikko voi palauttaa enemmän kuin yhden analyysituloksen. Esim. *kirjojen* sai kaksi perusmuotoa: *kirja* ja *kirjo*.

Toteutuksessani tallennan Voikon analyysistuloksista uniikit perusmuodot. Käyttämässäni datassa esiintyy sana *kaikenlaista*, jolle Voikko antaa peräti viisi erilaista analyysitulosta. Tuloksista poimitut perusmuodot ovat *kaikenlainen*, *kaikenlaki*, *kaikenlaki*, *kaikenlaki* ja *kaikenlaki*. Ohjelmani ottaa näistä talteen vain *kaikenlainen* ja *kaikenlaki.* Eri tulkinnat saadaan talteen, mutta vältetään toisto.

4.5 Tutkimuksia stemmauksesta ja lemmauksesta suomen kielessä

Esittelen lyhyesti kaksi tutkimusta, joissa on vertailtu stemmausta ja lemmausta suomenkielisissä teksteissä. Korenius ja muut [2004] vertailevat stemmauksen ja lemmauksen tuloksia suomenkielisistä sanomalehtiartikkeleista. Stemmauksessa oli käytetty Snowball-stemmeriä, mutta sulkusanoja ei ollut karsittu. Lemmauksessa oli vuorostaan käytetty FINTWOL-ohjelmaa ja lemmauksen yhteydessä sulkusanat oli karsittu. Heidän johtopäätöksensä on, että lemmaus on suomen kielessä parempi normalisointitapa kuin stemmaus.

Myös Alkulan [2000] mukaan perusmuotoistaminen on parempi vaihtoehto suomen kielessä. Hänen väitöskirjassaan vertaillaan kyselyjen tuloksia kuudessa erilaisessa testitietokannassa. Ensimmäinen on perinteinen tietokanta, johon oli tallennettu sanat taivutusmuodoissaan. Tätä kutsutaan tutkimuksessa taivutusmuotohakemistoksi. Kyselyssä hakija katkaisee itse hakusanat. Toinen on taivutusmuotohakemisto, mutta kyselyssä käytetään sanan vartaloita. Kolmas vaihtoehto on kuin toinen, mutta tulokset seulotaan perusmuotoon palauttavalla ohjelmalla. Neljäs on sanojen

perusmuotoistaminen ennen tallennusta. Perusmuotoja käytetään myös kyselyssä. Jos sanaa ei voitu perusmuotoistaa, se tallennetaan taivutusmuodossaan. Yhdyssanoista haettiin alkuosia. Viides on sanojen perusmuotoistaminen ennen tallennusta. Yhdyssanoista tallennetaan kaikki osat ja osien yhdistelmät. Kuudes on kaksoishakemisto, jossa kysely kohdistui perusmuotojen hakemistoon ja taivutusmuotojen hakemistoon.

Alkula [2000] toteaa, että perusmuotojen joukosta hakemalla tulosten tarkkuus on parempi kuin taivutusmuotojen joukosta hakemalla. Jos kyselyssä otetaan myös yhdyssanat huomioon, niin saanti paranee. Alkula mainitsee, että lemmauksen riskinä on sanojen väärintulkinta, jolloin tallennetaan vääriä sanoja.

4.6 Sulkusanojen karsinta

Sulkusanat (stop words) ovat kielessä useasti esiintyviä sanoja. Ne voidaan karsia pois tekstistä esikäsittelyn aikana. Tämä voidaan automatisoida. Esimerkiksi kun Solr-ohjelmistossa luo ytimen, Solr-instanssin, niin sieltä löytyy valmiina eri kielille sulkusanalistoja, myös suomen kielelle [Apache Software Foundation. 2021c]. Suomen kielen sulkusanalista on suhteellisen suppea. Se sisältää n. 235 sulkusanaa. Diazin [2016] kokoama sulkusanalista on laajempi, sillä se sisältää yli 800 sanaa. Tästä listasta kannattaa huomioida, että pieni osa (n. 20) ei ole suomen kielen sanoja (esim. *näissältä, näissästä*). Käytän Diazin sulkusanalistasta tekstitiedosto-versiota, joten laitoin virheelliset sulkusanat kommentin sisälle, joten ohjelmani voi jättää ne huomiotta, kun se lukee tiedoston sisältöä. Taulukossa 3 on pieni määrä sulkusanoja Diazin listalta. Jaottelin ne sanaluokan mukaan.

Sanaluokka	**Sulkusanoja**
Adjektiivi	*pieni, suuri*
Adverbi	*huomenna, täällä*
Konjunktio	*että, ja*
Lukusana	*tuhat, yksi*
Pronomini	*itse, joka, meidän, minä, tämä*
Verbi	*menet, menimme, olen, olisit*

Taulukko 3. Esimerkkejä sulkusanoista suomen kielessä

Erityisesti suomen kielen kohdalla tulee huomioida, että jos sulkusanat karsitaan tekstistä, joka sisältää taivutusmuotoja, tulee listan sisältää sulkusanan eri taivutusmuodoista ainakin yleisimmin käytetyt muodot. Tämä luonnollisesti kasvattaa sulkusanalistan pituutta. Solrin ja Diazin sulkusanalistat sisältävät sanojen eri taivutusmuotoja.

Jos teksti stemmataan, olisi hyödyllistä karsia sulkusanat tekstistä ennen stemmausta. Jos teksti lemmataan, niin käytännössä sulkusanat voi karsia ennen tai jälkeen lemmauksen. Jos karsinta on lemmauksen jälkeen, niin silloin myös sulkusanat voivat olla perusmuodossa. Koska toteutuksessa käyttämäni sulkusanalista on laaja, niin sulkusanat karsitaan ennen lemmausta.

Alkula [2000] mainitsee, että sulkusanalistan käyttöä suomenkielisissä tekstissä ei ole pidetty hyödyllisenä, mutta sen käyttöä kannattaisi tutkia. Testaan, onko sulkusanoilla vaikutusta tuloksiin.

Zobel ja Moffat [2006] toteavat, että hakukoneiden yhteydessä kaikki termit tulisi indeksoida, myös sulkusanat ja numerot. Esimerkiksi sulkusanat voivat olla hyödyllisiä fraasihauissa. Toteutuksessani ei tehdä kuitenkaan fraasihakuja.

5 MySQL:n kokotekstihaku

MySQL on avoimen lähdekoodin tietokantaohjelmisto, jota kehittää ja jakaa Oracle Corporation. Se on laajasti käytössä oleva relaatiotietokanta. MySQL-nimessä oleva SQL-lyhenne on muodostettu sanoista Structured Query Language, joka on standardoitu kyselykieli. SQL:n avulla voi antaa käskyjä tietokannalle, minkä perusteella tietokanta suorittaa toimintoja esim. lisää tai poistaa tietoja tai tekee hakukyselyitä. [Oracle Corporation 2022]

MySQL sisältää runsaasti ominaisuuksia, mutta tässä luvussa keskityn kokoteksti-indeksointiin ja kokotekstihakuun.

5.1 Kokoteksti-indeksi

MySQL-tietokannassa on yhtenä indeksivaihtoehtona InnoDB FULLTEXT -indeksi, joka on kokoteksti-indeksi. Kokoteksti-indeksi tarkoittaa käänteisindeksiä tekstitiedostolle. Sen tavoitteena on nopeuttaa kyselyjä, kun sarakkeen tyypiksi on määritelty TEXT, VARCHAR tai CHAR. [Oracle Corporation 2022]

Kokoteksti-indeksin voi lisätä tauluun luonnin yhteydessä tai jälkeen päin muokkaamalla taulua. Jos dataa on paljon, on nopeampaa tuoda data tauluun, jossa ei ole kokoteksti-indeksiä ja lisätä indeksi tauluun datan tuonnin jälkeen. Jos sarakkeelle on määritelty InnoDB FULLTEXT-indeksi, niin taulun luonnin yhteydessä MySQL luo indeksiä varten kuusi indeksitaulua (auxiliary index table). Nämä sisältävät käänteisindeksin. [Oracle Corporation 2022]

Kun dataa syötetään sarakkeeseen, jolla on kokoteksti-indeksi, MySQL tokenisoi tekstin, ja syöttää yksittäiset sanat ja lisätietoineen indeksitauluihin. Sanan lisäksi indeksitauluun tallennetaan tieto siitä, mihin dokumenttiin se liittyy ja missä sana sijaitsee tekstissä. MySQL antaa jokaiselle kokoteksti-indeksin dokumentille piilotetun uniikin tunnisteen (DOC_ID), jota käytetään indeksitauluissa yhdistämään sana ja dokumentti. [Oracle Corporation 2022]

Kun kokoteksti-indeksin sisältävään tauluun lisätään dokumentti, aiheuttaa se useita pieniä lisäyksiä indeksitauluihin. Tästä syystä MySQL käyttää välimuistia kokoteksti-indeksin muutosten käsittelyssä ja vie muutokset erä (batch) kerrallaan indeksitauluihin, kun välimuisti on täyttynyt. Dokumentin poisto kokoteksti-indeksistä aiheuttaisi samalla tavalla useita poistoja indeksitauluihin, joten poistetun dokumentin tunniste DOC_ID laitetaan väliaikaisesti FTS_*_DELETED-tauluun. Tiedot jäävät indeksiin, mutta ennen kyselytulosten palauttamista MYSQL tarkastaa, että dokumentin tunniste ei ole merkitty poistetuksi. [Oracle Corporation 2022]

5.2 Kokotekstihaku

Kokotekstihaku voi kohdistua yhteen tai useampaan sarakkeeseen, kunhan jokaiselle on määritelty FULLTEXT-indeksi. Kokotekstihaussa käytetään MATCH()… AGAINST() -syntaksia. Se muodostetaan MATCH()-funktiosta, jolle annetaan sarakkeet, joista haetaan ja AGAINST()-funktiosta, jonka parametrinä on merkkijono, jota haetaan. AGAINST()-funktioon voi lisätä IN A NATURAL MODE -määreen, mutta se ei ole välttämätöntä, sillä se on oletusarvo.

Jos IN A NATURAL MODE -määreen sijaan käyttää IN A BOOLEAN MODE -määrettä, voi tällöin käyttää plus- ja miinusmerkkejä sanan alussa määräämään pitääkö sanan olla mukana tuloksissa (esim. ’+mySQL’) vai ei (esim. ’-Solr’). Jos AGAINST()-funktiolle annetaan parametrinä on ’+mySQL -Solr IN A BOOLEAN MODE’, hakisi MySQL kaikki rivit, joissa esiintyy mySQL, mutta ei Solr. Luvun 3.3 esimerkki relevanssiarvon laskemisesta liittyy kokotekstihaun boolean-moodin. [Oracle Corporation 2022]

Kysely 1 on esimerkki kokotekstihausta. Teksti, jolle esimerkissä haetaan samankaltaisuuksia, on tunnisteen 10 vastausteksti testidatasta [Museovirasto ja Suomen Kotiseutuliitto 2015].

```
SELECT response_id FROM test_table_1a
WHERE MATCH (processed_text)
AGAINST ('paikkakunta lähihistoria säilyttää tärkeä');
```

Kysely 1. Esimerkki kokotekstihausta

Esimerkin kokotekstihaku kohdistuu toteutuksen test_table_1a-taulun processed_text-sarakkeeseen ja kysely palauttaa vastauksien tunnisteet (response_id). MySQL hakee indeksitauluista ’paikkakunta lähihistoria säilyttää tärkeä’ -merkkijonon tokeneita. Toteutuksessa käytetty kysely on rakenteeltaan samanlainen kuin kysely 1. Erona on se, että toteutuksessa on asetettu enimmäismäärä tuloksille. AGAINST()-funktiolle annetaan tarkoituksella merkkijono puolikkaissa lainausmerkeissä (') kokolainausmerkkien (") sijaa. Jos käytetään kokolainausmerkkejä, niin MySQL suorittaisi fraasihaun, mikä toisi vähemmän osumia kuin kysely 1.

Oracle Corporationin [2022] mukaan mitatakseen haettavan merkkijonon ja processed_text-sarakkeessa olevien tekstien samankaltaisuutta MATCH()-funktio laskee relevanssin taulun joka riville. Relevanssi on laskettu dokumentin kokonaissanamäärän ja uniikkien sanojen lukumäärän sekä taulun kaikkien dokumenttien kokonaissanamäärän ja kyseisen sanan sisältävien dokumenttien määrän perusteella. Kokotekstihaun tuloksissa ensimmäisenä on dokumentti, jolla on korkein relevanssi. [Oracle Corporation 2022]

Kokotekstihaku ei ota huomioon sanoja, joissa on alle kolme merkkiä. Tätä voi muuttaa konfiguraatiossa, mutta toteutuksessa käytän oletusarvoa. MySQL sisältää taulun englanninkielisille sulkusanoille, joita kokotekstihaku käyttää. Omat sulkusanat voi lisätä sulkusanojen tauluun kokotekstihaun käytettäväksi. [Oracle Corporation 2022]

Kokotekstihaussa sana, joka on yli minimipituuden ja ei ole sulkusanojen joukossa, saa painoarvon. Jos sana esiintyy monessa dokumentissa, saa se matalamman painoarvon kuin sana, joka esiintyy harvoissa dokumentissa. Tämä on hyödyllistä erityisesti isojen tekstimäärien kanssa. [Oracle Corporation 2022]

6 Solrin MoreLikeThis-kysely

Solr on tehokas hakukone, joka käyttää Apache Lucene -kirjastoa. Jotta Solrilla voi hakea jotain, tulee ensin luoda ydin (core), joka on yksittäinen Solr-instanssi. Ydin sisältää indeksin ja konfiguraation, mm. kaavion (skeema). Perinteisen tietokantataulun riviä Solrissa vastaa dokumentti, joka sisältää kentät ja niiden arvot. Kentät määritellään kaaviossa. Kun data indeksoidaan eli syötetään Solriin, missä se talletetaan indeksiin. Myös Solr käyttää käänteisindeksiä. [Apache Software Foundation 2021f]

Kuten MySQL:n kohdalla niin esittelen Solrin monista ominaisuuksista vain niitä, jotka ovat merkityksellisiä toteutukseni kannalta. Aluksi käyn läpi, kuinka MoreLikeThis-kyselyn voi rakentaa ja sitten esittelen parametreja, joita kyselyssä on mahdollista käyttää.

6.1 MoreLikeThis-kyselyn rakentaminen

Solrin MoreLikeThis-kysely hakee lähdedokumentin termejä vastaavia dokumentteja indeksistä. MLT-kyselyn voi rakentaa usealla tavalla. Yleisin tapa on käyttää sitä kutsun käsittelijänä (request handler), jolloin sitä voi käyttää tarvittaessa. Toinen tapa tehdä MLT-kysely on käyttää hakukomponenttia (search component). Tämä ei ole suositeltava, sillä se hidastaa hakua, koska se suorittaa MoreLikeThis-analyysiin kaikille dokumenteille, jotka kysely palauttaa. Tämä on kuitenkin oletuksena mukana Solrin konfiguraatiossa. Kolmas tapa on käyttää kyselyn jäsentäjää (query parser), jolloin MoreLikeThis-toiminnon voi lisätä erilaisiin kyselyihin, esim. suodatuskyselyihin (filter query). [Apache Software Foundation 2021d]

Käytän toteutuksessani kutsun käsittelijää, jonka olen määritellyt ytimen solrconfig.xml-tiedostossa (ks. määrittely 2). Toteutuksessani ei ole suurta merkitystä, miten MLT-kysely rakennetaan, koska ohjelmassa on vain yksi hakukysely, jolla haetaan ehdotuksia.

```
<requestHandler name="/mlt" class="solr.MoreLikeThisHandler">
  <str name="mlt.fl">response_text</str>
</requestHandler>
```

Määrittely 2. MoreLikeThis-kutsun käsittelijän määrittely

6.2 MoreLikeThis-kyselyn parametreja

Seuraavaksi käyn läpi suurimman osan parametreistä, joita MoreLikeThis-kutsun käsittelijän kanssa voi käyttää. Kutsun käsittelijälle voi asettaa useita eri parametreja, joista monet ovat samoja kuin mitä hakukomponentin kanssa voi käyttää. Kentät, joista haetaan samankaltaisuuksia, määritellään pakollisella mlt.fl-parametrillä. Mlt.mintf määrittää minimifrekvenssin termin esiintymiselle lähdedokumentissa ja sen oletusarvo on 2. Mlt.mindf määrittää minimifrekvenssin, sille kuinka monessa dokumentissa termi saa esiintyä ja sen oletusarvo on 5. Vastaavasti voi määrittää dokumenteille

maksimifrekvenssin mlt.maxdf-parametrillä. Dokumenteille voi määrittää toisenkin maksimifrekvenssin. Mlt.maxfpct on maksimidokumenttifrekvenssi. Parametrin arvo, esim. 75, tarkoittaa, että termiä ei huomioida, jos se esiintyy enemmän kuin 75 % dokumentteja. Mlt.maxqt määrittää maksimimäärän hakutermeille kyselyssä. Oletusarvo tälle on 25. Sanalle voi määrittää minimipituuden (mlt.minwl) ja maksimipituuden (mlt.maxwl). Mlt.maxntp on oletusarvoltaan 5000 ja se on maksimäärä tokeneita kentässä, joka ei käytä termivektoreita (termVectors). Mlt.interestingTerms lisää kyselyn tuloksiin osion, jossa on listattu osuvimmat termit, joita on käytetty kyselyssä. Mlt.boost-parametrilla voi tehostaa kyselyä mielenkiintoisten termien (interesting terms) relevanssilla. [Apache Software Foundation 2021d]

Koodikatkelmassa 3 on MoreLikeThis-kyselyn luonti toteutukseni Document-luokan ehdotusten hakufunktiossa. Lähdedokumentin tunniste on id-muuttujassa ja tulosten enimmäislukumäärä on limit-muuttujassa.

```
SolrQuery query = new SolrQuery();
query.setRequestHandler("/mlt");
query.set("mlt.fl", "response_text");
query.set("mlt.mintf", 1);
query.set("mlt.mindf", 1);
query.set("mlt.minwl", 3);
query.set(("fl", "id")
query.setQuery("id:" + id);
query.setRows(limit);
```

Koodikatkelma 3. SolrQuery-objektin luonti finxtextrec-ohjelmassa

Parametri fl voi sisältää yhden tai useamman kentän, jotka näytetään hakutuloksissa. Jos fl-parametriä ei ole määritelty, lisää Solr oletuksena kaikki kentät hakutuloksiin. Asetin mlt.minwl-parametriin arvon 3, jolloin sanassa tulee olla vähintään kolme merkkiä, jotta se otetaan huomioon. MySQL:n kokotekstihaussa on sama raja sanan minimipituudelle.

Koska yhdessä indeksissä on ainoastaan 119 dokumenttia ja osa vastausteksteistä on lyhyitä, niin asetin parametreille oletusarvoja pienempiä arvoja, jotta Solr pystyy tarjoamaan ehdotuksia. Isommalla tekstimäärällä ei ole suositeltavaa käyttää näin pieniä arvoja.

Yhdellä vastaustekstillä tehty testikysely näyttää, miltä mlt.interestingTerms-parametrin hakutuloksiin luoma lista näyttää. Tein kyselyn test_core_1-ytimeen, missä suodatetaan sulkusanat. Tämä testikysely eroaa koodikatkelman 3 kyselystä mlt.interestingTerms-parametrin lisäksi myös siinä, että fl-parametrin arvo on ”id,response_text”, jotta

vastaustekstit ovat mukana hakutuloksissa. Teksti sisältää toistuvia sanoja ja se on testauksessa käytetystä datasta [Museovirasto ja Suomen Kotiseutuliitto 2015]:

> "Kulttuuriperintömme on aika pitkälti näkyvissä rakennuskannassa. Toivoisin säilyttävämpää linjaa rakennusten pysymiseen "vanhoina". En tarkoita museoimista vaan asuttujenkin rakennusten korjaamisessa vaaditaan energiatehokkuutta ja uusien rakennusmääräysten noudattamista. Miksi vanhoista rakennuksista pitäisi tehdä uusia? "

Kysely ja sen vastaus json-muodossa ovat liitteessä 2. Vastauksessa on neljä avainta. ResponseHeader-avain sisältää tietoja kyselystä, match-avain tiedot lähdedokumentista, response-avain hakutulokset ja interestingTerms-avain listan termeistä. Vastauksen interestingTerms-listalla on 20 merkkijonoa: *kulttuuriperintö, rakennuks, pitäi, tehd, näkyv, korjaamis, rakennuskan, säilyttäv, toivois, rakennusmääräyst, vaad, noudattam, energiatehokkuut, asutu, museoim, linj, pysymis, pitkält, tarko, vanho,* ja *rakennust*. Näistä termeistä *kulttuuriperintö, rakennus* tai *vanha* esiintyvät eri taivutusmuodoissa monissa muissa testidatan vastausteksteissä. Tein saman interestingTerms-parametrin sisältävän kyselyn myös test_core_2:seen. Vastauksen interestingTerms-lista sisälsi 24 merkkijonoa. Koska test_core_2 ei suodata sulkusanoja, niin termien joukossa oli myös neljä sulkusanaa.

7 Toteutuksen esittely

Tutkielman tavoitteena on vertailla kokotekstihaun tuloksia Solrin MoreLikeThis-kyselyn tuloksiin. Tätä varten toteutin Javalla fintextrec-nimisen ohjelman [Seitamäki 2022]. Toteutuksen ei ole varsinaisesti tarkoitus toimia itsenäisesti, vaan toisen ohjelman lisäosana. Ohjelma on toteutettu vaihtoehtojen testausta varten, joten ohjelman virheenkäsittely on suppeaa

Aluksi käyn läpi ohjelman vaatimat riippuvuudet. Sitten esittelen tietokantataulujen ja kaavion rakennetta. Lopuksi kerron mitä toimintoja fintextrec-ohjelma sisältää.

7.1 Riippuvuudet

Ohjelmointikielenä on Java 8. Toteutus on tehty Apache Netbeans-ohjelmointiympäristössä. Ohjelman tarvitsemat riippuvuudet on asennettu Maven-repositoriosta. Riippuvuudet ja niiden versionumerot ovat ohjelman pom.xml-tiedostossa, jonka sisältö on liitteessä 3.

Ohjelmaan on asennettu Voikkoa varten libvoikko (4.1.1), joka piti asentaa myös koneelle. Solrista on käytössä viimeisin versio (8.1.1) ja solr-solrj-kirjastosta asensin projektiin vastaavan version. Käytän MySQL-ohjelmistoa MAMP:in (4.1) kautta ja käytössä on vanha versio (5.6.34). Ohjelmaan on asennettu myös mysql-connector-java-riippuvuus (8.0.28). Ohjelma vaatii toimiakseen yhteyden MySQL-tietokantaan ja Solriin.

7.2 Tietokantataulujen rakenne

Käytän MySQL-serveriä MAMP:in kautta, joten phpMyAdmin-hallintapaneelissa loin toteutustani varten cultural_heritage-nimisen tietokannan ja siihen taulut: response, test_table_1a, test_table_1b, test_table_2a, test_table_2b, test_table_3a ja test_table_3b.

Kun viitataan test_table-alkuisiin tauluihin, niin käytetään test_table_*-nimitystä. Liitteessä 4 on tietokantataulujen luonti- ja muokkauslauseet. Koska test_table_*-taulut on luotu ja muokattu samalla tavalla, on liitteessä 4 esimerkkinä ainoastaan test_table_1a-taulun luonti- ja muokkauslauseet.

Response-taulussa on kaksi saraketta: id ja response_text. Id-sarake on taulun pääavain. Se on tyypiltään INT ja sillä on juokseva laskuri. Alkuperäinen data ei sisältänyt tunnisteita, joten tällä saatiin vastauksille tunnisteet. Response_text -kenttään syötetään alkuperäinen, käsittelemätön vastausteksti. Se on TEXT-tyyppinen. Response-taulun kentät eivät saa olla tyhjiä.

Test_table_*-tauluissa on jokaisessa kolme saraketta: id, response_id ja processed_text. Taulun pääavain on id. Se on INT-tyyppiä ja sitä kasvatetaan automaattisesti. Response_id on vierasavain, joka viittaa response-taulun tunnisteeseen (id). Se on samaa

INT-tyyppiä kuin response-taulun id. Response_id:lle on määritelty uniikki avain. Processed_text-sarake on käsiteltyä tekstiä varten ja se on tyypiltään TEXT. Sille on lisätty InnoDB:n FULLTEXT-indeksi. Test_table_*-taulujen kentät eivät saa olla tyhjiä.

7.3 Ytimen luonti ja konfiguraation muokkaaminen

Solr valmisteltiin käyttöön seuraavalla tavalla. Kun Solr oli ajossa, loin komentorivillä ohjelmaani varten kaksi ydintä: test_core_1 ja test_core_2. Käsittelin kummankin ytimen konfiguraatiota samalla tavalla. Ainoastaan sulkusanojen käsittelyyn liittyvä kohta on määritelty ytimissä eri tavalla. Molempien ytimien schema.xml- ja solrconfig.xml-tiedostot ovat toteutuksen lähdekoodin kanssa samassa koodikannassa [Seitamäki 2022].

Solr luo automaattisesti ytimelle solrconfig.xml-tiedoston ja managed_schema-nimisen kaaviotiedoston. Tämä on datavetoinen tapa käyttää Solria, missä kaavio arvaa kenttätyypin datan perusteella [Apache Software Foundation 2021c]. Vaihdoin kaaviotiedoston nimeksi schema.xml ja poistin kaaviosta kenttiä kuten _root_ ja _nest_path, sillä tässä toteutuksessa ei ole lapsidokumentteja. Käyttääkseni perinteisempää vaihtoehtoa, missä muokataan suoraan kaaviota, asetin solrconfig.xml-tiedostossa käyttöön ClassicIndexSchemaFactoryn. Jos en olisi ottanut tätä käyttöön, niin Solr käyttäisi ManagedIndexSchemaFactorya. Kaavio on silloin oletusarvoisesti muuttuva (mutable) ja sitä muokataan Schema API:n kautta. [Apache Software Foundation 2021e]

Kaksi kenttää, id ja _version_, olivat jo valmiiksi määriteltyjä kaaviossa, joten annoin niiden jäädä paikalleen kaaviossa. Id on tyypiltään merkkijono (string) ja se oli valmiiksi asetettu dokumentin uniikiksi avaimeksi (uniqueKey). Toteutuksessani vastaustekstin tietokantatunniste talletetaan id-kenttään.

Lisäsin kaavioon response_text-kentän alkuperäistä vastaustekstiä varten (ks. määrittely 3). Kentän tyypiksi on määritelty suomenkielinen teksti (text_fi). Response-text-kentän indexed-ominaisuus on tosi, jotta kentän arvoa voi käyttää kyselyissä hakemaan tuloksia. Stored-ominaisuus on myös tosi, jotta Solr voi tarjota kentän arvon kyselyn tuloksissa. Kenttä on määritelty pakolliseksi required-ominaisuudella. [Apache Software Foundation 2021b]

```
<field name=" response_text"
       type="text_fi"
       indexed="true"
       stored="true"
       required="true"
       termVectors="true"/>
```

Määrittely 3. Vastausteksti-kentän määrittely schema.xml-tiedostossa

MoreLikeThis-kysely kohdistuu response_text-kenttään. Jotta Solr ylläpitäisi termivektoreita (term vector) joka dokumentille, on kentälle määritelty termVectors-ominaisuus. Tämä nopeuttaa hakua, mutta kasvattaa indeksin kokoa. Siksi TermVectors-ominaisuuden oletusarvo on epätosi. [Apache Software Foundation 2021b]

Liitteessä 5 on listattuna test_core_1:n käyttämän schema.xml-tiedoston kenttämäärittelyt sekä text_fi-kenttätyypin määrittely. Text_fi-kenttämäärittely sisältää tekstin analysointia varten analyzer-elementin, jonka lapsielementteinä ovat suodatin tokenisointia varten (StandardTokenizerFactory), suodatin tekstin pienaakkosia varten (LowerCaseFilterFactory), suodatin sulkusanoille (StopFilterFactory) ja suodatin stemmaukseen (SnowballPorterFilterFactorya), mille on määritelty kieleksi suomi. Solr voi analysoida tietoja sekä indeksoinnin että kyselyn yhteydessä. [Apache Software Foundation 2021a]

Test_core_2:n kaavion eroaa test_core_1:n kaaviosta siinä, että text_fi-kentällä ei ole StopFilterFactorya, sillä text_core_2:ssa ei suodateta pois sulkusanoja. Asetin test_core_1-ytimen käyttämään Diazin sulkusanat [2016] sisältävää tiedostoa, jotta ne vaihtoehdot, joissa sulkusanat karsitaan, käyttävät samoja sulkusanoja.

7.4 Ohjelman sisältämät toiminnot

Fintextrec-ohjelma koostuu seuraavista tiedostoista: App.java, DatabaseConnection.java, Document.java ja ProcessedText.java. App.java-tiedostossa sijaitsee main()-metodi. DatabaseConnection.java-tiedostossa on tietokantayhteyden muodostamista varten connect()-metodi. Document.java-tiedosto sisältää MLT-vaihtoehtojen metodit. ProcessedText.java-tiedosto sisältää FT-vaihtoehtoihin liittyvät metodit.

Ohjelmalla ei ole konfiguraatiotiedostoa. DatabaseConnection.java-tiedostoon tulee lisätä tietokannan käyttöä varten käyttäjä ja salasana. App.java-tiedostossa vakiomuuttujissa tulee määritellä tietokannan ja Solrin käyttöön tarvittavat polut sekä tietokannan, taulun ja ytimen nimet. Kenttien nimet ovat kuitenkin kovakoodattuja.

App.java-tiedostossa voi antaa polun tekstitiedostoon, jossa on sulkusanat listattuna, mutta se ei ole välttämätöntä. Jos polku on tyhjä merkkijono, ohjelma ei yritä lukea

sulkusanojen tiedostoa. Toteutus olettaa, että sulkusanatiedoston sisällä olevat kommenttirivit alkavat #-merkillä.

App.java-tiedostossa on tekstin esikäsittelyä varten lippumuuttujia. RemoveStopWords-muuttujalla määritellään, karsitaanko sulkusanat vai ei. Lemmatize-muuttujalla määritellään perusmuotoistetaanko sanat vai ei. AllowInflectedForm-muuttujan perusteella otetaan talteen alkuperäinen taivutusmuoto, jos sanalle ei löydy perusmuotoa, silloin kun arvo on tosi.

Koska ohjelmalla ei voi lisätä tietoja response-tauluun, niin toin datan tietokantaan csv-tiedostossa, missä jokainen rivi sisälsi yhden alkuperäisen, käsittelemättömän vastaustekstin.

Kun response-taulussa on vastaustekstejä, voi kutsua Document-luokan indexAll()-metodia, joka indeksoi datan. Se hakee kaikki response_text-sarakkeen alkuperäiset tekstit indeksoitavaksi. Uudelleen indeksointia varten Document-luokalla on deleteAll()-metodi, jolla voi poistaa kaikki dokumentit indeksistä, minkä jälkeen ne voi indeksoida uudelleen indexAll()-metodilla. Solria varten esikäsitellään teksti siten, että ohjelma ottaa alkuperäisen vastaustekstin ympärillä olevat lainausmerkit pois.

Vastaavasti ProcessedText-luokan processAll()-metodilla käsitellään kaikki response-taulussa olevat vastaukset. Se kutsuu processText()-metodia, jolle annetaan alkuperäinen teksti käsiteltäväksi. Metodin paluuarvo on merkkijono, joka sisältää käsitellyn tekstin.

ProcessText()-metodi käsittelee alkuperäisen vastaustekstin seuraavilla tavoilla. Ensin tekstistä poistetaan välimerkit. Tämän jälkeen teksti pilkotaan merkkijonoiksi tyhjän välin kohdalta eli tokenisoidaan. Merkkijonot lisätään taulukkoon, jotta ne voidaan käydä yksitellen läpi. Ne muutetaan pienaakkosiin sulkusanojen karsintaa varten. Jos sulkusanat on tarkoitus karsia ja sana löytyy sulkusanojen listalta, se karsitaan pois. Jos sulkusanoja ei ole tarkoitus karsia tai sana ei ole sulkusana, niin sana jatkaa eteenpäin. Jos tekstiä ei ole tarkoitus lemmata, niin otetaan suoraan taivutusmuoto talteen. Jos teksti on tarkoitus lemmata, niin sana jatkaa analyze()-metodiin Voikon analysoitavaksi.

Voikko suorittaa morfologisen analyysin, josta poimitaan perusmuoto BASEFORM-attribuutista. Myös perusmuoto muutetaan pienaakkosiin. Jos perusmuotoja on enemmän kuin yksi yhdelle sananmuodolle, niin perusmuodot kerätään samaan merkkijonoon välilyönnillä erotettuna. Kuitenkin jos sama perusmuoto esiintyy useammin kuin kerran, niin se lisätään merkkijonoon vain kerran. Koska ohjelma muuntaa Voikon antaman perusmuodon pienaakkosiin, niin esimerkiksi *satuja*-sanasta otetaan talteen vain ”satu”, vaikka Voikko antaa sille kaksi perusmuotoa *satu* (nomini) ja *Satu* (etunimi). Jos perusmuotoa ei löytynyt ja taivutusmuodot ovat sallittuja, niin otetaan perusmuodon sijaan taivutusmuoto talteen.

Kun alkuperäinen teksti on käyty läpi, niin talteen kerätyt merkkijonot tallennetaan test_table_*-taulun processed_text-kenttään insertProcessedText()-metodin avulla. Jos taulusta ei löydy vastauksen tunnistetta (response_id), niin lisätään uusi rivi tauluun. Jos vastauksen tunniste on jo test_table_*-taulussa, niin processed_text-kenttä vain päivitetään.

Taulukossa 4 esitän kahden esimerkkitekstin avulla miltä teksti näyttää processText()-metodin eri vaiheissa. Esimerkkitekstit (tunnisteet 92 ja 52) ovat testidatasta [Museovirasto ja Suomen Kotiseutuliitto 2015].

Vaihe	**Vaiheen tapahtuma**	**Tunnisteen 92 teksti**	**Tunnisteen 52 teksti**
1	Alkuperäinen teksti on haettu tietokannasta.	"Tarinoita, satuja, kansanperintöä, käsitöitä ja niiden ohjeita."	"Lippalioskeja ja Kivijalkakauppoja."
2	Tekstistä poistetaan välimerkit ja erikoismerkit.	Tarinoita satuja kansanperintöä käsitöitä ja niiden ohjeita	Lippalioskeja ja Kivijalkakauppoja
3	Teksti pilkotaan välilyöntien kohdalta merkkijonoiksi, jotka lisätään listaan.	[Tarinoita, satuja, kansanperintöä, käsitöitä, ja, niiden, ohjeita]	[Lippalioskeja, ja, Kivijalkakauppoja]
4	Sanat käydään läpi yksitellen ja ne muunnetaan pienaakkosiin.	tarinoita satuja kansanperintöä käsitöitä ja niiden ohjeita	lippalioskeja ja kivijalkakauppa
5	Tässä vaihtoehdossa sanat lemmataan. Taivutusmuoto otetaan, jos perusmuotoa ei löydy. Sulkusanoja ei karsita.	tarina satu kansanperintö käsityö ja ne ohje	lippalioskeja ja kivijalkakauppa
6	Käsitelty merkkijono tallennetaan tietokantaan.	tarina satu kansanperintö käsityö ja ne ohje	lippalioskeja ja kivijalkakauppa

Taulukko 4. Esimerkkitekstien käsittely yhden FT-vaihtoehdon eri vaiheissa

Ensimmäisessä vaiheessa teksti on siinä muodossa kuin se on response-taulun response_text-kentässä. Vaiheiden 2-4 esikäsittelyt tehdään kaikille kokotekstihakuvaihtoehdoille. Viidennessä vaiheessa on yhdistetty useampi toiminto.

Tässä vaiheessa suoritetaan kullekin vaihtoehdolle erikseen määritellyt tehtävät. Tässä vaihtoehdossa sanat perusmuotoistetaan ja jos perusmuotoa ei löydy, niin otetaan taivutusmuoto talteen. Voikon analyysin tulokset tunnisteen 92 tekstille ovat liitteessä 1. Tunnisteen 52 tekstissä olevassa *Lippalioskeja*-sanassa on kirjoitusvirhe, joten Voikko ei löydä sille perusmuotoa. Tässä vaihtoehdossa sulkusanoja ei karsita pois tekstistä. Viimeisessä vaiheessa merkkijono on siinä muodossa, jossa se tallennetaan test_table_*-taulun processed_text-kenttään

Document-luokalla ja ProcessedText-luokalla on molemmilla oma getSuggestions()-metodi. Metodeilla on seuraavat parametrit: ytimen tai testitaulun nimi, lähdedokumentin tunniste sekä enimmäislukumäärä tuloksille. Metodin paluuarvo on ehdotukset eli listan tunnisteita, joiden tekstit ohjelman mukaan sisältävät samankaltaisuuksia. Jotta ProcessedText -luokan ehdotusten hakumetodi voi löytää samankaltaisia tekstejä, tulee test_table_*-taulun processed_text-taulussa olla käsiteltyjä tekstejä. Myös Document-luokan ehdotusten hakumetodi vaatii, että ytimeen on syötetty dataa.

Koska joka FT-vaihtoehdolle on vain yksi taulu, johon käsitellyt tekstit tallennetaan, niin päätin pitää yhdyssanat kokonaisina. Eli en pilko yhdyssanoja osiin. Jos yhdyssanat pilkotaan, niin silloin voidaan saada osumia myös yhdyssanan osista, mutta silloin yhdyssanat ja niiden osat ovat enemmän edustettuna käsitellyssä tekstissä kuin yksittäinen sana.

8 Data ja testattavat vaihtoehdot

Testausta varten tarvitsin suomenkielistä dataa. Tässä luvussa esittelen aluksi testauksessa käyttämäni aineiston ja sitten eri vaihtoehdot.

8.1 Aineisto

Toteutuksen testaamiseen hain suomenkielistä tekstiä. Tavoitteenani oli löytää tekstiä, jonka tyylistä keskivertokielenpuhuja kirjoittaisi vapaamuotoisessa sähköpostiviestissä. Vältin tekstejä, jotka sisältävät runsaasti minkään erikoisalan sanastoa. Tekstit saavat kuitenkin sisältää yleiskielestä poikkeavia sanoja, esim. puhekielisiä ilmaisuja. Lisäksi tekstit saavat sisältää kirjoitusvirheitä.

Hain Yhteiskuntatieteellisen tietoarkiston [2021] Aila-palvelusta kvalitatiivisia aineistoja, sillä niiden joukosta olisi mahdollista löytää testaukseen sopivanpituisia tekstejä. Käytetyn aineiston nimi on Kaikkien yhteinen kulttuuriperintö 2014 ja sen tekijöinä ovat Museovirasto ja Suomen kotiseutuliitto [2015]. Aineiston käyttöehto on A eli se on vapaasti käytettävissä ilman rekisteröitymistä (CC BY 4.0). Aineisto on kerätty Kaikkien yhteinen kulttuuriperintö -kyselyssä Otakantaa.fi-sivustolla 15.4.-15.8.2014. Keruumenetelmänä on ollut itsetäytettävä verkkolomake. Kyselyllä on pyritty kartoittamaan suomalaisten suhdetta kulttuuriperintöönsä. Se on toiminut osana Faron sopimuksen ratifioinnin valmistelua. [FSD2981 aineisto-opas 2020] Opetus- ja kulttuuriministeriö [2017] tiedotti marraskuussa 2017, että Faron puiteyleissopimus on tullut voimaan Suomessa.

Aineistoon liittyvä data on rtf-tekstitiedostossa. Jokaisen kysymyksen alle on luettelomaisesti listattu vastaukset. Kysymykset ovat suomeksi ja niitä on 12 kappaletta. Vastauksia on 130 per kysymys, mutta myös tyhjät on laskettu vastauksiksi. Aineisto ei sisällä vastaajien taustatietoja. Vastaajan eri kysymyksiin antamien vastauksien välillä ei ole säilytetty yhteyttä. [Museovirasto ja Suomen Kotiseutuliitto 2015]

Valitsin dataksi ensimmäisen kysymyksen vastaukset, sillä siinä oli vain vähän tyhjiä vastauksia. Museoviraston ja Suomen Kotiseutuliiton [2015] keruulomakkeen ensimmäinen kysymysteksti on seuraavanlainen:

1. Kulttuuriperinnöllä tarkoitetaan menneisyydestä perittyjä aineellisia ja aineettomia voimavaroja, muuttuvia arvoja, uskomuksia, tietoja ja perinteitä. Millaista kulttuuriperintöä haluat siirtää eteenpäin?

Poistin tyhjät vastaukset (7 kpl) ja ruotsinkieliset vastaukset (4 kpl). Jäljelle jäi 119 vastaustekstiä, jotka ovat tietokantatunnisteiden kera listattuna liitteessä 6 [Museovirasto ja Suomen Kotiseutuliitto 2015]. Osa vastauksista on lyhyitä sekä luettelomaisia, ja osa

vastauksista vuorostaan koostuu useammasta lauseesta. Lyhyin vastaus on yhden sanan pituinen. Pisin vastaus on 299 sanaa pitkä. Useissa vastauksissa on kirjoitusvirheitä.

Koska data on tekstitiedossa, niin siirsin ensimmäisen kysymyksen vastaustekstit ensin taulukkoon. Alkuperäisessä datassa jokaisen vastaustekstin ympärillä on lainausmerkit, joiden annoin olla paikallaan. Tallensin sen csv-tiedostona utf-8-formaatissa, missä muodossa datan sai vietyä tietokantaan.

8.2 Vaihtoehtojen esittely

Testatessa haen ehdotuksia kahdeksalle eri vaihtoehdolle. Kokotekstihakuvaihtoehdot (FT) ovat taulukossa 5 ja MoreLikeThis-vaihtoehdot (MLT) taulukossa 6. FT-alkuisissa vaihtoehdoissa käsitelty teksti tallennetaan test_table_*-taulun processed_text-sarakkeeseen ja ehdotukset haetaan MySQL:n InnoDB FULLTEXT -haulla. MLT-alkuisissa vaihtoehdoissa alkuperäinen teksti indeksoidaan test_core_*-ytimen response_text-kenttään ja ehdotukset haetaan Solrin MoreLikeThis-kyselyllä. Pieni kirjain lyhenteessä perässä viittaa sulkusanojen karsintaan. Jos kirjain on a, niin sulkusanat karsitaan. Jos kirjain on b, niin sulkusanoja ei karsita.

Lyhenne	**Taulun nimi**	**Lemmaus**	**Sallitaanko taivutusmuodot?**	**Karsitaanko sulkusanat?**
FT1a	test_table_1a	Kyllä	Ei	Kyllä
FT1b	test_table_1b	Kyllä	Ei	Ei
FT2a	test_table_2a	Kyllä	Kyllä, jos perusmuotoa ei löydy.	Kyllä
FT2b	test_table_2b	Kyllä	Kyllä, jos perusmuotoa ei löydy.	Ei
FT3a	test_table_3a	Ei	Kyllä	Kyllä
FT3b	test_table_3b	Ei	Kyllä	Ei

Taulukko 5. Kokotekstivaihtoehdot

FT1a-vaihtoehdossa vastausteksti perusmuotoistetaan ja vain perusmuodot tallennetaan. Jos Voikko ei löydä sanalle perusmuotoa, sanaa ei huomioida. Sulkusanat karsitaan pois tekstistä esikäsittelyvaiheessa. Verrattuna muihin FT-vaihtoehtoihin tässä vaihtoehdossa käsitelty teksti muistuttaa vähiten alkuperäistä tekstiä.

FT1b-vaihtoehdossa vastausteksti perusmuotoistetaan ja vain perusmuodot tallennetaan. Jos Voikko ei löydä sanalle perusmuotoa, sanaa ei huomioida. Sulkusanoja ei karsita pois tekstistä.

FT2a-vaihtoehdossa vastausteksti perusmuotoistetaan. Jos Voikko ei tarjoa sanalle yhtään perusmuotoa, niin laitetaan tilalle taivutusmuoto eli sana alkuperäisessä muodossa, joka voi sisältää taivutuspäätteitä. Sulkusanat karsitaan pois tekstistä.

FT2b-vaihtoehdossa vastausteksti perusmuotoistetaan. Jos Voikko ei tarjoa sanalle yhtään perusmuotoa, niin tilalle laitetaan taivutusmuoto. Sulkusanoja ei karsita pois tekstistä. Tämä vaihtoehto on luvussa 7.4 käsitelty esimerkki.

FT3a-vaihtoehdossa vastaustekstiä ei perusmuotoisteta vaan alkuperäisen tekstin taivutusmuodot tallennetaan sellaisenaan. Sulkusanat karsitaan pois tekstistä.

FT3b-vaihtoehdossa vastaustekstiä ei perusmuotoisteta vaan taivutusmuodot tallennetaan sellaisenaan. Sulkusanoja ei karsita pois tekstistä. Verrattuna muihin FT-vaihtoehtoihin tässä vaihtoehdossa käsitelty teksti muistuttaa eniten alkuperäistä tekstiä.

Lyhenne	**Ytimen nimi**	**Stemmaus**	**Suodatetaanko sulkusanat?**
MLTa	test_core_1	Kyllä	Kyllä
MLTb	test_core_2	Kyllä	Ei

Taulukko 6. MoreLikeThis-vaihtoehdot

MLT-vaihtoehdoissa sanat stemmataan käyttäen SnowballPorterFilterFactorya. MLTa-vaihtoehdon ytimen skeemassa on text_fi-tyypin kentälle määritelty StopFilterFactory. Se suodattaa sulkusanat pois kyselyistä. MLTb-vaihtoehdossa sulkusanat jätetään tekstiin, joten MLT-vaihtoehdoista tämän vaihtoehdon teksti muistuttaa eniten alkuperäistä tekstiä.

9 Tulokset

Tässä luvussa esittelen mitä tuloksia sain testatessa kahdeksaa eri vaihtoehtoa fintextrec-ohjelmalla liitteessä 6 olevalla testidatalla. Tein neljä testiä eli hain neljälle vastaustekstille ehdotukset test_table_*-tauluista ja test_core_*-ytimistä. Ehdotusten enimmäislukumäärä per kysely oli viisi, joten ehdotuksia tuli enintään 40 per testi. Olen kiinnostunut siitä, mitä tunnisteita esiintyy ehdotuksissa, mutta en juurikaan anna painoarvoa ehdotusten järjestykselle.

Ennen ehdotusten hakemista ohjelmalla kävin läpi testidataksi valitsemani vastaustekstit, joista valitsin enintään viisi vastaustekstiä, jotka oman arvioni perusteella ovat samankaltaisia kuin lähdedokumentti. Arvioni perustuu intuitioon. En järjestänyt omia ehdotuksiani osuvuuden mukaan, mutta mainitsen erikseen, jos joku ehdotuksista on erityisen osuva.

Käyn läpi tekemäni testit ja niistä saadut tulokset neljässä alaluvussa. Testit ovat siinä järjestyksessä kuin tein ne. Viidennessä alaluvussa vertailen omia ehdotuksia vaihtoehtojen ehdotuksiin. Viimeisessä alaluvussa vertailen testeissä saatuja ehdotuksia vaihtoehtopareittain.

9.1 Ensimmäinen testi: ehdotukset tunnisteelle 116

Ensimmäisessä testissä haetaan ehdotuksia tunnisteen 116 vastaustekstille, joka ei ole kovin pitkä verrattuna muihin vastausteksteihin. Tämä koostuu kuitenkin useammasta virkkeestä.

(1) "Taitoja: miten tehdään saunavihta tai karjalanpiirakoita. Tietoja: Miten ennen on eletty ja millaisia tarinoita eri paikoilla on. Ymmärrystä: Miten menneisyys vaikuttaa nykypäivään. Paikkoja ja rakennuksia: vanhaa pitää säilyttää."

Valitsin omiksi ehdotuksiksi tunnisteet 1, 7, 9, 62 ja 76, jotka löytyvät liitteestä 6. Tunnisteen 76 vastausteksti on oman arvioni perusteella osuvin, koska siinä käsitellään tarinoita, paikkoja, rakennuksia ja menneisyyttä tuntemusta. Vaihtoehtojen ehdotukset ovat taulukossa 7 ja ehdotuksissa esiintyvät tunnisteet ja niiden frekvenssit ovat taulukossa 8. Ehdotuksissa esiintyy 17 eri tunnistetta.

MLTa	MLTb	FT1a	FT1b	FT2a	FT2b	FT3a	FT3b
113	53	76	73	76	73	113	113
9	20	19	65	94	65	20	20
20	7	94	94	19	34	96	96
25	113	65	113	65	113	70	53
62	9	7	107	7	107	24	70

Taulukko 7. Ensimmäisen testin ehdotukset vaihtoehdoittain

Tunnisteet	Frekvenssi
113	6
20, 65	4
7	3
9, 19, 34, 53, 70, 73, 76, 94, 96, 107	2
24, 25, 62	1

Taulukko 8. Ensimmäisen testin ehdotettujen tunnisteiden frekvenssit

Kolmen eniten esiintyneen tunnisteen vastaustekstit ovat taulukossa 9. Tunnisteen 113 vastausteksti on näistä vastausteksteistä aiheeltaan lähimpänä lähdedokumentin aihetta.

Id	Vastausteksti
113	"Suomi on valtiona nuori ja sen kulttuurinen menneisyys on ollut enemmän hävittävää kuin säilyttävää. Sanotaan, että kansallisen identiteetin vahvistuminen tapahtui vasta toisen maailmansodan aikana. Vahva usko teknologian edistyksellisyyteen on myös vaikuttanut siihen, että uusinta teknologiaa on myös pidetty parempana kuin vanhaa. Suomalaiseen modernismiin ei siihenkään liity vahvaa perinteiden ymmärrystä. Haluan edistää kulttuuria, jossa vanhan ymmärtäminen on edellytys uuden luomiselle. Jossa menneisyys ja nykyisyys yhdistyvät toisiaan kunnioittavalla tavalla "
20	"En halua siirtää eteenpäin mitään erikoisempaa kulttuuriperintöä, vaan paremminkin viestittää, että eletään yksinkertaisesti hetkessä ja löydetään jokainen itse oma totuutemme. Menneisyys on menneisyyttä eikä sen tule vaikuttaa siihen mitä me tällä hetkellä tai tulevaisuudessa olemme. Yhteisöllisestä taustastaan on hyvä toki ymmärtää jotain, mutta se on sitten ihan eri asia mitä vaikka uskomusten ja perinteiden siirtäminen eteenpäin. "
65	"Suomalaisia rituaaleja ja seremonioita, myös niitä, jotka eivät ole minullekaan siirtyneet. Vanhoista uskomuksista juontavia tapoja, jotka eivät rajoita nykyihmisen elämää mutta rikastuttavat sitä antamalla kokemuksen omista juurista. Kaikista tärkeintä on kuitenkin kansanmusiikki ja tanssi kaikissa muodoissaan. Tätä on poljettu alas niin kauan, että minä,olen musiikkitaustasta huolimatta vasta kolmikymppisenä saanut tietää, mitä aito suomalainen kansanmusiikki on nykypäivänä. Tätä ei pitäisi enää piilotella, vaan tuoda esiin positiivisella ja viihdyttävällä tavalla kaikissa sopivisa yhteyksissä. Kansanmusiikin on tarkoitus olla hauskaa yhdessä tekemistä, johon voi jokainen jollain tavalla osallistua. "

Taulukko 9. Ensimmäisen testin kolme eniten ehdotettua vastaustekstiä

FT1a:lla ja FT2a:lla on viisi samaa ehdotusta. Näissä molemmissa perusmuotoistetaan ja karsitaan sulkusanat. MLT-vaihtoehdoilla on kolme samaa tunnistetta ehdotuksissa. Samoin on MLTb:llä ja FT3b:llä, joissa ei kummassakaan perusmuotoisteta eikä karsita sulkusanoja. FT1b:n ja FT2b:n ehdotuksissa on kolme samaa tunnistetta. Näissä kahdessa vaihtoehdossa perusmuotoistetaan, mutta ei karsita sulkusanoja. FT3-vaihtoehtojen ehdotuksissa on myös kolme samaa tunnistetta.

Omista ehdotuksistani neljä löytyy vaihtoehtojen ehdotusten joukosta. Vastausteksti, jonka arvioin osuvimmaksi, esiintyy ehdotuksissa kahdesti. Se esiintyi lemmausvaihtoehtojen, joissa karsitaan sulkusanat, ehdotuksissa. Kolme eniten ehdotettua tunnistetta eivät ole omien ehdotuksieni joukossa.

9.2 Toinen testi: ehdotukset tunnisteelle 92

Toiseen testiin valitsin lyhyen vastaustekstin, jonka tunniste on 92. Lähdedokumentin teksti on lyhyt ja luettelomainen, joten omien ehdotusten hakeminen oli sujuvampaa kuin ensimmäisessä testissä. Tekstissä olevat sanat (pl. *ohje*) esiintyvät myös muissa vastausteksteissä.

(2) "Tarinoita, satuja, kansanperintöä, käsitöitä ja niiden ohjeita."

Omat ehdotukseni ovat tunnisteet 1, 22, 33, 44 ja 96 (ks. liite 6). Vaihtoehtojen ehdotukset ovat taulukossa 10 ja ehdotuksissa esiintyvät tunnisteet ja niiden frekvenssit ovat taulukossa 11. Ehdotuksissa esiintyy 12 eri tunnistetta.

MLTa	MLTb	FT1a	FT1b	FT2a	FT2b	FT3a	FT3b
96	96	1	1	1	1	1	1
1	80	33	33	33	9	96	96
84	1	9	22	9	22	9	80
95	43	22	96	22	96	50	43
9	84	96	9	96	33	70	47

Taulukko 10. Toisen testin ehdotukset vaihtoehdoittain

Tunnisteet	Frekvenssi
1, 96	8
9	6
22, 33	4
43, 80, 84	2
47, 50, 70, 95	1

Taulukko 11. Toisen testin ehdotettujen tunnisteiden frekvenssit

Kokotekstihaun lemmausvaihtoehdot (FT1a, FT1b, FT2a, FT2b) antavat samat ehdotukset. MLTb- ja FT3b-vaihtoehdoilla on neljä samaa vaihtoehtoa. Kummassakaan ei karsita sulkusanoja. MLTb-vaihtoehdossa stemmataan ja FT3b-vaihtoehdossa säilytetään alkuperäiset taivutusmuodot.

MLT-vaihtoehtojen ehdotuksissa on kolme samaa tunnistetta. Vaikka tekstissä esiintyy ainoastaan kaksi sulkusanaa (*ja*, *niiden*), niin silti FT3a- ja FT3b-vaihtoehdoissa on vain kaksi samaa ehdotusta.

Taulukossa 12 on ehdotusten tekstit, joiden frekvenssi on kuusi tai suurempi. Kaksi eniten ehdotettua ovat aiheeltaan lähellä lähdedokumenttia. Arvioin tunnisteiden 1 ja 96 tekstit hyviksi osumiksi aiheen kannalta.

Id	**Vastausteksti**
1	"Haluan siirtää suullisen kulttuurin rikkautta: muistitietoa, ajankohtaisia näkökulmia aikaan, eri väestöryhmien kokemuksia muuttuvasta Suomesta, mutta pidän tärkeänä myös vanhempaa perinnekulttuuria: kalevalaista runoutta, satuja, tarinoita, uskomuksia, koko vanhan suullisen kulttuurin rikkautta, joka antaa sisältöjä ja luomistyön aineksia myös nykyajalle Myös rakennuskulttuuri, käden taidot, vanhat työtavat ovat asioita, joiden kautta hahmotetaan historiaa ja ihmisen elämää. Ilman näitä olisimme irrallisia ja juurettomia, eivätkä tulevat sukupolvet olisi tietoisia niistä elämänsisällöistä, jotka koskevat erityisesti tavallisten ihmisten elinkaarta. Luonnon ja maiseman oennaiset kohteet kuuluvat myös säilytettävään kulttuuriperintöön. "
96	"Vanhoja rakennuksia eli arkkitehtuuria, kirkkoja, arvokkaita käsitöitä (myös kotitekoisia!) ja taidetta, lauluja, sanontoja, suvun tarinoita, perinteitä, valokuvia suvusta, käsityötaitoja (kangaskäsityöt, lankakäsityöt, punamultamaalin keitto, pärekaton tekeminen, päreiden höyläämineen, hirsitalon kunnossapito,...), suvun perinteiset marjapaikat, suvun perinteiset leivontareseptit, hyvä suomen kielitaito ja sanasto,..."
9	"kaikki kultturiperintö sisältää kokemuksia ja tietoa, jota mahdolisesti myös tulevaisuudessa tarvitaan. Lisäksi perintö vahvistaa identiteettiä. Koska kaikkea ei ehkä voi säilöä, on aineellinen perintö oleellista ja silloin myös esineisiin / rakennuksiin yms. liittyvä käyttötieto. Rakennettu kulttuuriperintö sisältää paljon tietoa ja kokemusta sekä työtavoista että rakennusaineista. Lisäksi rakennukset ja rakennelmat vahvistavat paikallista muistia - niihin liittyy tarinoita ja muistuksia, jotka helposti jäävät unholaan, jos mikään ei tarinoista ja ihmisistä muistuta. "

Taulukko 12. Toisen testin kolme eniten ehdotettua vastaustekstiä

Omista ehdotuksistani neljä tunnistetta löytyy vaihtoehtojen ehdotusten joukosta. Tunnisteet 1 ja 96 esiintyvät kaikkien vaihtoehtojen ehdotuksissa ja tunnisteet 22 ja 33 esiintyvät lemmausvaihtoehtojen ehdotuksissa. Tässä testissä kolmen eniten ehdotetuimman tunnisteen joukossa on kaksi omaa ehdotusta. Koska lähdedokumentin teksti on lyhyt, niin siitä oli helpompi poimia avainsanoja, joita hain muiden vastaustekstien joukosta, mikä todennäköisesti edesauttoi samojen osumien tulemista.

9.3 Kolmas testi: ehdotukset tunnisteelle 57

Kolmanteen testiin valitsin tunnisteen 57 vastaustekstin, joka on luettelomainen. Siinä esiintyy sanoja, jotka toistuvat muissa teksteissä, kuten *vanhat* ja *rakennukset* sekä sanoja, joita ei muissa ole, kuten *hietikot* ja *naavametsiköt*. Oletan saavani samansuuntaisia tuloksia kuin toisessa testissä.

(3) "Kalliomaalaukset, aarnimetsät, puhtaat rannat, kansanuskomukset, marja- ja sienipaikat, hienot hietikot, naavametsiköt, vanhat rakennukset."

Omiksi ehdotuksiksi valitsin tunnisteet 7, 8, 40, 54 ja 96 (ks. liite 6). Vaihtoehtojen ehdotukset ovat taulukossa 13 ja ehdotuksissa esiintyvät tunnisteet ja niiden frekvenssit ovat taulukossa 14. Ehdotuksissa esiintyy 15 eri tunnistetta.

MLTa	MLTb	FT1a	FT1b	FT2a	FT2b	FT3a	FT3b
7	7	7	7	7	7	54	54
116	116	54	54	54	40	1	1
8	54	40	42	40	54	9	9
54	8	1	75	1	1	49	49
25	17	24	83	24	24	53	53

Taulukko 13. Kolmannen testin ehdotukset vaihtoehdoittain

Tunnisteet	Frekvenssi
54	8
7	6
1	5
24, 40	3
8, 9, 49, 53, 116	2
17, 25, 42, 75, 83	1

Taulukko 14. Kolmannen testin ehdotettujen tunnisteiden frekvenssit

Kokotekstihakujen pelkät taivutusmuodot sisältävät vaihtoehdot (FT3a, FT3b) antoivat viisi samaa ehdotusta. FT1a-, FT2a- ja FT2b-vaihtoehdoilla on myös viisi samaa ehdotusta. MLT-vaihtoehtojen ehdotuksissa esiintyy neljä samaa tunnistetta.

Vastaustekstissä on vain yksi sulkusana, joten oletin, että sulkusanojen karsinnalla ei olisi suurta merkitystä ehdotuksissa. Tämä pätee muiden kohdalla, paitsi ei FT1-vaihtoehtojen kanssa. FT1a:lla on vain kaksi samaa ehdotusta FT1a:n kanssa.

Omien ehdotusteni tunnisteista neljä on samoja (7, 8, 40, 54) kuin vaihtoehdoilla. Näistä kaksi esiintyy vaihtoehtojen ehdotuksissa eniten ja ne ovat taulukossa 15. Arvioin tunnisteen 54 vastaustekstin näistä kahdesta paremmaksi osumaksi.

Id	**Vastausteksti**
54	"Rakennettua kulttuuriperintöä: asuinsijat, viljely- ja teollisuusympäristöjä, liikenteen väyliä (tiet, polut, kanavat, joet, satama, venepoukamat); suomalaiseen muinaisuskoon ja luonnonuskoon liittyviä paikkoja (esim. uhrikivet, lehdot, karsikot, kalliomaalaukset), arkeologisia jäänteitä, muinaisia asuinpaikkoja, tämän päivän rakennuskulttuuria, kansanviisautta, puutietoutta - taitoa elää ilman sähköä ja öljypohjaisia polttoaineita. "
7	"Alueelliset eri paikkoihin liittyvät tarinat ja henkilöt ovat tärkeää kulttuuriperintöä. Olisi myös hienoa, ettei kaikkea vanhaa pureta pois eli arvostettaisiin ja huolehdittaisiin vanhoista rakennuksista. Kylien raiteilla näkyisi vuosikymmenten eri rakennustyylit - sekä kauniit että kamalat. Lisäksi vahva yhteisöllisyys on hienoa kulttuuriperintöä useassa pienemmässä kylässä - sen toivoisi säilyvän. "

Taulukko 15. Kolmannen testin kaksi eniten ehdotettua vastaustekstiä

FT3a:n ja FT3b:n ehdotuksissa ei esiinny tunnistetta 7, vaikka se on muissa vaihtoehdoissa ensimmäisenä ehdotuksena. Tämä johtuu siitä, että FT3-vaihtoehdoissa tallennetaan pelkät taivutusmuodot ja eri taivutusmuodossa olleita sanoja ei siten löydetty.

9.4 Neljäs testi: ehdotukset tunnisteelle 40

Neljänteen testiin valitsin tunnisteen 40 tekstin, koska siinä on toistuvia sanoja (*rakennus*, *vanha*, *uusi*).

(4) "Kulttuuriperintömme on aika pitkälti näkyvissä rakennuskannassa. Toivoisin säilyttävämpää linjaa rakennusten pysymiseen "vanhoina". En tarkoita museoimista vaan asuttujenkin rakennusten korjaamisessa vaaditaan energiatehokkuutta ja uusien rakennusmääräysten noudattamista. Miksi vanhoista rakennuksista pitäisi tehdä uusia? "

Omat ehdotukseni lähdedokumentille ovat 1, 7, 9, 49 ja 76 (ks. liite 6). Tunnisteen 7 vastausteksti on oman arvioni perusteella osuvin, sillä siinä on yhtenä aiheena vanhat

rakennukset. Vaihtoehtojen ehdotukset ovat taulukossa 16 ja ehdotuksissa esiintyvät tunnisteet ja niiden frekvenssit ovat taulukossa 17. Ehdotuksissa esiintyy 16 eri tunnistetta.

MLTa	MLTb	FT1a	FT1b	FT2a	FT2b	FT3a	FT3b
7	7	83	73	83	73	7	7
116	49	68	113	68	83	81	65
84	116	7	83	7	113	37	81
56	84	73	65	73	65	65	37
37	81	1	34	1	34	13	80

Taulukko 16. Neljännen testin ehdotukset vaihtoehdoittain

Tunnisteet	Frekvenssi
7	6
65, 73, 83	4
37, 81	3
1, 34, 68, 84, 113, 116	2
13, 49, 56, 80	1

Taulukko 17. Neljännen testin ehdotusten frekvenssit

FT1a- ja FT2a-vaihtoehdoilla on viisi samaa ehdotusta. Myös FT1b:llä ja FT2b:llä on samat ehdotukset. FT3-vaihtoehdoilla on neljä samaa tunnistetta ehdotuksissa. MLT-vaihtoehdoilla on kolme samaa tunnistetta.

Lähdedokumentin teksti on pisin testeissä olleista teksteistä, ja tekstin toistuvat sanat esiintyvät myös muissa vastauksissa, mutta silti ehdotukset jakautuvat kuten muissakin testeissä.

Vaihtoehtojen ja omien ehdotusteni joukossa on kolme samaa tunnistetta. Osuvimmaksi arvioimani tunniste 7 esiintyy eniten vaihtoehtojen ehdotuksissa. Sen teksti on taulukossa 18. Taulukossa on toisena tunniste 83, koska sillä on parhaimmat sijoitukset tunnisteista, joiden frekvenssi on neljä. Tunnisteen 83 teksti ei ole aiheeltaan lähellä lähdedokumentin tekstiä.

Id	Vastausteksti
7	"Alueelliset eri paikkoihin liittyvät tarinat ja henkilöt ovat tärkeää kulttuuriperintöä. Olisi myös hienoa, ettei kaikkea vanhaa pureta pois eli arvostettaisiin ja huolehdittaisiin vanhoista rakennuksista. Kylien raiteilla näkyisi vuosikymmenten eri rakennustyylit - sekä kauniit että kamalat. Lisäksi vahva

	yhteisöllisyys on hienoa kulttuuriperintöä useassa pienemmässä kylässä - sen toivoisi säilyvän. "
83	"Kaikenlaista. Jos esim. difficult heritage näkökulma uupuu kansallisen kulttuuriperinnön käsittelystä, ollaan vääristyneen kansallis-identiteettin ja historian vääristelyn asialla. Kulttuuriperinnön tulee olla terveen identiteetin ja kansallistunteen rakentaja. Tällöin kaikella kulttuuriperinnöllä on sijansa. Kulttuuriperintö elää ja siirtyy formaalin kentän (esim. museot) ohella arjessa. Arjessa siirtyvän kulttuuriperinnön ei tule "alistua" kenenkään "halulle" siirtää tai olla siirtämättä sitä eteenpäin. Kulttuuriperinnön tulee antaa elää ja hengittää vapaasti. Se täytyy kuitenkin alistaa terveelle itsereflektiolle: kyvylle ymmärtää esim. menneisyyden kontekstia (vaikkapa tiettyyn aikaan liittyviä kansallisia haasteita kun vaikkapa uutta kansakuntaa ja sen kansallisidentiteettiä synnytetään). Kulttuuriperinnön moninaisuus tulisi huomioida tällä hetkellä korostetusti. Ajassamme elävä käsitys "puhtaasta suomalaisuudesta" tulisi ohjata näkemään kansallisidentiteettimme monikulttuuriset juuret. "

Taulukko 18. Neljännen testin kaksi eniten ehdotettua vastaustekstiä

9.5 Omat ehdotukset verrattuna vaihtoehtojen ehdotuksiin

Vertailen lyhyesti, kuinka paljon samoja ehdotuksia omissa ehdotuksissani on verrattuna vaihtoehtojen ehdotuksiin. Taulukossa 19 on listattuna, kuinka monta samaa tunnistetta on minun ja vaihtoehdon ehdotuksissa per testi. Viimeisessä sarakkeessa on keskiarvo minun ja kyseisen vaihtoehdon samoille ehdotuksille.

Vaihtoehto	Testi 1	Testi 2	Testi 3	Testi 4	Keskiarvo
MLTa	2	2	3	1	2,0
MLTb	2	2	3	2	2,3
FT1a	2	4	3	2	2,8
FT1b	0	4	2	0	1,5
FT2a	2	4	3	2	2,8
FT2b	0	4	3	0	1,8
FT3a	0	2	1	1	1,0
FT3b	0	2	1	1	1,0

Taulukko 19. Samojen tunnisteiden lukumäärät

Kun vertailen testeittäin, niin toisessa ja kolmannessa testissä on eniten samoja tunnisteita. Ensimmäisessä testissä on eniten nollia eli ei yhtään samaa tunnistetta. Kun vertailen vaihtoehdoittain, niin omissa ehdotuksissani on eniten samoja tunnisteita

FT1a:n ja FT2a:n kanssa. MLT-vaihtoehtojen kanssa on myös jonkin verran samoja. Omat ehdotukseni eroavat eniten FT3-vaihtoehtojen ehdotuksista.

Oletin, että samoja ehdotuksia olisi ollut tasaisemmin yksi tai kaksi per vaihtoehto, mutta samojen tunnisteiden lukumäärät ovatkin vaihtelevammat. Toisen testin samojen tunnisteiden lukumäärät ovat selkeästi suuremmat kuin ensimmäisessä tai neljännessä testissä.

9.6 Vaihtoehtoparien ehdotusten vertailu

Tässä alaluvussa tarkastelen vaihtoehtoja pareittain. Taulukossa 20 on joka rivillä kaksi vaihtoehtoa ja niiden ehdotuksissa esiintyvien samojen tunnisteiden lukumäärät testeittäin. Taulukon viimeisessä sarakkeessa on parin samojen tunnisteiden lukumäärien keskiarvo.

Vaihtoehtopari		Testi 1	Testi 2	Testi 3	Testi 4	Keskiarvo
MLTa	MLTb	3	3	4	3	3,3
MLTa	FT1a	0	3	2	1	1,5
MLTa	FT1b	0	3	2	1	1,5
MLTa	FT2a	0	3	2	1	1,5
MLTa	FT2b	1	3	2	1	1,8
MLTa	FT3a	2	3	1	2	2,0
MLTa	FT3b	2	2	1	2	1,8
MLTb	FT1a	1	2	2	1	1,5
MLTb	FT1b	0	2	2	0	1,0
MLTb	FT2a	0	2	2	1	1,3
MLTb	FT2b	1	2	2	0	1,3
MLTb	FT3a	2	2	1	2	1,8
MLTb	FT3b	3	4	1	2	2,5
FT1a	FT1b	2	5	2	2	2,8
FT1a	FT2a	5	5	5	5	5,0
FT1a	FT2b	1	5	5	2	3,3
FT1a	FT3a	0	3	2	1	1,5
FT1a	FT3b	0	2	2	1	1,3
FT1b	FT2a	2	5	2	1	2,5
FT1b	FT2b	3	5	2	5	3,8
FT1b	FT3a	0	3	1	1	1,3
FT1b	FT3b	0	2	1	1	1,0
FT2a	FT2b	0	5	5	2	3,0
FT2a	FT3a	0	3	2	1	1,5
FT2a	FT3b	0	2	2	1	1,3
FT2b	FT3a	1	3	2	1	1,8
FT2b	FT3b	1	2	2	1	1,5
FT3a	FT3b	3	2	5	4	3,5

Taulukko 20. Eri vaihtoehtoparien samojen ehdotusten lukumäärät

FT3a ja FT3b-vaihtoehdot saivat keskenään enemmän samoja ehdotuksia kuin muiden kanssa. MLT-vaihtoehdot saivat myös keskenään eniten samoja ehdotuksia, vähintään kolme samaa tunnistetta per testi. MLTa sai kokotekstihakuvaihtoehdoista eniten samoja tunnisteita FT3a:n kanssa ja MLTb FT3b:n kanssa. MLT-vaihtoehdoissa tekstit stemmataan ja FT3-vaihtoehdoissa on tallennettu sanat alkuperäisissä taivutusmuodoissaan. Kummassakaan ei perusmuotoisteta sanoja.

Vaihtoehdot, joissa ei lemmata löysivät parhaat parinsa omiensa joukosta. Vastaavasti vaihtoehdot, joissa lemmataan löysivät parhaat parinsa omasta joukostaan. FT1a- ja FT2a- vaihtoehdoilla oli kaikissa testeissä samat ehdotukset. Näissä vaihtoehdoissa karsitaan sulkusanat ja perusmuotoistetaan sanat. Erona on se, että FT1a:ssa sallitaan vain perusmuodot. FT1b ja FT2b saivat myös keskenään eniten samoja tunnisteita. Toiseksi paras pari FT1a:lle ja FT2a:lle on FT2b eli vaihtoehto, jossa otetaan taivutusmuoto, jos perusmuotoa ei löydy ja sulkusanat karsitaan. Toiseksi eniten samoja tunnisteita FT1b ja FT2b saivat, kun parina oli FT1a. Lemmausvaihtoehdoissa vaikuttaa paras pari olevan se, jossa sulkusanat käsitellään samalla tavalla.

10 Yhteenveto

Kaikissa testeissä kaikilta vaihtoehdoilta tuli ehdotuksia ja oman arvioni perusteella osa niistä on relevantteja. Testieni perusteella eniten vaikutusta on lemmauksella ja stemmauksella. Lemmausvaihtoehdon paras pari on todennäköisesti toinen lemmausvaihtoehto. Stemmausvaihtoehdon paras pari on vastaavasti joko toinen stemmausvaihtoehto tai pelkkiä taivutusmuotoja sisältävä vaihtoehto. Paras pari pelkkiä taivutusmuotoja sisältävälle vaihtoehdolle on toinen taivutusmuotovaihtoehto. Sulkusanojen karsinnalla vaikuttaa olevan sen verran merkitystä, että lemmausvaihtoehdoissa paras pari on se, jossa sulkusanat käsitellään samalla tavalla. Tämä tuli todennäköisesti esille lemmausvaihtoehtojen kohdalla siitä syystä, että lemmausvaihtoehtoja oli enemmän kuin muita vaihtoehtoja.

Omat ehdotukseni olivat lähimpänä lemmausvaihtoehtoja, joissa karsitaan sulkusanat. Tämä oletettavasti johtui siitä, että manuaalisessa arvioinnissa sulkusanoja ei oteta huomioon. Kuten aiemmissa tutkimuksissa on tullut ilmi, niin myös näissä testeissä lemmauksella saadaan paremmin käsiteltyä suomenkielinen teksti kyselyitä varten.

Testauksessa käyttämäni data sisälsi vain yhden tekstikentän, joten toteutusta olisi mielenkiintoista laajentaa. Ohjelma, tietokanta ja kaavio voisivat ottaa huomioon enemmän kuin yhden kentän, johon kokotekstihaku tai MoreLikeThis-kysely kohdistuu. Tätä voisi testata useammalla kentällä sekä isommalla dokumenttien määrällä ja mahdollisesti heterogeenisemmällä datalla. Tässä toteutuksessa en pilkkonut yhdyssanoja, joten olisi myös kiinnostavaa nähdä miten yhdyssanojen pilkkominen osiin vaikuttaisi ehdotuksiin.

Tässä työssä esitelty ohjelma on suunniteltu osaksi järjestelmää, jossa ilmoitetaan ja käsitellään havaintoja, esimerkiksi työturvallisuuteen liittyen. Työturvallisuusilmoitus on vapaamuotoista tekstiä ja ohjelman avulla voidaan etsiä samankaltaisia ilmoituksia ja niihin liittyviä toimenpiteitä. Näin käyttäjälle voidaan informoida, kuinka aikaisemmin on toimittu samankaltaisten tapauksien kohdalla.

11 Viiteluettelo

Ahonen, Mikael. 2019. Finnish stemming and lemmatization in Python. Haettu osoitteesta: *https://data.solita.fi/finnish-stemming-and-lemmatization-in-python/* (18.4.2022).

Alaterä, Anu ja Kai Halttunen. 2002. *Tiedonhaun perusteet – osa lukutaitoa*. BTJ Kirjastopalvelu, 31-32.

Alkula, Riitta. 2000. *Merkkijonoista suomen kielen sanoiksi. Suomen kielen morfologisten tulkintaohjelmien liittäminen tekstitiedonhakujärjestelmään ja liittämisen vaikutukset tekstin tallennukseen ja haluun.* Väitöskirja. Informaatiotutkimuksen laitos, Tampereen yliopisto, 5-6, 271-272, 276.

Apache Software Foundation. 2021a. Analyzers. Haettu osoitteesta: *https://solr.apache.org/guide/8_11/analyzers.html* (27.4.2022).

Apache Software Foundation. 2021b. Defining fields. Haettu osoitteesta: *https://solr.apache.org/guide/8_11/defining-fields.html* (27.4.2022).

Apache Software Foundation. 2021c. Installing Solr. Haettu osoitteesta: *https://solr.apache.org/guide/8_11/installing-solr.html* (15.4.2022).

Apache Software Foundation. 2021d. MoreLikeThis. Haettu osoitteesta: *https://solr.apache.org/guide/8_11/morelikethis.html* (15.4.2022).

Apache Software Foundation. 2021e. Schema factory definition in solrconfig. Haettu osoitteesta: *https://solr.apache.org/guide/8_11/schema-factory-definition-in-solrconfig.html* (15.4.2022).

Apache Software Foundation. 2021f. Solr glossary. Haettu osoitteesta: *https://solr.apache.org/guide/8_11/solr-glossary.html* (4.5.2022).

Corevoikko. 2017. Morphological analysis. Haettu osoitteesta: *https://github.com/voikko/corevoikko/blob/master/libvoikko/doc/morphological-analysis.txt* (15.4.2022).

Diaz, Gene. 2016. Finnish stopwords collection. Haettu osoitteesta: *https://github.com/stopwords-iso/stopwords-fi/blob/master/stopwords-fi.txt* (10.4.2022).

FSD2981 aineisto-opas. 2020. Kaikkien yhteinen kulttuuriperintö 2014, aineisto-opas. Yhteiskuntatieteellinen tietoarkisto. Haettu osoitteesta: *http://urn.fi/urn:nbn:fi:fsd:T-FSD2981* (11.2.2021).

Hakulinen, Auli, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen ja Irja Alho. 2008. Ison suomen kieliopin verkkoversio. Suomalaisen Kirjallisuuden Seura. Haettu osoitteesta: *https://scripta.kotus.fi/visk* (15.4.2022), § 97.

Järvelin, Kalervo ja Eero Sormunen. 2010 Tiedon tallennus ja haku. Teoksessa: Serola, Sami (toim.). *Ote informaatiosta: johdatus informaatiotutkimukseen ja interaktiiviseen mediaan.* BTJ, 164-167, 184-185.

Karlsson, Fred. 2009. Yleinen kielitiede. 4. painos. Gaudeamus Helsinki University Press, 2-3.

Kielitoimiston sanakirja. 2021a. Perusmuoto. Kotimaisten kielten keskuksen verkkojulkaisuja 35. Haettu osoitteesta: *https://www.kielitoimistonsanakirja.fi/#/perusmuoto?searchMode=all* (20.4.2022).

Kielitoimiston sanakirja. 2021b. Sana. Kotimaisten kielten keskuksen verkkojulkaisuja 35. Haettu osoitteesta: *https://www.kielitoimistonsanakirja.fi/#/sana?searchMode=all* (20.4.2022).

Kielitoimiston sanakirja. 2021c. Sane. Kotimaisten kielten keskuksen verkkojulkaisuja 35. Haettu osoitteesta: *https://www.kielitoimistonsanakirja.fi/#/sane?searchMode=all* (1.5.2022)

Kielitoimiston sanakirja. 2021d. Vartalo. Kotimaisten kielten keskuksen verkkojulkaisuja 35. Haettu osoitteesta: *https://www.kielitoimistonsanakirja.fi/#/vartalo?searchMode=all* (28.4.2022).

Korenius, Tuomo, Jorma Laurikkala, Kalervo Järvelin and Martti Juhola. 2004. Stemming and lemmatization in the clustering of Finnish text documents. In: *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management,* 625.

Museovirasto ja Suomen Kotiseutuliitto. 2015. *Kaikkien yhteinen kulttuuriperintö 2014*. Versio 1.0. Yhteiskuntatieteellinen tietoarkisto. Haettu osoitteesta: *http://urn.fi/urn:nbn:fi:fsd:T-FSD2981* (11.2.2021).

NLTK Project. 2022. Sample usage for tokenize. Haettu osoitteesta: *https://www.nltk.org/howto/tokenize.html* (18.4.2022).

Oikofix. 2022a. Lue ja analysoi tekstiä. Haettu osoitteesta: *https://oikofix.com/analysis* (28.4.2022).

Oikofix. 2022b. Tietoja Oikofix-palvelusta. Haettu osoitteesta: *https://oikofix.com/contact* (16.4.2022).

Opetus- ja kulttuuriministeriö. 2017 Faron sopimus voimaan Suomessa. Haettu osoitteesta: *https://okm.fi/-/faron-sopimus-voimaan-suomessa* (16.4.2022).

Oracle Corporation. 2022. MySQL 5.6 reference manual. Haettu osoitteesta: *https://downloads.mysql.com/docs/refman-5.6-en.a4.pdf* (9.4.2022), 4, 1422-1426, 1429-1431, 1897-1901.

Petrović, Đorđe and Milena Stanković. 2019. The influence of text preprocessing methods and tools on calculating text similarity. In: *Facta universitatis (NIŠ). Series: Mathematics and informatics.* Vol. 34, No 5, 973-974, 976-980.

Porter, Martin. 2001. Snowball: A language for stemming algorithms. Haettu osoitteesta: *https://snowballstem.org/texts/introduction.html* (17.4.2022).

Seitamäki, Sirkku. 2022. Fintextrec-ohjelma. Haettu osoitteesta: *https://gitlab.com/SirkkuS/fintextrec* (19.5.2022).

Snowball. 2022a. Demo. Haettu osoitteesta: *https://snowballstem.org/demo.html#Finnish* (15.4.2022).

Snowball. 2022b. Finnish stemming algorithm. Haettu osoitteesta: *https://snowballstem.org/algorithms/finnish/stemmer.html* (17.4.2022).

Snowball. 2022c. Snowball. Haettu osoitteesta: *https://snowballstem.org/* (8.5.2022).

Snowball. 2022d. Stemming algorithms. Haettu osoitteesta: *https://snowballstem.org/algorithms/* (8.5.2022).

Tieteen termipankki. 2013. Finite-state transducer. Haettu osoitteesta: *https://tieteentermipankki.fi/wiki/Language_Technology:finite-state-transducer* (15.4.2022).

Voikko. 2021. Free linguistic software and data for Finnish. Haettu osoitteesta: *https://voikko.puimula.org/* (21.4.2021).

Yhteiskuntatieteellinen tietoarkisto. 2021. Aila. Haettu osoitteesta: *https://services.fsd.tuni.fi/catalogue/index?study_language=fi* (11.2.2021).

Zobel, Justin and Alistair Moffat. 2006. Inverted files for text search engines. In: *ACM Computing Surveys.* Vol. 38, Issue 2, 3, 8.

Liite 1: Voikon analyysin tulokset esimerkkitekstille

Esimerkkiteksti on Kaikkien yhteinen kulttuuriperintö 2014 -aineistosta [Museovirasto ja Suomen kotiseutuliitto 2015]: "Tarinoita, satuja, kansanperintöä, käsitöitä ja niiden ohjeita.".

```
[{NUMBER=plural, STRUCTURE==pppppppppp, BASEFORM=tarina,
SIJAMUOTO=osanto, CLASS=nimisana,
FSTOUTPUT=[Ln][Xp]tarina[X]tarino[Sp][Nm]ita,
WORDBASES=+tarina(tarina)}]

[{NUMBER=plural, STRUCTURE==pppppp, BASEFORM=satu, SIJAMUOTO=osanto,
CLASS=nimisana, FSTOUTPUT=[Ln][Xp]satu[X]satu[Sp][Nm]ja,
WORDBASES=+satu(satu)}, {NUMBER=plural, STRUCTURE==ippppp,
BASEFORM=Satu, SIJAMUOTO=osanto, CLASS=etunimi,
FSTOUTPUT=[Lee][Xp]Satu[X]satu[Sp][Nm]ja, WORDBASES=+Satu(Satu)}]

[{NUMBER=singular, STRUCTURE==pppppp=pppppppp, BASEFORM=kansanperintö,
SIJAMUOTO=osanto, CLASS=nimisana,
FSTOUTPUT=[Ln][Xp]kansa[X]kans[Sg][Ny]an[Bh][Bc][Ln][Xp]perintö[X]peri
ntö[Sp][Ny]ä, WORDBASES=+kansan(kansa)+perintö(perintö)}]

[{NUMBER=plural, STRUCTURE==pppp=ppppp, BASEFORM=käsityö,
SIJAMUOTO=osanto, CLASS=nimisana,
FSTOUTPUT=[Ln][Xp]käsi[X]kä[Sn][Ny]si[Bh][Bc][Ln][Xp]työ[X]tö[Sp][Nm]i
tä, WORDBASES=+käsi(käsi)+työ(työ)}]

[{NUMBER=plural, STRUCTURE==ppppppp, BASEFORM=ohje, SIJAMUOTO=osanto,
CLASS=nimisana, FSTOUTPUT=[Ln][Xp]ohje[X]ohje[Sp][Nm]ita,
WORDBASES=+ohje(ohje)}]
```

Liite 2. MoreLikeThis-kysely ja vastaus, joka sisältää interestingTerms-listan

http://localhost:8983/solr/test_core_1/mlt?q=id:40&mlt.fl=response_text&mlt.mintf=1&mlt.mindf=1&mlt.minwl=3&fl=id,response_text&mlt.interestingTerms=list&rows=5


```
{
  "responseHeader":{
    "status":0,
    "QTime":2},
  "match":{"numFound":1,"start":0,"numFoundExact":true,"docs":[
      {
        "id":"40",
        "response_text":"Kulttuuriperintömme on aika pitkälti
näkyvissä rakennuskannassa. Toivoisin säilyttävämpää linjaa
rakennusten pysymiseen \"vanhoina\". En tarkoita museoimista vaan
asuttujenkin rakennusten korjaamisessa vaaditaan energiatehokkuutta ja
uusien rakennusmääräysten noudattamista. Miksi vanhoista rakennuksista
pitäisi tehdä uusia? "}]
  },
  "response":{"numFound":61,"start":0,"numFoundExact":true,"docs":[
      {
        "id":"7",
        "response_text":"Alueelliset eri paikkoihin liittyvät tarinat
ja henkilöt ovat tärkeää kulttuuriperintöä. Olisi myös hienoa, ettei
kaikkea vanhaa pureta pois eli arvostettaisiin ja huolehdittaisiin
vanhoista rakennuksista. Kylien raiteilla näkyisi vuosikymmenten eri
rakennustyylit - sekä kauniit että kamalat. Lisäksi vahva
yhteisöllisyys on hienoa kulttuuriperintöä useassa pienemmässä kylässä
- sen toivoisi säilyvän. "},
      {
        "id":"116",
        "response_text":"Taitoja: miten tehdään saunavihta tai
karjalanpiirakoita. Tietoja: Miten ennen on eletty ja millaisia
tarinoita eri paikoilla on. Ymmärrystä: Miten menneisyys vaikuttaa
nykypäivään. Paikkoja ja rakennuksia: vanhaa pitää säilyttää."},
      {
        "id":"84",
        "response_text":"Maisemien, rakennusten, tarinoiden ja
musiikkiperinnön vaaliminen on merkityksellistä jälkipolville."},
```

```
      {
        "id":"56",
        "response_text":"Rakennusperinnön korjaamisen aineellista ja
aineetonta perintöä. Esimerkiksi ylläpitämällä pienten käsityöalojen
ammatti- ja erikoisammattitutkintoja (esim. kultaajakisälli- ja
kultaajamestari)."},
      {
        "id":"37",
        "response_text":"Eurooppalaiset kulttuuriset juuremme:
kristinusko, analyyttinen filosofia (kriittinen ajattelu),
keskusteleva demokratia, suvaitsevaisuus ja toinen toisesta
välittäminen ovat kulttuuriperintömme vaalittavia asioita. Myös
moderni taide ja rakennuskanta, menneitä teknologisia käytänteitä
koskevan tiedon osaamisen säilyttäminen ja ylläpito ovat periaatteessa
asioita, joita tulisi jotenkin siirtää eteenpäin."}]
  },
  "interestingTerms":["kulttuuriperintö",
    "rakennuks",
    "pitäi",
    "tehd",
    "näkyv",
    "korjaamis",
    "rakennuskan",
    "säilyttäv",
    "toivois",
    "rakennusmääräyst",
    "vaad",
    "noudattam",
    "energiatehokkuut",
    "asutu",
    "museoim",
    "linj",
    "pysymis",
    "pitkält",
    "tarko",
    "vanho",
    "rakennust"]}
```

Liite 3. Ohjelman riippuvuudet pom.xml-tiedostossa

```
<?xml version="1.0" encoding="UTF-8"?>
<project xmlns="http://maven.apache.org/POM/4.0.0"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://maven.apache.org/POM/4.0.0
http://maven.apache.org/xsd/maven-4.0.0.xsd">
  <modelVersion>4.0.0</modelVersion>
  <groupId>com.app</groupId>
  <artifactId>fintextrec</artifactId>
  <version>1.0-SNAPSHOT</version>
  <packaging>jar</packaging>
  <dependencies>
    <dependency>
      <groupId>org.apache.solr</groupId>
      <artifactId>solr-solrj</artifactId>
      <version>8.11.1</version>
    </dependency>
    <dependency>
      <groupId>mysql</groupId>
      <artifactId>mysql-connector-java</artifactId>
      <version>8.0.28</version>
    </dependency>
    <dependency>
      <groupId>org.slf4j</groupId>
      <artifactId>slf4j-jdk14</artifactId>
      <version>1.7.36</version>
    </dependency>
    <dependency>
      <groupId>org.puimula.voikko</groupId>
      <artifactId>libvoikko</artifactId>
      <version>4.1.1</version>
    </dependency>
  </dependencies>
  <properties>
    <project.build.sourceEncoding>UTF-8</project.build.sourceEncoding>
    <maven.compiler.source>1.8</maven.compiler.source>
    <maven.compiler.target>1.8</maven.compiler.target>
  </properties>
  <name>fintextrec</name>
</project>
```

Liite 4: Tietokantataulujen luonti- ja muokkauslauseet

```
CREATE TABLE `response` (
            `id` INT NOT NULL AUTO_INCREMENT,
            `response_text` TEXT NOT NULL,
            PRIMARY KEY (`id`)
) ENGINE = InnoDB;

CREATE TABLE `test_table_1a` (
            `id` INT NOT NULL AUTO_INCREMENT,
            `response_id` INT NOT NULL,
            `processed_text` TEXT NOT NULL,
            PRIMARY KEY (`id`),
            UNIQUE (`response_id`),
            FULLTEXT (`processed_text`)
) ENGINE = InnoDB;

ALTER TABLE `test_table_1a`
            ADD FOREIGN KEY (`response_id`)
            REFERENCES `response`(`id`);
```

Liite 5. Kenttämäärittelyjä test_core_1-ytimen kaaviosta

```
<field name="id"
       type="string"
       indexed="true"
       stored="true"
       required="true"
       multiValued="false"/>

<field name="_version_"
       type="plong"
       indexed="false"
       stored="false"/>

<field name=" response_text"
       type="text_fi"
       indexed="true"
       stored="true"
       required="true"
       termVectors="true"/>

<fieldType name="text_fi"
           class="solr.TextField"
           positionIncrementGap="100">
  <analyzer>
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class=="solr.StopFilterFactory"
            ignoreCase="true"
            words="lang/stopwords_fi.txt"/>
    <filter class="solr.SnowballPorterFilterFactory"
            language="Finnish"/>
  </analyzer>
</fieldType>
```

Liite 6. Vastaustekstit

Vastaukset ovat Kaikkien yhteinen kulttuuriperintö 2014 -aineistosta [Museovirasto ja Suomen kotiseutuliitto 2015].

Tunniste 1

"Haluan siirtää suullisen kulttuurin rikkautta: muistitietoa, ajankohtaisia näkökulmia aikaan, eri väestöryhmien kokemuksia muuttuvasta Suomesta, mutta pidän tärkeänä myös vanhempaa perinnekulttuuria: kalevalaista runoutta, satuja, tarinoita, uskomuksia, koko vanhan suullisen kulttuurin rikkautta, joka antaa sisältöjä ja luomistyön aineksia myös nykyajalle Myös rakennuskulttuuri, käden taidot, vanhat työtavat ovat asioita, joiden kautta hahmotetaan historiaa ja ihmisen elämää. Ilman näitä olisimme irrallisia ja juurettomia, eivätkä tulevat sukupolvet olisi tietoisia niistä elämänsisällöistä, jotka koskevat erityisesti tavallisten ihmisten elinkaarta. Luonnon ja maiseman oennaiset kohteet kuuluvat myös säilytettävään kulttuuriperintöön. "

Tunniste 2

"Elävää, historiatietoisuutta lisäävää sekä nykyhetkeä ja tulevaisuutta palvelevaa kulttuuriperintöä ."

Tunniste 3

"Omassa perheessäni isovanhemmilta opittua ruokakulttuuria, sanoja ja tottumuksia. On myös negatiivista kulttuuriperintöä ja tapoja, joita haluan tunnistaa, mutta en halua välittää lapsilleni. Oman sukuni ruokakulttuuria olen käsitellyt valokuvataiteeni kautta koko perheen yhteisessä ruokavalokuvausprojektissa:
[[poistettu www-osoite]]

Tunniste 4

"Haluan siirtää eteenpäin paikalliskulttuuriin liittyviä kertomuksia, jotka elävöittävät menneisyyttä ja historiaa. Mielestäni se tekee elinympäristössä liikkumisen kiinnostavammaksi. Muistomerkit ym. kertovat tietenkin myös oman tarinansa, mutta ihmisten muistot merkitsevät sitäkin enemmän. Haluan siirtää eteenpäin myös asenteita: erilaiset tavat tekevät elämästä rikasta, kunnioittakaamme siten kulttuurista kirjoa. Ei ole yhtä oikeaa ja ainoaa kultturiperintöä eikä yhtä ja ainoaa oikeaa tapaa välittää kulttuuriperintöä."

Tunniste 5

"Kaikenlaista"

Tunniste 6

"Haluan siirtää eteenpäin kultuuriperinnön koko kirjoa eli sekä positiivia että negatiivisia asioita. Kertomukset, kieli, (vähäinen) tapakultuuri, tiedot ja taidot ovat kultuuriperintöä, jota kansalaiset siirtävät eteenpäin. Fyysisen ympäristön ja materian (esineistö tms), joka liittyy kansalaisten kulttuuriperintöön,säilyttäminen,suojelu ja vaaliminen on ennen kaikkea yhteisöjen kuten museoiden tehtävä."

Tunniste 7

"Alueelliset eri paikkoihin liittyvät tarinat ja henkilöt ovat tärkeää kulttuuriperintöä. Olisi myös hienoa, ettei kaikkea vanhaa pureta pois eli arvostettaisiin ja huolehdittaisiin vanhoista rakennuksista. Kylien raiteilla näkyisi vuosikymmenten eri rakennustyylit - sekä kauniit että kamalat. Lisäksi vahva yhteisöllisyys on hienoa kulttuuriperintöä useassa pienemmässä kylässä - sen toivoisi säilyvän. "

Tunniste 8

"Henkilökohtaisesti toivon erityisesti, että vanha rakennusperintömme säilyy tuleville sukupolville, koska Suomessa vanhaa rakennuskantaa on niin vähän. "

Tunniste 9

"kaikki kultturiperintö sisältää kokemuksia ja tietoa, jota mahdolisesti myös tulevaisuudessa tarvitaan. Lisäksi perintö vahvistaa identiteettiä. Koska kaikkea ei ehkä voi säilöä, on aineellinen perintö oleellista ja silloin myös esineisiin / rakennuksiin yms. liittyvä käyttötieto. Rakennettu kulttuuriperintö sisältää paljon tietoa ja kokemusta sekä työtavoista että rakennusaineista. Lisäksi rakennukset ja rakennelmat vahvistavat paikallista muistia - niihin liittyy tarinoita ja muistuksia, jotka helposti jäävät unholaan, jos mikään ei tarinoista ja ihmisistä muistuta. "

Tunniste 10

"Paikkakunnan lähihistorian säilyttäminen on tärkeää"

Tunniste 11

"Hevoskulttuuria.. hevosurheilukeskus"

Tunniste 12

"Kehittävää, kannustavaa, omaa identiteettiä ja perintöä korostavaa."

Tunniste 13

"Vaasalainen jalkapallokulttuuri on menneisyydestä perittyä ja arvokasta kulttuuria paikallisille, joille Hietalahden stadion on antanut puitteet. Se on kuitenkin katoamassa jos asioille ei tehdä mitään. "

Tunniste 14

"Tasa-arvoisuutta, yhteistä vastuunkantoa, sivistystä, uskoa koulutukseen ja ajattelutavan, jossa asioilla voi olla muuta kuin rahallista arvoa."

Tunniste 15

"Muinaisjäännöksiä ja arkeologisia kohteita, luontoarvoja, Suomen luontoa ja ympäristöä monipuolisesti niin kasvistoa kuin eläimistöäkin, perinnetietoa niihin liittyen, kansanparannustietoa ja tapoja, paikkojen nimiä, ihmisten nimiä. "

Tunniste 16

"Ympäristöä, josta näkyy sen historiallisen juuret ja jota yhteisö arvostaa. Rakennus- ja elämäntapoja, joilla sitä pidetään yllä. Myös muita tapoja, ruokia, uskomuksia, perjantaisaunan ja SUVIVIRREN (vaikka olen ateisti) haluan säilyttää."

Tunniste 17

"Isonvihan aikaiset venäläisten julmuudet Suomessa. Petäjäveden vanha kirkko."

Tunniste 18

"Monipuolista, kirjava katsaus perinnöstämme: arvoja, uskomuksia, tietoa, suullista perimätieto, perinteitä...."

Tunniste 19

"Musiikkia ja vanhoja tapoja ja uskomuksia, koska vaikka aika ja ympäristö muuttuu, ne säilyvät ja antavat paikalle erityisen vivahteen. Vähän niin kuin paikka itsessään olisi puuro ja musiikki, tavat, uskomukset ja tarinat ovat ne sokuri, kaneli ja voisilmä siihen päälle. "

Tunniste 20

"En halua siirtää eteenpäin mitään erikoisempaa kulttuuriperintöä, vaan paremminkin viestittää, että eletään yksinkertaisesti hetkessä ja löydetään jokainen itse oma totuutemme. Menneisyys on menneisyyttä eikä sen tule vaikuttaa siihen mitä me tällä hetkellä tai tulevaisuudessa olemme. Yhteisöllisestä taustastaan on hyvä toki ymmärtää jotain, mutta se on sitten ihan eri asia mitä vaikka uskomusten ja perinteiden siirtäminen eteenpäin. "

Tunniste 21

"Laitakaupungin historiaa, köyhien alueiden menneisyyttä ja perimätietoa siitä."

Tunniste 22

"Mm. kesäasunto, sekä käsityönä tehtyjä käyttö- ja koriste-esineitä. Myös perinteiset rituaalit kuten perunannosto tai talon kunnostus, ja taideartikkelit kuten perinteiset laulut ja lorut, tarinat, tai vaikka valokuvat ovat tärkeitä."

Tunniste 23

"Keksikää joku kansanomaisempi käsite kuin kulttuuriperintö. Se on liian akateeminen ja edellyttää vastaajan olevan koulutettu ja ymmärtävän hiukan hienommin asiat. Tosin kulttuuriperintö- sana ei avaudu mitenkään edes koulutetulle henkilölle. Sitäkö haluatte tälläkin kyselyllä viestiä? Haluan siirtää eteenpäin hyvää toimintaa, sukuni ja omani näköistä."

Tunniste 24

"Vanhoja työtapoja, uskomuksia, perinteitä, tietoja, tietoa vanhoista asuinpaikoista ja kylien vanhimmista taloista/tiloista, sukututkimusta, väestöpohjan rakennetta ajan myötä, RUOKA JA VILLIVIHANNEKSET! Suomen metsissä kasvaa paljon ravitsevaa ruokaa, josta me olemme täysin vieraantuneita. Paluu takaisin auttaisi varmasti painonhallinnassa..."

Tunniste 25

"Yllä mainitut asiat ovat tärkeitä kulttuuriperinteelle ja hyvä säilyttää. Myös luonto ja sen monimuotoisuus ovat tärkeitä niinkuin rakennuksetkin."

Tunniste 26

"arkielämän tapoja, kuten ruuanlaittoa, polttopuiden pilkkomista tms; sitä että arjessa tehdään asioita itse"

Tunniste 27

"Suomalaisia perinteitä, suomalaisen yhteiskunnan arvomaailmaa, myös kristillistä perintöä, joka on Suomessa todella vahva. Mielestäni sitä ei tulisi syrjiä sen vuoksi, että Suomessa on maahanmuuttajia muista kulttuureista, vaan arvostaa myös meidän omaa perintöämme. Haluaisin siirtää eteenpäin myös aineellista kulttuuriperintöä, siinä museot ovat avainasemassa. "

Tunniste 28

"sekä materiaalista, esineellistä että ei-materiaalista mm. suomen kieli, kansantanssit ect"

Tunniste 29

"Rakennusperintöä"

Tunniste 30

"Suomalaista"

Tunniste 31

"Kaikkea vanhaa, ennen 1950-lukua rakennettuja ympäristöjä, koska Suomessa on hävitetty luvattoman paljon kulttuuriperintöämme."

Tunniste 32

"Aineellinen ja Henkinen kulttuuriperintö, kansallisella ja paikallisella tasolla, näiden erot ja samankaltaisuudet. Esim pääsiäisperinteet, virpominen ja trullittelu ovat sulautumassa yhteen, mutta ovat kaksi eri perinnettä. Perinteen ominaisuus on muutos mutta olisi hyvä muistaa aikaisempi perinnetapa, koska se on mahdollista..yleensäkin kaikkien pienten perinnenyanssien tallennus paikallisella tasolla"

Tunniste 33

"Laajasti suomalaista elämäntapaa. Yhtenä osana käsityö, joka viittaa käsityön tekemisen prossessiin ja käsityönä valmsitettuihin tuotteisiin. Käsityötaito on arvokasta osaamista, josta on iloa ja hyötyä jokapäiväisessä elämässä. Käsityö tuottaa hyvinvointia tekijälleen ja myös tuotteiden käyttäjille. On tärkeää siirtää innostus käsityötaitoihin uusille sukupolville. "

Tunniste 34

"Suomalaiset ovat pieneksi kansaksi olleet mukana monessa maailmalla, tästä on todisteena sekä esinekokoelmia että historiallisia lähteitä, mutta tämä näkyy huonosti julkisuudessa. Todella harva tietää esim. suomalaisten tärkeästä roolista Alaskassa Venäjän vallan aikana ja sen ajan vaikutuksista vaikkapa telakkateollisuuteen maassamme. Tässä on mielenkiintoisia yhteyksiä suomalaisten ja vaikkapa Alaskan

alkuperäasukkaiden välillä, sekä mielenkiintoisia elämäntarinoita. Kultturiperintöä pitäisi tuoda esiin siis syvällisemmin: myös esineisiin, kuviin ja tapahtumiin liittyvät tarinat ja tieto ovat tärkeitä. Suomen historia varhaisina aikoina ja vaikkapa yhteydet viikinkeihin kiinnostaa myös, esim. tämä miekkamiehen löytö. Nämä kummatkin aiheet ovat kansainvälisesti tapetilla myös. Mutta tässäkin kaivataan sitä tutkittua tietoam eli sisältö on tärkeä, eikä vaan ne raamit, eli enemmän rahaa tutkimukseen ja asiantuntijoiden palkkaamiseen. "

Tunniste 35

"Turun ja Varsinais-Suomen tarinaa"

Tunniste 36

"Klassinen ja kansanmusiikki, uskonto, käden taidot, maalaustaide, rakentaminen, luonto"

Tunniste 37

"Eurooppalaiset kulttuuriset juuremme: kristinusko, analyyttinen filosofia (kriittinen ajattelu), keskusteleva demokratia, suvaitsevaisuus ja toinen toisesta välittäminen ovat kulttuuriperintömme vaalittavia asioita. Myös moderni taide ja rakennuskanta, menneitä teknologisia käytänteitä koskevan tiedon osaamisen säilyttäminen ja ylläpito ovat periaatteessa asioita, joita tulisi jotenkin siirtää eteenpäin."

Tunniste 38

"Kansanmusiikki, kansantaide, perinteinen puurakentaminen eri ilmenemismuodoissa, puistot ja puutarhat...."

Tunniste 39

"Arkielämään, esineisiin, rakennettuihin miljöisiin, rakennuksiin, kieleen, kirjalliseen, suulliseen ja visuaaliseen perintöön jaoteltavaa, taitoja, arkistomateriaalia, käyännön elämää ja kulttuurihistoriaa."

Tunniste 40

"Kulttuuriperintömme on aika pitkälti näkyvissä rakennuskannassa. Toivoisin säilyttävämpää linjaa rakennusten pysymiseen "vanhoina". En tarkoita museoimista vaan asuttujenkin rakennusten korjaamisessa vaaditaan energiatehokkuutta ja uusien rakennusmääräysten noudattamista. Miksi vanhoista rakennuksista pitäisi tehdä uusia? "

Tunniste 41

"Suomalaista kulttuuria, luontoa, luottamusta, puhtauta. Arvostusta omaan perimään ja kieleen."

Tunniste 42

"Puhdasta vesistömaisemaa, monimuotoista metsämaisemaa, kauniita puistoja eri-ikäisine puineen, eri-ikäisiä hoidettuja rakennettuja ympäristöjä. "

Tunniste 43

"Naisten kulttuurista perintöä on vähätelty koska taiteena ja kulttuurina on arvostettu lähinnä miesten kontribuutioita. Naisten käsitöissä kautta maailman on yhteisiä ekososiaalisen keskinäisriippuvaisuuden

elämänpuun ja sukupolvien ja lajien ketjun narratiiveja ja kuvioita, motiiveja ja symboleja. Niiden tutkimusta tulisi rahoittaa ja edistää sillä ne palauttavat meidän ei vain suomalaisen vaan maailmanlaajuisen elämänmyönteisen ekososiaalisesti kestävän menneisyyden parhaiden käytäntöjen äärelle. Suomalainen kansanrunous pitää sisällään paljon viisautta mutta tarvitaan laajempaa tutkimusta ja näkyväksi nostamista kuin iänikuinen Kalevala miessankareineen. Tarvitsemme kaikkialle ekomyyttejä, joilla lapset saataisiin omaksumaan koko planeetan kannalta ekosysteemikeskeisiä arvoja ja vastuuta. Niitä on kerättävä ja tehtävä esimerkeiksi kestävästä elämäntavasta sotien valloituksen ja luonnonherruuden sijaan. "

Tunniste 44

"Haluan siirtää eteenpäin aineellista ja aineetonta kulttuuriperintöä. Itseäni ja työtäni lähellä ovat taiteen, käsityön ja koko kansankulttuurin arvot, unohtamatta luontoon liittyvää kulttuuriperintöä. "

Tunniste 45

"Käsityökulttuuria"

Tunniste 46

"Luonnon, esivanhempien työn ja yhteisöllisyyden arvostamisen kulttuuriperintöä"

Tunniste 47

"Haluan siirtää eteenpäin mahdollisimman moniäänistä ja monimuotoista kulttuuriperintöä. Erityisesti muut kuin tämän hetken vallitsevat äänet ansaitsevat huomiota, sillä niiden säilyminen on epävarmempaa. Esimerkiksi saamelaisten kulttuuriperintöä tulee pyrkiä säilyttämään."

Tunniste 48

"vanhaa suomalaisuutta, Suomen oloihin vaikuttaneita ilmiöitä, elävää nykyoloa"

Tunniste 49

"Ylipäätänsä haluaisin olla lenkkinä kulttuuriperinnön viemisessä uusille sukupolville. Erityisesti minua kinnostaa kaupunkikulttuuri ja rakennettu ympäristö, erilaiset, eri aikakausien rakennukset ja ympäristökokonaisuudet. Kirjallisuus ja musiikki ovat lähellä omaa sydäntä- suomalaista kirjallista ja musiikillista perintöä."

Tunniste 50

"Arkkitehtuuria, tarinoita kirjallisuuden, musiikin, arjen perimätiedon, kirkon, median, arkistomateriaalin, museoiden kautta. Hyvä esimerkki inkeriläistiedon siirtymisestä meille jälkipolville on mittava 1920-luvulta saakka kertynyt arkistomateriaali, jota voi kuka tahansa pyytää arkistoista, selata verkosta perimätietoa oman vuosikymmeniä ilmestyneen lehden sivuilta. Suvussa tieto ei siirtynyt, vaikka yritystä oli mutta aihe oli yksinkertaisesti liian vaikea ja koettiin jopa vaaralliseksi. Onneksi arkistolaitos ja aktiivit ovat huolehtineet, että me jälkipolvet pääsemme tutustumaan papereihin nyt kun ihmisistä on aika jättänyt."

Tunniste 51

"Aitoa kulttuuriperintöä uusvanhan kitsin sijaan. Edellinen määrittyy tuon perinnön historian kautta ja siihen kuuluu myös modernismi, eikä sitä voi erikseen määritellä erossa tuosta perinteestä ja sen konkretiasta. "

Tunniste 52

"Lippalioskeja ja Kivijalkakauppoja."

Tunniste 53

"Pääkaupunkiseudulla on kaikki ympäristö rakennettua ympäristöä joka sekin on jatkuvassa vaarassa tuhoutua. Haluaisin että kaikki yli 100-vuotta vanhat kulttuurikerrostumat kartoitetaan ja jokaiselle tehdään hoitosuunnitelma ja määritelmä miten se saadaan säilytettyä tai ennallistettua jälkipolville?"

Tunniste 54

"Rakennettua kulttuuriperintöä: asuinsijat, viljely- ja teollisuusympäristöjä, liikenteen väyliä (tiet, polut, kanavat, joet, satama, venepoukamat); suomalaiseen muinaisuskoon ja luonnonuskoon liittyviä paikkoja (esim. uhrikivet, lehdot, karsikot, kalliomaalaukset), arkeologisia jäänteitä, muinaisia asuinpaikkoja, tämän päivän rakennuskulttuuria, kansanviisautta, puutietoutta - taitoa elää ilman sähköä ja öljypohjaisia polttoaineita. "

Tunniste 55

"Paikalliskulttuurien erityislaatuisuus niin kielen, osaamisen kuin aineellisenkin perinteen osalta. Ennen toista maailmansotaa tehdyt rakennukset ja erityisesti kansanrakentamiseen liittyvän osaamisen."

Tunniste 56

"Rakennusperinnön korjaamisen aineellista ja aineetonta perintöä. Esimerkiksi ylläpitämällä pienten käsityöalojen ammatti- ja erikoisammattitutkintoja (esim. kultaajakisälli- ja kultaajamestari)."

Tunniste 57

"Kalliomaalaukset, aarnimetsät, puhtaat rannat, kansanuskomukset, marja- ja sienipaikat, hienot hietikot, naavametsiköt, vanhat rakennukset."

Tunniste 58

"Kykyä muutokseen yhteistyössä toisten kanssa."

Tunniste 59

"Saunakulttuurin monimuotoisuudessaan, kyläkoulut ja -yhteisöt, luonnonlääkintää, kansantarinoita, omat kansalliset perinteet (joulu, juhannus, pääsiäinen) kunniaan ilman että niiden tilalle tuodaan jenkkiläistä hapatusta (halloween tai vieraat juhlaperinteet)"

Tunniste 60

"Talkoohenkeä ja yhdessä tekemistä. Ei aina voida piiloutua rahan puutteen taakse, joskus pitää myös kääriä hihat ja liata kädet. Toisaalla se on vielä voimissaan, toisaalla kadonnutta kulttuuriperintöä. "

Tunniste 61

"Arkipäivän ja tavallisten ihmisten kokemuksia. "

Tunniste 62

"Maaseutumaisen asumisen kulttuuriperintöä. Työn tekemisen kulttuuria ja arvostusta. Ruokakulttuuria. Tarinoiden ja paikallishistorian tallentaminen. Yhteisöllisyys, joka on katoamassa, kun yksityisyys valtaa alaa myös maaseudulla. Uskoa ja voimia siihen, että maa tuottaa elannon perheelle. Maanviljely on muokannut maisemaa, avaraa peltomaisemaa eli kulttuurimaisemaa. Tietoa, taitoa ja nöyryyttä luonnon ja säiden kunnioittamiseen. Suomalaisten juuret ovat maaseudulla, kaupunkikulttuuri on nuorempaa kulttuuria. Myös tätä vanhaa kulttuuria tulee arvostaa."

Tunniste 63

"Kädentaidot, esim. kangaspuilla kutominen, ovat mielestäni kaikki erittäin tärkeitä eteenpäin siirrettäviä taitoja. Yhteiskuntaa ei voi pyörittää pelkillä kylmillä (tieto)koneilla."

Tunniste 64

"Omaa alueellista kulttuurihistoriaa"

Tunniste 65

"Suomalaisia rituaaleja ja seremonioita, myös niitä, jotka eivät ole minullekaan siirtyneet. Vanhoista uskomuksista juontavia tapoja, jotka eivät rajoita nykyihmisen elämää mutta rikastuttavat sitä antamalla kokemuksen omista juurista. Kaikista tärkeintä on kuitenkin kansanmusiikki ja tanssi kaikissa muodoissaan. Tätä on poljettu alas niin kauan, että minä,olen musiikkitaustasta huolimatta vasta kolmikymppisenä saanut tietää, mitä aito suomalainen kansanmusiikki on nykypäivänä. Tätä ei pitäisi enää piilotella, vaan tuoda esiin positiivisella ja viihdyttävällä tavalla kaikissa sopivisa yhteyksissä. Kansanmusiikin on tarkoitus olla hauskaa yhdessä tekemistä, johon voi jokainen jollain tavalla osallistua. "

Tunniste 66

"Eri ihmisillä ja kansalaisryhmillä on erilaisia kulttuuriperintöön liittyviä arvoja. Haluan säilyttää monimuotoisuuden ja edistää niin juhlan kuin arjen perinteiden tallentamista ja hyväksikäyttöä uusissa muodoissa. "

Tunniste 67

"Kulttuuriperintöni on sukuni ja perheeni tarina, aineellisesti köyhä, mutta rikas selviytymistarina autonomian ajoista tähän päivään. Siihen sisältyy "muistojen Karjala", ei kliseinen hössöttävä, vaan se tummempi raita, jossa suvaitaan erilaisuutta, ei vihata eikä syytetä venäläisyyttä , ei mitään kansanryhmää. Otettiin kulkijat vastaan ja tarjottiin ruokaa ja yösijaa. Luonnon kauneus ja antimet merkitsevät paljon. Vanhoissa valokuvissamme on ihmisillä käsissään tuomenkukka, koivunoksa, soitin. Ymmärrys siitä, että sietämättömissä olosuhteissa on jokaisella maailman ihmisellä oikeus lähteä etsimään parempaa elämää ja tulla ystävällisesti vastaanotetuksi."

Tunniste 68

"Merkityksellistä, kestävää, moniarvoista ja käytettävää. Toivon että kulttuuriperintöä hedelmällisesti avaava tieteellinen keskustelu (ks. esim. uudet Muuttuva kulttuuriperintö ja Mitä on kulttuuriperintö? -julkaisut; Laurajane Smith, Uses of Heritage) rantautuisi myös kulttuuriperintötoimijoiden arkeen, ja kulttuuriperintö nähtäisiin enemmän prosessina ja merkityksinä kuin fyysisinä kappaleina, museoesineinä ja monumentteina. Smithin mukaan kaikki kulttuuriperintö on aineetonta, mutta sen esiin tuominen vaatii esim. suomalaisilta museoilta aiempaa parempaa kokoelmahallintaa ja kokoelmien merkitysten analysointia ja esiin tuomista."

Tunniste 69

"Kulttuuri-ilmiöt muuttuvat ajassa, paikassa ja sosiaalisessa/ekologisessa kontekstissa. On myös kulttuuriperintöä, joka kuuluu museoon, kuten noitavainot, eläinten kevätmetsästys, nykyinen teollinen suhtautuminen tuotantoeläimiin, patriarkaatti, homofobia jne."

Tunniste 70

"Omia perheen perinteitä ja yleisiä paikallista ja kansallisia kultuuriperinteitä, tapoja, tarinoita ja uskomuksia. Maailma on rikas kulttuureiltaan ja erilaisuus viehättäää ja ruokkii mielikuvitusta sekä tuo ymmärrystä ja edistä tasavertaisuutta ja siten oikeudenmukaisuutta globaalisti. Kun tuntee omat juurensa, voi suhteuttaa siihen maailmalta oppimaansa, verrata ja parastaa omaa elävää kulttuuriperintlöä. Ja se että ihminen kykenee sopeutumaan täysin erilaiseen kulttuuriin ja ilmastoon, kertoo siitä että ihminen voi muuttua ja ymmärtää, hyväksyä erilaisuuden. "

Tunniste 71

"Kaikkia mainittuja niin, että kulttuuriperintöä ei voi kaapata tukemaan nationalistista ja "muut" poissulkevaa politiikkaa vaan päin vastoin se kertoo siitä, miten ihmiset ja ideat ovat olleet vuorovaikutuksessa nykyisten valtioiden rajojen yli."

Tunniste 72

"Ruokaperinnettä ja murretta, rakennetun kulttuuriperinnön arvostamista. "

Tunniste 73

"Kulttuuriperinnön siirron tulisi sisältää tasapuolisesti aineettomia ja aineellisia voimavaroja. Kulttuurievoluutiosta ajatuksia lainaten: Viimeisen tiedon (Juha Valste, Ihmislajin synty, 2012) mukaan esi-isämme olivat vähällä hävitä sukupuuttoon 70 000-120 000 vuotta sitten. Kaikki maailman yli seitsemän miljardia ihmistä ovat yhden ainoan tuolloin Afrikassa eläneen väestön jälkeläisiä. Siihen kuului noin 500 lisääntyvää naista, ja yhteensä yksilöitä tässä väestössä arvellaan olleen noin 2 000. Eri väestöihin kuuluvien ihmisten väliset geenierot ovat pienempiä kuin juuri millään toisella suvullisesti lisääntyvällä ja kohtuullisen runsaalla eläinlajilla. Suomalaisen maanviljelijän, melanesialaisen helmenpyytäjän, kongolaisen pygmin ja Perun Andeilla laamoja kasvattavan intiaanin geenit eroavat toisistaan vähemmän kuin esimerkiksi Helsingin Kruunuhaassa, Tukholman Bandhagenissa ja Hampurin satamassa elävien kotihiirien geenit toisistaan. Kulttuurievoluutio ei perustu geeneihin samalla tavalla kuin biologinen evoluutio. Sille on luonteenomaista hankittujen ominaisuuksien periytyminen ja informaation määrän

kasvu sukupolvi sukupolvelta. Ominaisuudet siirtyvät sekä pystysuunnassa sukupolvelta seuraavalle että vaakasuunnassa samaan aikaan eläviltä yksilöiltä toisille. Ihmisen kulttuurievoluutiolle on luonteenomaista sen nopeuden lisääntyminen. Mitä enemmän informaatiota on kertynyt, sitä vauhdikkaammin sitä kertyy lisää. Kulttuurievoluution myötä ihmiselle on kehittynyt tapoja säilyttää ja siirtää informaatiota: puhuttu symboleja käyttävä kieli, suullinen perimätieto, kirjoitustaito, kirjapainon kehittäminen, sähköinen tiedonvälitys, tietokoneet ja langaton osaksi vuorovaikutteinen tietoverkko. Kulttuurievoluution tuloksena informaatio kumuloituu. Tämä johtaa siihen, että yksilöt pystyvät hallitsemaan valtavasta tiedon määrästä vain pieniä ja koko ajan pieneneviä osia. Kulttuurin avulla ihminen pystyy sopeutumaan elinympäristön muutoksiin tai elämiseen uudenlaisssa elinympäristöissä. Kulttuurievoluutiota tapahtuu koko ajan. Ensimmäinen selvä ero nykyihmisen ja muiden ihmistyyppien välillä on erilaisten abstraktia ajattelua osoittavien koristeiden ja merkkien ilmestyminen. Afrikasta on tavattu simpukankuorista, luusta ja strutsinmunankuoren palasista valmistetuja helmiä, jotka ovat 120 000 vuotta vanhoja. Jostakin syystä 60 000-50 000 vuotta sitten erilaiset korut alkoivat alkoivat nopeasti yleistyä. Tätä on kutsuttu ihmisen kulttuurin ”alkuräjähdykseksi”. Samoihin aikoihin tehtiin vanhimmat tunnetut kalliopiirrokset Australiassa ja asuinpaikoista alkoi löytyä yhä enemmän merkkejä myös värien käytöstä. Euroopan vanhimmat luolamaalaukset ovat noin 34 000 vuoden ikäisiä."

Tunniste 74

"Maisema,- ja rakennusperinnön"

Tunniste 75

"Ensimmäisenä tulee mieleen luonto. Sitten kysymys, "niin, onko luonto kulttuuriperintöä", vastaus: kyllä minusta. Se on osa suomalaista kulttuuria, sitä missä elämme, mistä inspiroidumme, mistä elämme. Haluan siirtää eteenpäin suomalaista taidetta, musiikkia, sitä miten musiikkia opetetaan ja opiskellaan (esim suomalaiset musiikkileirit!), maisemia, koskematonta kalliota meren rannalla, kirkasta vettä keskisuomalaisessa järvessä ja veneilyä siellä, arkeologisia löytöjä ja muinaisia paikkoja (Sysmän kulttuurimaisemat tulevat ekana mieleen)... "

Tunniste 76

"Paikkoihin ja maisemiin liittyvät tarinat ja kertomukset ovat olennainen osa kulttuuriperintöä. Ilman paikan kerrotun menneisyyden tuntemusta, sen merkitys, ainakaan sen laajuus, ei avaudu. Paikkoihin ja rakennuksiin liittyvät kertomukset, niin suulliset kuin kirjoitetut, ovatkin olennainen osa kulttuuriperintöä."

Tunniste 77

"Hyvin laaja-alaisesti erilaista kulttuuriperintöä, jossa näkyvät myös vähemmän tunnetut, piilossa olevat ja ehkä osin jo unohdetut aspektit. Muutakin kuin suomalaiskansallisesti "kanonisoitua" perinnettä. Tietoa myös ikävistä asioista historiassamme. Perinnön monimuotoisuutta ja sen arvostusta. Perinteisten menetelmien ja kulttuuriperinnön hoidon osaamista. "

Tunniste 78

"Minulle sydäntä lähellä ovat perinteiset käsityötavat, kansanmusiikki, kansallispuvut, kansantanssi ja kaikenlainen kansanperinne. Koen tärkeäksi myös muunlaisen positiivisen perinteen säilyttämisen. Toisaalta kaunaa ja luutuneita tapoja haluaisin aktiivisesti unohtaa."

Tunniste 79

"Haluan siirtää kestäviä, muuttuviinkin kulttuuritottumuksiin sopeutuvia voimavaroja, tietotaitoa ja perinteitä sukupolvelta toiselle. Museoihin ja kokoelmiin tallentunut otos on edelleen vajavainen arkiseen tietoyhteiskuntaan liittyneen materiaalisen perinnön ja siihen liittyvien käyttökokemusten osalta. 1900-luvun jälkipuoliskoa ja 2000-luvun ja 2010-luvun matriaalivarantoa pitäisi selventää ja päättää mitä säilytetään. Ilmaston lämpeneminen muuttaa materiaalista maailmaa ja perinteitä, samoin sodat ja siihen liittyvä perinteiden turmeleminen sekä luonnonvarojen käyttö ja tietotekniikka joka jää harvojen etuoikeudeksi ellei ... tuleville polville on voitava jäädä edes tieto siitä mitä menetetään ja miten niihin voidaan luoda side kustakin kulttuurisesta kehitysvaiheesta."

Tunniste 80

"Uskomuksia ja niiden taustoja -miksi perinteet ovat sellaisia kuin ovat. Ihan konkreettisia perinteita ja niiden takana olevia arvoja."

Tunniste 81

"Yleisesti näkisin, että Suomessa tulisi kiinnittää erityisesti huomio aineettoman kulttuuriperinnön siirtämisen edistämiseen. Aineellisen perinnön tallentamisen, tutkimuksen ja siirtämisen saralla on edistytty huimasti viime vuosien ja vuosikymmenten aikana. Erityisesti tulisi monipuolisesti juhlistaa aineetonta ja koko ajan uudistuvaa monikulttuurista aineetonta perintöä. Kiinnostus olisi herätettävä kouluissa ja oppilaitoksissa. Ylisukupolvinen dialogi eri ryhmien välillä voi tehdä aineettomasta perinnöstä elävää, tämä tarkoittaisi käytännössä niin paikallisen kuin kansallisten toimijoiden osaamisen hyödyntämisen opetuksessa. Kansalaisjärjestöjen ja ruohonjuuritason toimijoiden näkyvyyttä kouluissa tulisi vahvistaa. Rinnakkaiset perinteet pitäisi tuoda samanaikaisesti ja tasavertaisesti esille. "

Tunniste 82

"Rakennusperintöä, kulttuurimaisemaa, tapoja, tottumuksia, uskomuksia eri ajoilta, 'kouluperintöä' (laulut, leikit, opettamistavat, koululaisena oleminen), arkipäivän elämärytmiä ja siihen liittyviä asioita maalla/kaupungissa..., vapaa-aikaan liittyviä toimintoja maalla/kaupungissa, vieraista kulttuureista omaksuttuja asioita, tarinaperinnettä, ruokakulttuuria"

Tunniste 83

"Kaikenlaista. Jos esim. difficult heritage näkökulma uupuu kansallisen kulttuuriperinnön käsittelystä, ollaan vääristyneen kansallis-identiteettin ja historian vääristelyn asialla. Kulttuuriperinnön tulee olla terveen identiteetin ja kansallistunteen rakentaja. Tällöin kaikella kulttuuriperinnöllä on sijansa. Kulttuuriperintö elää ja siirtyy formaalin kentän (esim. museot) ohella arjessa. Arjessa siirtyvän kulttuuriperinnön ei tule "alistua" kenenkään "halulle" siirtää tai olla siirtämättä sitä eteenpäin. Kulttuuriperinnön tulee antaa elää ja hengittää vapaasti. Se täytyy kuitenkin alistaa terveelle itsereflektiolle:

kyvylle ymmärtää esim. menneisyyden kontekstia (vaikkapa tiettyyn aikaan liittyviä kansallisia haasteita kun vaikkapa uutta kansakuntaa ja sen kansallisidentiteettiä synnytetään). Kulttuuriperinnön moninaisuus tulisi huomioida tällä hetkellä korostetusti. Ajassamme elävä käsitys "puhtaasta suomalaisuudesta" tulisi ohjata näkemään kansallisidentiteettimme monikulttuuriset juuret. "

Tunniste 84

"Maisemien, rakennusten, tarinoiden ja musiikkiperinnön vaaliminen on merkityksellistä jälkipolville."

Tunniste 85

"Suomalainen kulttuuripeirintö on laaja ja moninainen. Siihen kuuluvat kansalliset, alueelliset, kulttuuriset ja kielelliset osiot ja niitä pitäisi aina tarkastella, päivittää ja suhteuttaa myös monikulttuuriseen maailmaan."

Tunniste 86

"Ymmärryksen siitä, että arvot, uskomukset ja perinteet, tietokin, muuttuu aikojen kuluessa. Niiden suhteellisuuden voi ymmärtää vain ymmärtämällä, että Suomessakin on joskus ajateltu eri tavoin, ja tullaan ajattelemaan vielä enemmän eri tavoin tulevaisuudessa. Miten tämä kaikki näkyy meidän nykyisessä fyysisessä ja henkisessä maailmassamme, on tärkeää hahmottaa."

Tunniste 87

"Tässä tuhansien järvien ja pitkän merirantaviivan maassa, itse saaressa asuvana, haluan nostaa framille ja siirtää seuraaville sukupolville luonnonvesissä uimisen. Joissain kulttuureissa ja joillain kielialueilla sitä kutsutaan "villiuimiseksi", mutta meille suomalaisille sen pitäisi olla arkipäivää (tosin juhlavaa sellaista) ja itsestäänselvyys varsinkin kesällä, mutta mielellään myös talvella."

Tunniste 88

"Suomalaista ja suomenruotsalaista kulttuuriperintöä. "

Tunniste 89

"Käsityötaitoja, -tekniikoita, -malleja, tuotteita, tekijöiden/yhteisöjen tekemis- ja tuotehistoriaa"

Tunniste 90

"kädentaitoja, luonnon kiertokulkuun liittyviä sanontoja, kristillisiä perinteitä"

Tunniste 91

"Realistista kuva ja videomateriaalia sekä sanallisia kertomuksia (ääninäytteitä, tekstejä) nykyajasta eri paikoista ja tilanteista, arkisistakin: esim. päiväkodin arkea, tien päällystysurakointia, opettaja ja oppilaat työssään luokassa, toimistotyön arkea, rakennustyömaan toimintaa, toimintaa satamassa ja lentokentällä, auto korjaamossa, kaupan kassa, taksinkuljettaja työssään ja uutistoimittaja työssään."

Tunniste 92

"Tarinoita, satuja, kansanperintöä, käsitöitä ja niiden ohjeita."

Tunniste 93

"juuri tuon määritelmän mukaista perintöä"

Tunniste 94

"Perinnekäsitöitä (langan kehräys, kasvivärjäys, kankaankudonta, tuohityöt yms), vanhat tarinat ja uskomukset, lääkintäyrttien käyttö. Luonnossa liikkuminen ja suunnistus/erätaidot. Perinneruoat, esim. karjalanpiirakoiden leivonta."

Tunniste 95

"Yleisesti lienee kestävintä "siirtää" tai ennemminkin uudelleentulkita kulttuuriperintöä mahdollisimman laaja-alaisesti tuhoamatta materiaalista kulttuuria tai katkaisematta jatkumoita silloin kun se ei ole välttämätöntä. Omakohtaisesti haluan välittää eteenpäin minua henkilökohtaisesti koskettavaa tietoa ja ohjata esimerkiksi lastani lukemaan kriittisesti institutionaalista kulttuuriperintöä."

Tunniste 96

"Vanhoja rakennuksia eli arkkitehtuuria, kirkkoja, arvokkaita käsitöitä (myös kotitekoisia!) ja taidetta, lauluja, sanontoja, suvun tarinoita, perinteitä, valokuvia suvusta, käsityötaitoja (kangaskäsityöt, lankakäsityöt, punamultamaalin keitto, pärekaton tekeminen, päreiden höylääminen, hirsitalon kunnossapito,...), suvun perinteiset marjapaikat, suvun perinteiset leivontareseptit, hyvä suomen kielitaito ja sanasto,..."

Tunniste 97

"Haluaisin siirtää eteenpäin sellaista kulttuuriperintöä joka auttaa tulevia sukupolvia maadoittumaan oman alueensa menneisyyteen. Tavoite olisi oppia arvostamaan ja kunnioittamaan menneiden sukupolvien aikaansaannoksia siten että oma paikka maailmassa tuntuisi merkitykselliseltä. Lapset oppisivat olemaan osa sukupolvien ketjua ja ymmärtäisivät olevansa yksi lenkki pitkässä ketjussa. Tavoite olisi hiljalleen kasvattaa asiassa globaali näkökulma ja ymmärtää että me ihmiset jaamme saman planeetan. "

Tunniste 98

"Erityisesti aineellista ja kaikkien ympärillä olevaa rakennettua ympäristöä, arjen kulttuuriperintöä"

Tunniste 99

"Suomalaista tasa-arvoisuuden perinnettä. Harkitsevaisuutta, kohtuullisuutta ja jalat maassa -asennetta. Läheistä suhdetta luontoon. Omaa kieltä. Materiaalista kulttuuria niin, ettei kaikkea vanhaa pureta. Juhlatapoja, vaikkei välttämättä uskonnollisessa mielessä koska en ole kristitty. Tärkeimmät kristinuskon periaatteet ja tarinat pitäisi kuitenkin olla jokaiselle tuttuja, samoin se mitä suomalaisesta muinaisuskosta vielä tiedetään. "

Tunniste 100

"Sellaista, joka kertoo kansan ja alueen historiasta, ei pelkästään suurmiesten. Ei pelkästään hyvää ja kaunista."

Tunniste 101

"Kielen, perinteiset työtavat, perinnemaisemat, rakennuskulttuurin edustavat näytteet alueellisesti eri aikakausilta"

Tunniste 102

"Muinaisia uskomuksia henkiolennoista ja muista vastaavista, ennustamista ja taikoja"

Tunniste 103

"Muiden ihmisten erilaisuuden hyväksymistä ja kunniottamista. Kohteliasta käytöstä. Itsenäistä ja kriittistäkin ajattelua. Tasapuolisuutta ja tasa-arvoisuutta. Oman itsensä hyväksymistä ja arvostamista. Historian ja kulttuurin tuntemusta."

Tunniste 104

"Elinvoimaisen maiseman, tarinat, kauniin rakennusperinnön. Tärkeintä on kuitenkin suhde luontoon ja taito selviytyä näissä olosuhteissa."

Tunniste 105

"Arkistoammattilaisena pistää silmään, että esineellinen kulttuuri mainitaan ja tuodaan esille ja museot nimetään niiden säilyttäjätahoksi, mutta vastaavasti loogisesti kaksi muuta muistiorganisaatiota, joilla on vastaava kulttuuriperinnön säilytystehtävä, jäävät mainitsematta. Toivottavasti valmistelussa otetaan tasapuolisuuden ja kulttuuriperinnön kokonaisuuden vuoksi huomioon myös arkistot ja kirjastot toimijoina."

Tunniste 106

"Käsityötaitoja ja perinteitä."

Tunniste 107

"Olen yrittänyt siirtää omilla lapsille tietoisuutta "suomalaisesta kultturiperinteestä". Harrastan itse kansantanssia ja -musiikkia. Tämän maailman pojat ovat perinneet äidinmaidossaan. Vaikka tiedän, että tämä osa kulttuurista ei ole "muodissa" yritän saada heidät ymmärtämään vanhaa tanssi- ja musiikkiperinnettä. Lisäksi pyrin kotona ylläpitämään ainakin jonkilaista vanhaa juhlaperinnettä. Minusta on tärkeää, että lapset eivät häpeä omaa kansanperinnettä. Esiintymismatkat ulkomaille ovat lisänneet lasten kunnioitusta suomalaista perinnettä kohtaa, sillä suomalaiset kansallispuvut, tanssit ja musiikki saavat ulkomailla yleensä suuren suosion ja kiinnostavat katsojia. Vain me suomalaiset häpeämme kauniita pukujamme ja tanssejamme."

Tunniste 108

"Kaikenlaista tasapuolisesti. Erityisesti haluan, että näkyvät kohteet säilyisivät ja hiljalleen tuhoutuvista kohteista olisi tieto mahdollisimman hyvin tallessa."

Tunniste 109

"ennen kaikkea käsityökulttuuria, myös omavaraisen ruuan tuotannon perinne on tärkeää"

Tunniste 110

"Kulttuuriperintöä omasta ajastani ja varhaisemmista, ilmiöitä, aineistoja kulttuuriympäristöä, tietoa, tarinoita, jotka kuvastavat olennaisia piirteitä kustakin ajasta/ihmisen toiminnasta. "

Tunniste 111

"Suomessa asuvien suomalaisten, joihin maahan eri aikoina muuttaneet myös luonnollisesti kuuluvat, tapojen, yhteisömuodostuksen, estetiikan ja arvostusten, uskonnon tai uskonnottomuuden tiedostamista. Myös edellämainittujen muutoksen tiedostamista ja arvostamista."

Tunniste 112

"Suomalainen, demokraattinen metsästys-eräkulttuuri."

Tunniste 113

"Suomi on valtiona nuori ja sen kulttuurinen menneisyys on ollut enemmän hävittävää kuin säilyttävää. Sanotaan, että kansallisen identiteetin vahvistuminen tapahtui vasta toisen maailmansodan aikana. Vahva usko teknologian edistyksellisyyteen on myös vaikuttanut siihen, että uusinta teknologiaa on myös pidetty parempana kuin vanhaa. Suomalaiseen modernismiin ei siihenkään liity vahvaa perinteiden ymmärrystä. Haluan edistää kulttuuria, jossa vanhan ymmärtäminen on edellytys uuden luomiselle. Jossa menneisyys ja nykyisyys yhdistyvät toisiaan kunnioittavalla tavalla "

Tunniste 114

"Tietoja ja perinteitä. Esimerkiksi Suvivirren esittämistä koulujen kevätjuhlissa. "

Tunniste 115

"Elävän laulun kulttuuria! Kalevalamittainen runous ei ole tarkoitettu vain kirjoissa säilöttäväksi, vaan haluamme saada suomalaiset laulamaan sitä uudelleen ja luomaan sen pohjalta myös uutta runoutta, vaikkapa räppiä. Kalevala-mitta ei ole kuollut! Se elää suomen kielen rytmissä, nimissämme, fraaseissa, joita käytämme. Hyvän itsetuntemuksen kulttuuriperintöä, aineetonta kulttuuria, joka elää sielussa, juurissa, myönteisessä minäkuvassa. Se voi olla fiktiota, faktaa tai taidetta, mutta sillä on ajallinen kytkös. Se kumpuaa menneestä ja elää vuropuhelua tulevaisuuden kanssa."

Tunniste 116

"Taitoja: miten tehdään saunavihta tai karjalanpiirakoita. Tietoja: Miten ennen on eletty ja millaisia tarinoita eri paikoilla on. Ymmärrystä: Miten menneisyys vaikuttaa nykypäivään. Paikkoja ja rakennuksia: vanhaa pitää säilyttää."

Tunniste 117

"Tärkeitä, arvokkaista ja moniarvoisia. "

Tunniste 118

"Haluan siirtää sellaista kulttuuriperintöä, joka rauhoittaa ja eheyttää sekä tekee elämän kokonaisemmaksi."

Tunniste 119

"Vaikka elämme kansainvälisessä maailmassa, haluaisin, että suomalaisen kulttuurin pohja musiikkeineen, tansseineen, loitsuineen, ruokineen olisi paremmin ihmisten tietoisuudessa. Se, mistä suomalaisuus on ponnistanut vuosisatoja sitten, vaikka on oltu milloin minkäkin vallan alla, tulisi pitää arvokkaana aarrearkkuna, josta ammentaa ja kertoa tuleville sukupolville."