Yleiskatsaus nimien vertailuun

Kuvittele, että sinun vastuullasi on selvittää, vastaako jonkin asiakkaasi, toimittajasi tai muun sidosryhmässäsi olevan tahon nimi pakotelistalla olevia nimiä. Tehdäksesi tämän tarkasti, tarvitset ohjelmiston koska nimiä on tuhansia. Käyttämäsi ohjelmiston tulee olla tehokas ja joustava, koska nimet voivat sisältää muunnelmia, kirjoitusvirheitä tai muita puutteita.

DOKS Sanctions API

DOKS Sanctions API on rajapintapalvelu, jonka avulla nimien vertailu pakotelistohin sujuu helposti ja luotettavasti. Ratkaisu perustuu kahden nimen väliseen pisteytykseen, prosenttilukuun.

DOKS Sanctions API on helposti liitettävissö sellaisenaan ulkoisiin tietojärjestelmiin. Se on myös sisäisesti käytössä DOKSin käyttöliittymäversiossa.

Miksi pisteytys? Miksi ei vain kyllä tai ei?

Kun yrität selvittää, esiintyykö jokin listallasi oleva nimi pakotelistoilla, oletetaan usein, että ohjelmisto palauttaa ”kyllä tai ei” ilman muita tietoja. Jos listasi on lyhyt ja riski saada se väärin on pieni, saatat olla tyytyväinen tähän tulokseen ja toimia sen mukaisesti.

Mutta pakotelistat ovat pitkiä ja näissä tapauksissa myös virheestä koituva riski on suuri.

Jos saisit vastaukseksi vain ”kyllä”, alat ehkä miettimään onko tämä sittenkään oikein. Ethän tietenkään halua alkaa isoihin toimenpiteisiin ilman tarkempia tietoja. Saatat miettiä, että ehkä käyttämässäsi nimessä oli jokin virhe.

Jos taas saisit vastaukseksi vain ”ei”, voit jäädä pohtimaan, että mitä jos antamasi nimi onkin kirjoitettu eri tavalla tai sisältää kirjoitusvirheen. Puuttuikohan sukunimi? Entä jos etunimet olivatkin väärässä järjestyksessä?

Pisteytyksen käytön perusteena on se, että nimien vertailu ja täsmäytys on erittäin vaikeaa!

Nimet ovat harvoin täsmällisiä
Nimien vertailussa joudutaan aina pohtimaan seuraavia tapauksia ja niiden aiheuttamia toimenpiteitä:

1. Todelliset positiiviset osumat
2. Todelliset negatiiviset osumat
3. Väärät positiiviset osumat. Tapaukset jotka olet valmis tarkistamaan manuaalisesti.
4. Väärät negatiiviset osumat. Tapaukset jotka olet valmis jättämään huomioimatta.

Nimien vertailussa kaikki nämä neljä osa-aluetta ovat aina läsnä. Laadukkaat ohjelmistot kuitenkin pystyvät maksimoimaan tapaukset 1-2 ja vastaavasti minimoimaan tapaukset 3-4. Joissakin tapauksissa voidaan sietää enemmän vääriä positiivisia osumia huolimatta siitä, että tämä aiheuttaa manuaalista työtä. Hyvä esimerkki tästä on esimerkiksi rajavalvonta.

Toisaalta joskus halutaan nimenomaan minimoida kaikki manuaalisen työn tarve ja saavuttaa mahdollisimman pieni määrä vääriä positiivisia osumia. Tämä päätös voidaan tehdä arvioimalla riskiä siitä, kuinka suuri todennäköisyys on väärälle negatiiviselle osumalle ylipäätänsä ja mitkä ovat siitä aiheutuvat vahingot.

Koska nimien vertailun tavoitteet vaihtelevat eri tapauksissa, tarvitaan ohjelmistoja jotka osaavat kertoa myös muuta kuin "kyllä tai ei". Pisteytys antaa objektiivisen ja johdonmukaisen tavan mitata todennäköisyyttä tai luottamusta siihen, että vertailtavat nimet vastaavat toisiaan. Pisteytys auttaa myös automatisoimaan päätöksentekoprosessin.

Otetaan esimerkiksi pakotelistalla oleva nimi **Vladimir Vladimirovich PUTIN:**

Pisteet	**Nimi**	**Tapaus**
100	Vladimir Vladimirovich PUTIN	Täydellinen osuma
97	Vladimir V. Putin	Toinen nimi kirjoitettu lyhenteellä
93	Vladimir Vladimirovich PUTIM	Kirjoitusvirhe sukunimessä
90	Putin, Vladimir Vladimirovich	Nimet ovat eri järjestyksessä
85	Vladimir Putin	Toinen nimi puuttuu kokonaan
68	Fladim MUTIN	Kokonaan eri nimi

Päätöksentekoon voidaan asettaa haluttu kynnys, esim. 85%, jonka ylittävät tulokset läpikäydään manuaalisesti. Kynnys voidaan asettaa halutessa myös ylärajaan.

Mitä menetelmiä pisteiden laskemiseen voidaan käyttää?

Ohjelmistoihin on kehitetty useita eri tapoja nimien vertailua varten. Kaikissa tapauksissa tavoitteena on laskea eroavaisuus ja esittää se pistelukuna.

Muokkaus-etäisyys

Kuinka monta merkkiä sinun pitäisi vaihtaa, jotta nimet täsmäävät? Periaatteessa pienempien muutosten pitäisi osoittaa suuremman osuvuuden todennäköisyyttä.

Levenshtein-etäisyys on algoritmi, joka perustuu yhden merkin muutosten määrään, joka tarvitaan, jotta merkkijono on identtinen toisen kanssa. Menetelmä on helppo toteuttaa ja nopea lyhyillä merkkijonoilla.

Muokkausetäisyys sopii kuitenkin paremmin tekstijonojen kuin nimien vertailuun. Jos esimerkiksi vertaat nimeä ”Jack” nimiin ”Jac” ja ”Mack”, on molempien Levenshtein-etäisyys 1. Ihmisen loogisesti tekemä tarkastelu tietysti päättelee helposti, että jopa väärin kirjoitettuna ”Jac” on todennäköisempi osuma kuin ”Mack”,

Luettelointi

Kuinka monta eri tapaa on kirjoittaa ja translitteroida jokainen nimen komponentti? Tämä menetelmä vaatii luettelon tekemisen kaikista muunnelmista ja sen jälkeen kyselyn kaikista mahdollisista komponenttien yhdistelmistä.

Kattavan luettelon luominen ja ylläpitäminen vaatii kuitenkin suuria resursseja ja laskentatehoa. Lisäksi menetelmä ei käsittele sellaisia nimiä jotka eivät ole luettelossa ja huomioon on myös otettava kielten ja kirjoitusten väliset erot.

Samankaltaisuus kuultuna

Kuinka samanlaisilta nimet kuulostavat? Yleiset tähän käytetyt algoritmit, kuten Soundex, muuntavat nimet avaimeksi. Ajatuksena on vertailla näitä avaimia. Menetelmä riippuu kuitenkin siitä, miltä nimi kuulostaa viitekielellä, joka on yleensä englanti. Monet muiden kielten äänet on ensin muunnettava tälle viitekielelle joka vaikeuttaa vertailutyötä.

Tilastollinen samankaltaisuus

Kuinka samanlaisia nimet ovat tilastollisesti? Tilastollinen menetelmä käyttää satojen tai tuhansien vastaavien nimiparien luetteloa tekoälyn mallin opettamiseksi tunnistamaan nimien samankaltaisuus. Tuloksena oleva malli laskee todennäköisyyden sillä että vertailtavat nimet vastaavat toisiaan ja laskee sille pisteet. Mitä suurempi määrä nimipareja on tekoälyn koulutukseen käytetyssä tietojoukossa, sitä suurempi on tuloksena olevan mallin tilastollinen tarkkuus. Menetelmä kuitenkin vaatii huomattavaa manuaalista työtä, ennen kuin ohjelmisto voi tehdä työnsä. Lisäksi järjestelmä, joka käyttää pelkän tekoälymallia etsimiseen suuresta luettelosta osumia, voi olla käytännössä liian hidas.

Mikään yksittäinen menetelmä ei riitä erittäin tarkkaan ja nopeaan nimien täsmäytykseen. Onnistunut ohjelmisto lähestymistapa yhdistää useita menetelmiä.

DOKS Sanctions API

DOKS Sanctions APIn sydämenä on Rosette Name Indexer. Se on huipputeknologiaa, joka yhdistää kaikki yleiset vertailumenetelmät käyttäen kustakin sen parhaat puolet. Rosetten on kehittänyt BasisTech ja heidän vastaava teknologia on käytössä mm. USA:n ja Iso-Britannian viranomaisilla. Suomen Tunnistetieto Oy on ainoa toimija pohjoismaissa, joka tarjoaa tämän teknologian asiakkailleen rajapintapalveluna tai graafisena käyttöliittymänä.

Phonetic similarity	Jesus ↔ Heyzeus ↔ Haezoos
Transliteration spelling differences	Abdul Rasheed ↔ Abd al-Rashid
Nicknames	William ↔ Will ↔ Bill ↔ Billy
Missing spaces or hyphens	MaryEllen ↔ Mary Ellen ↔ Mary-Ellen
Titles and honorifics	Dr. ↔ Mr. ↔ Ph.D.
Gender	Jon Smith ↔ John Smith (but not Joan Smith)
Truncated name components	McDonalds ↔ McDonald ↔ McD
Missing name components	Phillip Charles Carr ↔ Phillip Carr
Out-of-order name components	Diaz, Carlos Alfonzo ↔ Carlos Alfonzo Diaz
Initials	J. E. Smith ↔ James Earl Smith
Names split inconsistently across database fields	Dick Van Dyke ↔ Dick Van . Dyke
Same name in multiple languages	Mao Zedong ↔ Mao Цзэдун ↔ 毛泽东 ↔ 毛澤東
Semantically similar names	Eagle Pharmaceuticals, Inc. ↔ Eagle Drugs, Co.
Semantically similar names across languages	Nippon Telegraph and Telephone Corporation ↔ 日本電信電話株式会社
Organizational Aliases	IBM ↔ Big Blue