

Processing and linking address data

Madis Nõmme
Karl Kilgi

Task description

Goal:

- build an address matching engine
- take addresses in textual form
- match them against structured DB
- compare results with official search

Solution available from

github.com/madis/ses-2013

Data used

Real data from Maanteeamet (Estonian Road Administration) was used

2 datasets in CSV format:

1. Tallinn - 395,638 rows
2. Whole Estonia - 2,611,625 rows

37 fields for every row

	A	B	C	D	E	F	G	H	I	J	K
1	ADOB_ID	ADS_OID	ADOB_LIIK	ORIG_TUNNUS	ETAK_ID	ADS_KEHTIV	OLEK	ADR_ID	KOODAADRESS	TAISAADDRESS	LAHIAADDRESS
2	118006	LP00012285	LP	2119541		19.02.2004 06:40:40	K	108838	377840524000004DS0000000000000000	Harju maakond, Tallinna linn, NVümme linnaosa, Kalevala tn	Kalevala tn
3	121233	LP00016245	LP	2614698		28.05.2004 09:55:18	K	110991	377840596000004JP0000000000000000	Harju maakond, Tallinna linn, Pirita linnaosa, Sarapiku tee	Sarapiku tee
4	121263	LP00016282	LP	2706927		28.05.2004 09:55:04	K	117726	377840176000004H3000000000000000	Harju maakond, Tallinna linn, Haabersti linnaosa, Kudu tn	Kudu tn
5	121411	LP00016463	LP	2104757		19.02.2004 01:25:50	K	113229	377840387000004BU0000000000000000	Harju maakond, Tallinna linn, Lasnamõõse linnaosa, Tuulemvõõse tn	Tuulemvõõse tn
6	121465	LP00016529	LP	2120362		19.02.2004 06:58:51	K	113777	377840524000003RC0000000000000000	Harju maakond, Tallinna linn, NVümme linnaosa, Uku tn	Uku tn
7	120020	LP00014757	LP	2105767		19.02.2004 01:52:14	K	115634	377840387000004BG0000000000000000	Harju maakond, Tallinna linn, Lasnamõõse linnaosa, VVõõse-Paala tn	VVõõse-Paala tn
8	120241	LP00015028	LP	2114760		19.02.2004 04:42:09	K	108500	377840298000004AH0000000000000000	Harju maakond, Tallinna linn, Kesklinna linnaosa, Rahukohtu tn	Rahukohtu tn
9	120247	LP00015035	LP	2120075		19.02.2004 06:52:54	K	116726	377840524000004E2000000000000000	Harju maakond, Tallinna linn, NVümme linnaosa, Kasteheina tn	Kasteheina tn
10	120270	LP00015063	LP	2119750		19.02.2004 06:46:20	K	108877	377840524000003Z9000000000000000	Harju maakond, Tallinna linn, NVümme linnaosa, Puraviku tn	Puraviku tn
11	118504	LP00012896	LP	2612376		28.05.2004 09:50:50	K	115414	377840298000005ZG0000000000000000	Harju maakond, Tallinna linn, Kesklinna linnaosa, Loite tn	Loite tn
12	120568	LP00015428	LP	2114000		19.02.2004 04:22:56	K	112801	377840298000003KK0000000000000000	Harju maakond, Tallinna linn, Kesklinna linnaosa, Vana-Veerenni tn	Vana-Veerenni tn
13	118732	LP00013177	LP	2107313		19.02.2004 02:21:08	K	108324	377840596000003S6000000000000000	Harju maakond, Tallinna linn, Pirita linnaosa, Kedriku tn	Kedriku tn

Importing & structuring data

With optimizations ~2000 rows/s

Wrote a migration script

- took ~5 min to import ~400k rows
- data in relational DB

Searching approaches

Simple query

- SQL WHERE x LIKE '%...%'

Full text search -

- Sphinx
- Elasticsearch

Validating results

Validated against

<http://xgis.maaamet.ee/adsavalik/ads>

Example queries

Maakond	Harju maakond
Omavalitsus	Tallinna linn
Asustusüksus/Linnaosa	Kesklinna linnaosa
Aadressi normaliseerimata osa tekstilisel kujul	Müürivahe 14 *
Objektiliik	<div>?</div>
Täisaadress	Harju maakond, Tallinna linn, Kesklinna linnaosa, Müürivahe tn 14
Optimeeritud aadress	Harju maakond, Tallinna linn, Kesklinna linnaosa, Müürivahe tn 14
Objekti normaliseeritud lähiaadress stringina	Müürivahe tn 14

Validating results

Results

Jkn	ADR-ID	Objekti versiooni id	Liik	Tase 1	Tase 2	Tase 3	Tase 4	Tase 5	Tase 6	Tase 7	Tase 8	Koodaadress	Aadressi normaliseerimata osa tekstilisel kujul
1	390906	3551857	mitteelukondlik hoone	Kood: 37 Nimi: Harju maakond Nimi liigisõnaga: Harju maakond	Kood: 784 Nimi: Tallinna linn Nimi liigisõnaga: Tallinna linn	Kood: 0298 Nimi: Kesklinna linnaosa Nimi liigisõnaga: Kesklinna linnaosa		Kood: 03QJ Nimi: Müürivahe tn Nimi liigisõnaga: Müürivahe tänav		Kood: OPEN Nimi: 14 Nimi liigisõnaga: 14		377840298000003QJ00000PEN00000000	
2	390906	3061437	katastriüksus	Kood: 37 Nimi: Harju maakond Nimi liigisõnaga: Harju maakond	Kood: 784 Nimi: Tallinna linn Nimi liigisõnaga: Tallinna linn	Kood: 0298 Nimi: Kesklinna linnaosa Nimi liigisõnaga: Kesklinna linnaosa		Kood: 03QJ Nimi: Müürivahe tn Nimi liigisõnaga: Müürivahe tänav		Kood: OPEN Nimi: 14 Nimi liigisõnaga: 14		377840298000003QJ00000PEN00000000	
3	390906	3853189	elukondlik hoone	Kood: 37 Nimi: Harju maakond Nimi liigisõnaga: Harju maakond	Kood: 784 Nimi: Tallinna linn Nimi liigisõnaga: Tallinna linn	Kood: 0298 Nimi: Kesklinna linnaosa Nimi liigisõnaga: Kesklinna linnaosa		Kood: 03QJ Nimi: Müürivahe tn Nimi liigisõnaga: Müürivahe tänav		Kood: OPEN Nimi: 14 Nimi liigisõnaga: 14		377840298000003QJ00000PEN00000000	
4	390906	4808103	elukondlik hoone	Kood: 37 Nimi: Harju maakond Nimi liigisõnaga: Harju maakond	Kood: 784 Nimi: Tallinna linn Nimi liigisõnaga: Tallinna linn	Kood: 0298 Nimi: Kesklinna linnaosa Nimi liigisõnaga: Kesklinna linnaosa		Kood: 03QJ Nimi: Müürivahe tn Nimi liigisõnaga: Müürivahe tänav		Kood: OPEN Nimi: 14 Nimi liigisõnaga: 14		377840298000003QJ00000PEN00000000	

Our solution

Demo

Conclusion

Results vary with different search methods

Primitive SQL search

- leaves out results
- includes too many results
- it's much slower (if not optimized)

Full text search

- reasonably accurate results
- faster

We are unsure of why Maanteamet's query leaves out some results