Project2023

Meeskond:
*) Ellar Seidelberg

Projekti link:
(https://github.com/EllarSeidelberg/Project2023)

1. Sissejuhatus

Eks alati võib tekkida küsimus, kui oleks vaja uut tänavat nimetada, et milline üks õige ja ilus tänava nimi peaks olema. Oleme me ju kõik kuulnud ütlust, et ühes õiges poisslapse nimes peab olema “R” täht. Siit ka koorus välja mõte teha uuring tähtede kasutusest tänavanimedes.

Positiivne tulem oleks see, kui suudame mõne uue tänavanime puhul hinnata, kas seal on niinimetatud vajalikud tähed olemas.

Negatiivne tulem võiks olla see, kui just enim esinevaid tähti tulevikus tänavanimedes vältida (kuna neid on näiteks seni liiga palju juba kasutatud).

Õnnestumise tõenäosus oleks see, kui umbes 20-30% tähestiku tähtedest tõuseb esile.

Ebaõnn oleks see, kui kõiki tähti oleks kasutatud sama sagedalt.

2. Andmete allikad ja kirjeldus

Kasutame https://avaandmed.eesti.ee/ leheküljelt saadud Tallinna tänavate andmebaasi.
Täpsemalt lingilt:
https://avaandmed.eesti.ee/datasets/tallinna-aadresside-tabelid-asumite-kaupa

Ja veelgi täpsemalt valime sealt "Tallinna aadresside tabelid asumite kaupa 29.11.2015 10 MB (XLSX)" ehk "aadressid20151129.xlsx"

Kuna kättesaadav oli Tallinna tänavate nimede andmebaas, siis kitsendame oma uuringut Tallinna tänava nimedele.

Andmete andja on kirjeldanud, et “Metaandmete tõlkimiseks on kasutatud masintõlget ning nende kvaliteet võib seega olla kohati ebakorrektne”. Seda on tõesti näha ja tunda. Andmete puhastamiseks ja korduste väljavõtmiseks läheb palju aega ja vaeva.

Eemaldame kõikidelt inimeste auks nimetatud tänavatelt nimedest eesnimed või siis eesnime tähed, et jääksid ainult perekonnanimed. Näiteks “Jaan Koorti” tänav esines nii “J.Koorti”, “J. Koorti” kui ka “Jaan Koorti” nimekujul.

Eemaldame ka kõik tüüni liitmed nagu näiteks tänav, manatee. Seejuures esialgu eemaldame “tn” ning “mnt” ja seejärel vaatame käsitsi järgijäänud erandid üle.

Ja lõpuks saab eraldatatavatest liidetest selline kooslus:

aas	mnt	saar
aed	ots	tänav
allee	õu	tee
haljak	park	tiik
jalg	plats	tn
järv	põik	turg
kael	pst	umbtänav
käik	puiestee	vaateplats
liivakarjäär	puiestik	väljak
maantee	raudteejaam	

Lisaks on algandmetes just kui kirjeldatud ühe reana tänav ja siis kõik majanumbrid sellel tänaval eraldi lisareane. Need majanumbritega read viskame välja. Õigupoolest teeme nii, et kõigepealt viskame nende ridade lõpust välja numbri osa, nimetuse jätame alles. Seejärel jätame sarnastest nimetustest ainult ühe alles.

Kui meil algselt on andmebaasis 272'075 kirjet, siis peale puhastamist jääb alles ainult 1470 kirjet. Seda on küllaltki vähe, vahepaeal on isegi mõte, et peaks teemat vahetame või miskit, aga kuna ühest küljest paremat teemat ka mõttesse ei tule ning eks elus võib teinegi kord juhtuda, et andmeid on raske saada või neid on vähe, siis jätkame siiski sama teemaga ja samade andmetega.

3. Meetodid ja tulemused

Reaalselt alustame tööd andmebaasiga sellest, et loeme sisse ".xlsx" faili ja salvestame selle ".csv" formaati. Seejärel tühjendame mälu ja loeme sisse ".csv". Mõte ja problem mida lahendame on see, et ".xlsx" faili lugemine teeb arvuti aeglaseks. Aga ".csv" failiga probleeme ei ole ja kiirust jätkub. Ning selle salvestamise ja uuesti lugemisega saavutame ka salvestuspunkti, et ei pea Jupyter'i Notebook'i algusest peal iga kord uuesti läbi jooksutama. Algus oli meil see aeglane ".xlsx" failist lugemise osa.

Seejärel puhastame, mitu mitu korda ja järsjest täiuslikumalt. Sellest juba rääkisime eelmises punktis. Ja see ilmselt oli ka kõige ajamahukam osa.

Seejärel on küsimus, kas uurida sõnu tervikuna või jagada need eraldi tähdeteks ja siis uurida. Läheme seda teed, et jätame andmekogusse ainult tänavanimede tulba (lisaks olid linnaosad jne) ning sisestame lisaks eesti tähestiku tähtetega tulbad lisaks ja seejärel loeme igas reas esimese tulba "Lühiaadress" ehk tänavanimi üle ja paneme igasse tähestiku tähe tulpa antud real numbrilise tähtede arvu, mitu korda antud tähte antud rea tänavanimes esines.

Ja lisasime ka "Lühiaadress" ehk tänavanime ja tähestikutähtede vahele tulba "Pikkus" kuhu sisestame sõna pikkuse. Sest töö käigus tuli mõte, et lisaks võiks ka teha statistikat, kui pikk üks tänavanimi olema peaks. Ja leidsime et parim selleks on lähtuda olemasolevate nimede pikkusest.

Seejärel liitsime iga tähe tulba ülalt-alla kokku ja saimegi mis tähte on kõige rohkem kasutatud ja millist kõige vähem. Selle liitmise käigus tekkis küsimus, kui ühes sõnas on 3x "A" täht, kas siis lugeda kogumisse "3" või ainult "1" ehk "esineb" või nii öelda "tõene" või "väär" loogika alusel.

Kuna ei suutnud lõplikult otsustada ja arvutit kokkulugemine väga ei koormanud, siis tegime mõlemad. Ja huvitavam saigi, sest tekkis juba väike võrdlus, kas ühe või mitmekordne arvestamine muudab tulemust. Tulemus on nähtav "Tabelis-1".

Tabel-1(osa1):

	A	B	C	D	E	F	G	H	I
Koos kordustega	1396	96	0	183	914	8	127	190	1087
Kordused eemaldatud	907	93	0	181	670	8	126	186	811

Tabel-1(osa2):

	J	K	L	M	N	O	P	Q	R
Koos kordustega	146	620	653	381	404	308	258	0	607
Kordused eemaldatud	146	540	546	346	331	223	244	0	564

Tabel-1 (osa3):

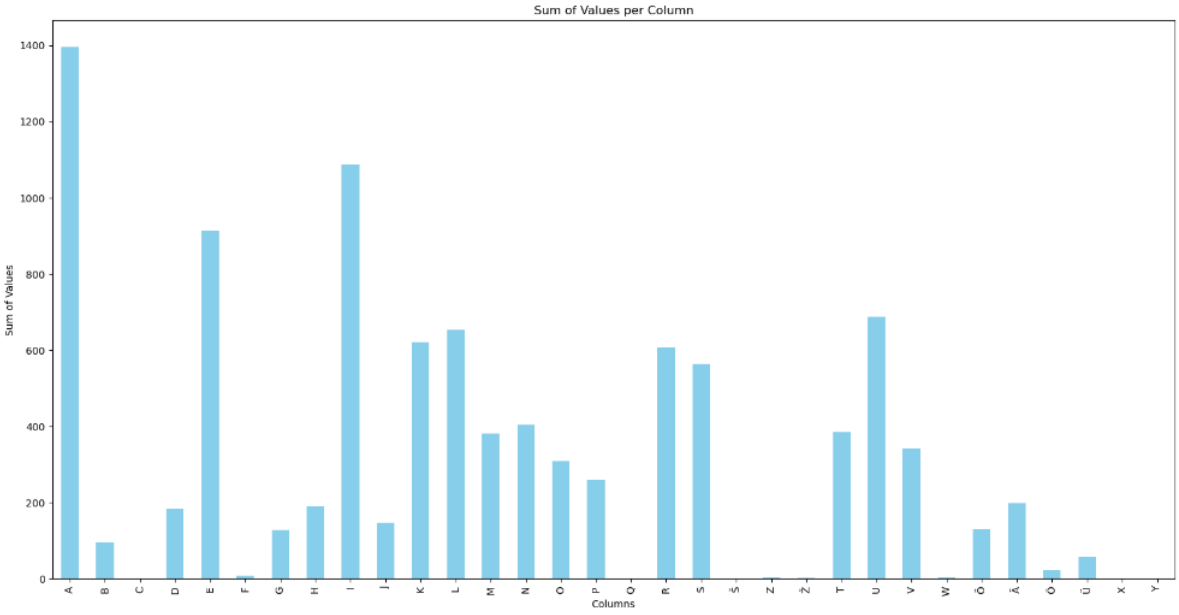
	S	Š	Z	Ž	T	U	V	W	Õ
Koos kordustega	564	0	4	2	384	688	341	4	130
Kordused eemaldatud	493	0	4	2	357	549	319	4	124

Tabel-1 (osa4):

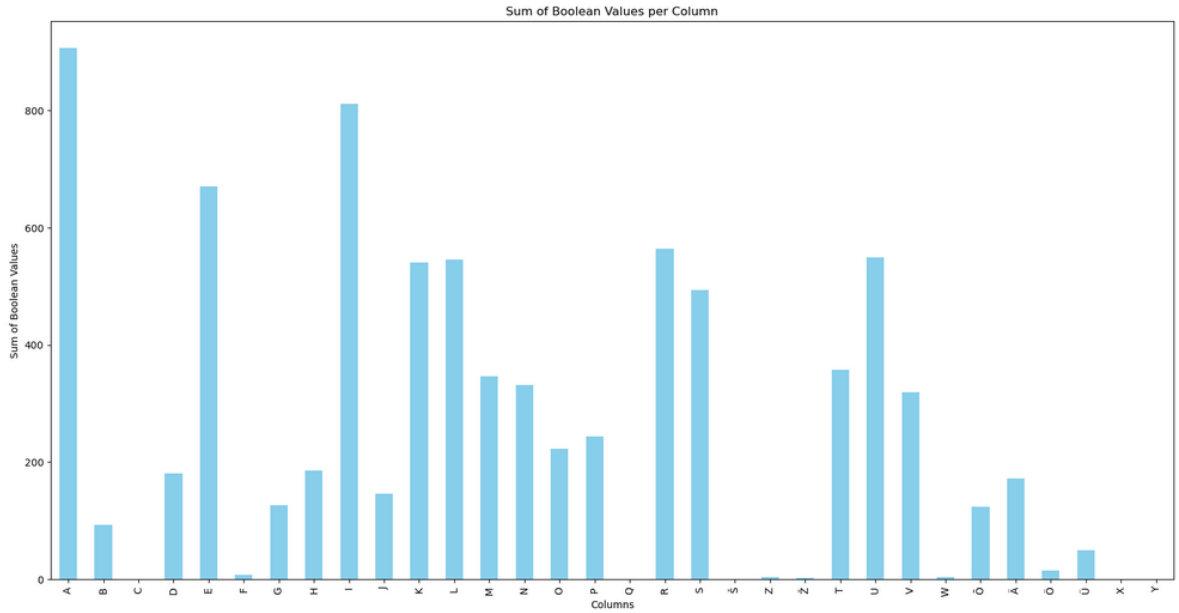
	Ä	Ö	Ü	X	Y
Koos kordustega	197	23	58	0	0
Kordused eemaldatud	172	15	50	0	0

Ja parema ülevaate huvides kujutasime tulemust ka graafiliselt:

Kõigepealt siis arvestades igat korduvat tähte:



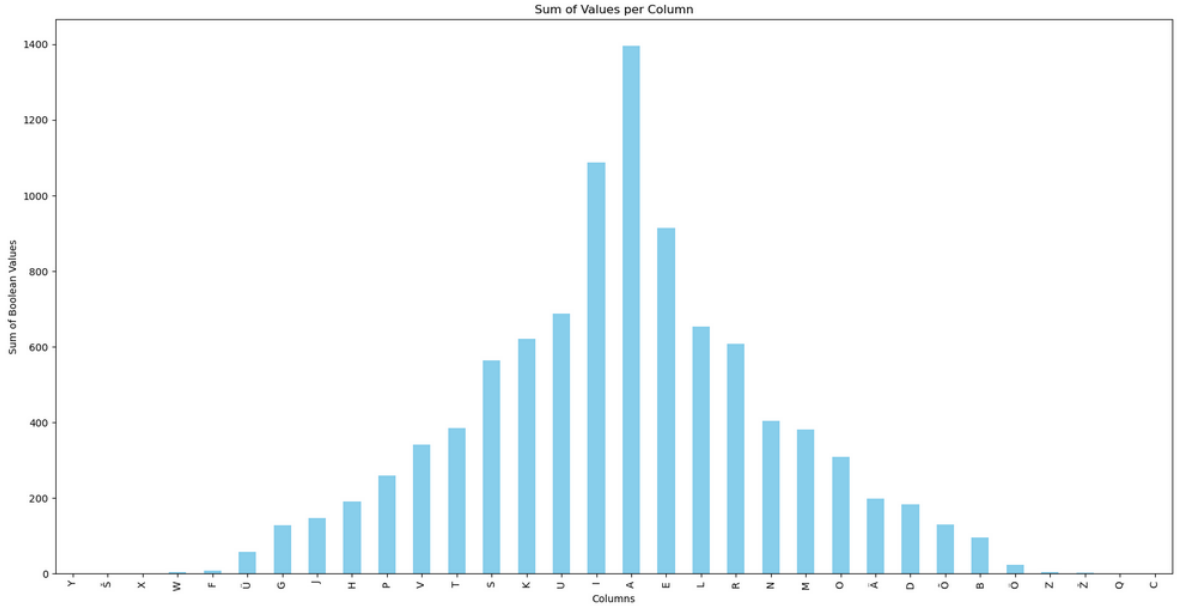
Ning seejärel ka ilma korduvaid tähti arvestamata:



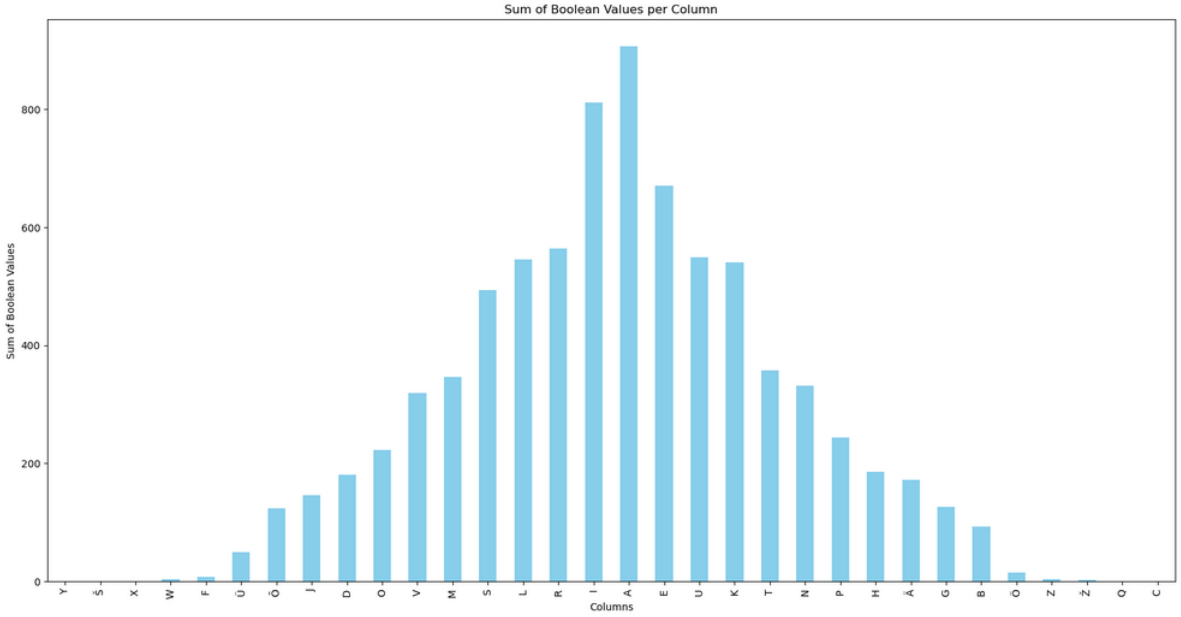
Ja näeme, et jaotus on suhteliselt sama. (Huvitav oleks tegelikult võrdlus eesti keelega üldiselt või siis eesti keele nimisõnadega, et kas oleme saanud lihtsalt näiteks eesti keele nimisõnadele omase tähtede esinemissageduse ja sellega kinnitanud lihtsalt reeglit või on tänavanimedes siiski teine tähtede esinemissagedus, kui eesti keele nimisõnades keskmiselt)

Aga et oleks parem ja huvitavam ülevaade, proovisime anda graafikutele normaaljaotuse kuju, ehk kõige enim esinenud väärtused keskele ja harvem esinenud äärtesse. Ja saime järgnevad graafikud:

Normaaljaotus kuju (arvestades igat korduvat tähte):



Normaaljaotus kuju (ilma korduvaid tähti arvestamata):



Ja näeme, et jaotus vastab suhteliselt hästi normaaljaotuse üldkujule.

Enim esinenud tähed Tallinna tänavanimedes on (kordustega/ilma kordusteta):
(meeldetuletuseks, et valim peale puhastamist oli 1470 kirjet)
A = 1396/907 ehk (95% / 62%)
I = 1087/811 ehk (74% / 55%)
E = 914/670 ehk (62% / 45%)

Ja tänavanime pikkus on keskmisel 7, minimaalselt 2 ja maksimaalselt 16 tähemärki.

4. Järeldused ja arutelu

Ühes ilusas Tallinna tänavanimes on kindlasti olemas “A” täht. Veel parem kui esineks ka “I” ja “E”. Ühtlasi näeme, et meie uuring on õnnestunud, meil kerkisid esile selged liidrid.

Proovisime otsida ka eesti nimisõnade andmebaasi, aga ei leidnud. Meid oleks huvitanud näiteks kõik nimisõnad omastavas käändes. Siis oleks saanud teha statistikat, kui palju protsentuaalselt on juba kasutusel tänavanimedena ja kui palju veel nii öelda varuks.

Edaspidi võiks otsida, kas keegi on teinud eesti nimisõnade omastava käände tähekasutuse statistikat ja kui on siis kuidas seal tähtede jaotus on. Ning kui ei ole, siis saab seda teha. Ning kui tõsti pole veel olemas eesti keele nimisõnade andmebaasi, kus nad esineksid juba valmiskujul kõigis 14 käändes, siis võiks selle luua. (see aitaks kindlasti kaasa ka eesti keelsele masinõppele ehk masinal eesti keelt õppida) Me otsisime, ei leidnud, ainus küsimus on, kas otsisime piisavalt põhjalikult.

Kui oleks eesti nimisõnade andmebaas, siis saaks võtta välja nimekirja neid sõnu, mida täna veel pole tänavanimedena kasutatud. Seejärel need siiski inimfaktoriga üle vaadata (kas on sobilikud tänavanimeks) ja kui on sobilikud, siis panna varusse uute tänavate nimetamiseks.

Võiks lasta masinõppel antud tänavanimede andmebaasi pealt mudel koostada ja siis lasta masinal genereerida mudeli baasil 25/50/100 sõna pikkusega 7+/-2 tähemärki nii öelda uuteks potensiaalseteks tänavanimedeks. Kindlasti peaks tulemi üle kontrollima esiteks meie tänavanimede andmebaasist, et sellist nime juba ei esine ning teiseks ka inimlikult, et selline genereeritud sõna oleks sobilik, nii keeleliselt kui ka mõtteliselt. Eks ta ikkagi peaks olema mõni olemasolev sõna, mida veel pole tänavanimene kasutatud.

Ehk kokkuvõtvalt mõned mõtted edasiarendusteks:

*) Kui leiaks eesti keele nimisõnade andmebaasi, kus oleksid kõik käänded (meil oleks omastavat vaja) olemas, siis saaksime võrrelda, kui palju on juba kasutusel ja kui palju veel vabu ja kasutamata.

*) Kui sellist andmebaasi ei ole, siis võiks selle luua. Et teha eesti keele nimisõnade andmebaas, kus sõnad on juba valmis 14 käändes käänatud. Kindlasti aitab see eesti keele masinõppele ka kaasa.

*) Võib proovida masinõpet ja siis lasta genereerida teatud arvu (25/50/100) sõnu ning kontrollida, kas need on juba olemas/kasutusel ja kui ei ole, siis inimfaktoriga need läbi töötada ning mõelda, kas need sobiksid uuteks tänava nimedeks.

*) Võib antud uuringut teostada mõnes muus omavalitsuses.

*) Võib antud uuringut laiendada tervele Eestile, sellisel juhul ei tohiks ilmselt valimist välja arvata korduvat sõna, kui kordused esinevad eri omavalitsustes.

*) Miks mitte kontrollida ka ütlust, et ühes õiges poisslapse nimes (eestis) on alati „R“ täht.