Монгол хэлний нийлмэл хэлцийн таниур

*Монгол хэлний нийлмэл хэлцийн таниурын хэрэгжүүлэлт

Э.Багабанди (19B1NUM0700)
ХШУИС - МКУТ,
МУИС
Монгол улс, Улаабаатар хот
19B1NUM0700@stud.num.edu.mn

I. Удиртгал

Дэлхийн өндөр хөгжилтэй улс орны хиймэл оюун болон машин сургалт ашиглан хэл боловсруулах технологи нь манай улсаас нилээн хэдэн жилээр түрүүлж явж байгаа бөгөөд тухайн өндөр хөгжилтай улс орнуудын хэл дээр ниймэл хэлц (Multi-word expression) – ийн таниурыг хэрэгжүүлсэн байдаг. Харин Монгол хэлэн дээр ниймэл хэлцийн таниур хараахан хэрэгжээгүй байна. Өөр Монгол хэлтэй төстэй залгамал бүтэцтэй хэлэн дээр хэрэгжүүлэгдсэн нийлмэл хэлцийн таниурыг Монгол хэлэн дээр хэрэгжүүлэхэд хугацаа бага зарцуулах хэдий ч хэлний онцлог хэв шинжээр хамаарч нэмэлт хөгжүүлэлт орох, хувирсан нийлмэл хэлцийг олохгүй байх гэх зэрэг дутагдалтай. Энэхүү ажлаар тухайн нийлмэл хэлц нь ямар нэг нөхцөл, дагавраар хувирсан эсэхээс үл хамаарч олдог Монгол хэлний нийлмэл хэлцийн таниурын хэрэгжүүлэлт болон түүний хэрэглээг хуулийн баримт бичгүүд дэх нэр томьёоны жишээн дээрээр хийгдсэн үн дүнг танилцуулна.

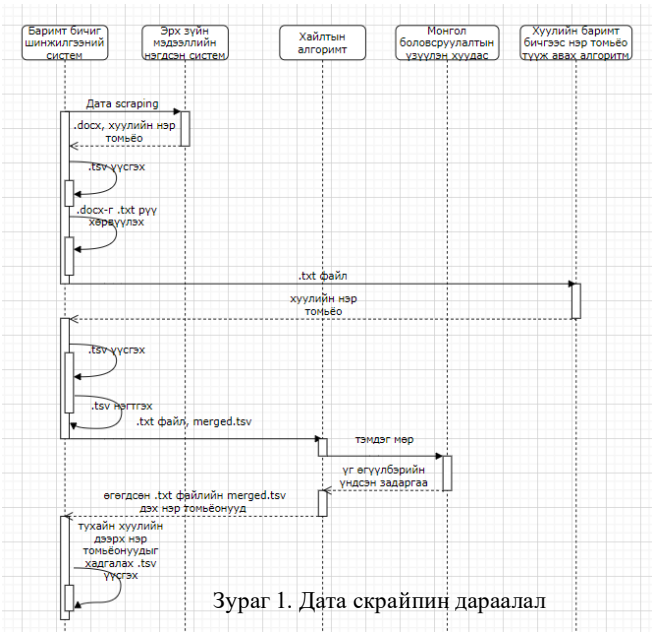
Монгол хэлтэй ижил залгамал бүтэцтэй хэлэн дээр хэрэгжүүлсэн нийлмэл хэлцийн таниур нь нийлмэл хэлцийг язгуур болон үгийн бүтцээр нь задалж таньдаггүй учраас Монгол хэлэн дээр нэвтрүүлэхэд хүндрэлтэй байна.

II. Араг зүй

Монгол хэлний нийлмэл хэлцийн таниурыг хуулийн баримт бичиг дээрх томьёон хэрэгжүүлэх үйл явц:

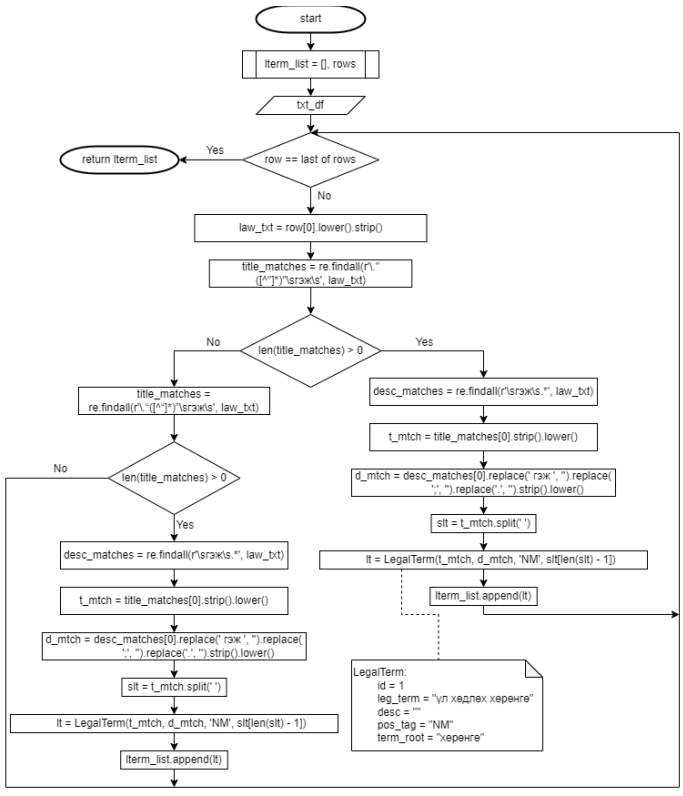
A. Эрх зүйн мэдээллийн нэгдсэн системээс бүх хуулийн нэр томьёог дата скрайпин(scraping) хийж авчрих

Эрх зүйн мэдээллийн нэгдсэн систем(legalinfo.mn) – ийн “Хуулийн нэр томьёо” цэсэнд байрлах бүх хуулийн нэр томьёог татаж аван “Хуулийн нэр томьёо, нэр томьёоны тодорхойлолт, үгийн аймгийн тэмдэглэгээ(ниймэл хэлц гэх таних тэмдэглэгээ), үндсэн үг(нийлмэл хэлцийн хамгийн сүүлийн үг)” гэх загвараар хадгалж TSV(Tab-separated values) файл болгон хадгална. Тухайн цэснээс хуулийн нэр томьёог татаж авчрихтай зэрэгцэн хуулийн бүх docx өргөтгөлтэй файлуудыг татаж авчрих бөгөөд мөн түүнийг текст файл руу хөрвүүлнэ.Үүний дараа “Хуулийн баримт бичгийн хуулийн нэр томьёоны хэсгээс нэр томьёог ялгаж авах” алгоритм ашиглан тухайн хуулийн баримт бичиг дээрээс бүх хуулийн нэр түүж аван дээр дурдсан загвараар TSV файл үүсгэж хадгална. Үүсгэсэн хоёр TSV файлыг нэгтгэж нэгдсэн нэг TSV файл үүсгэнэ. Эцэст нь хадгалж авсан хуулийн нэр томьёо, үгийг үндсээр задлаж өгөх API болон баримт бичгээс нийлмэл хэлц хайх алгоритмыг ашиглан хуулиуд дээр дурдагдсан нэр томьёонуудыг хайж олон хадгална.



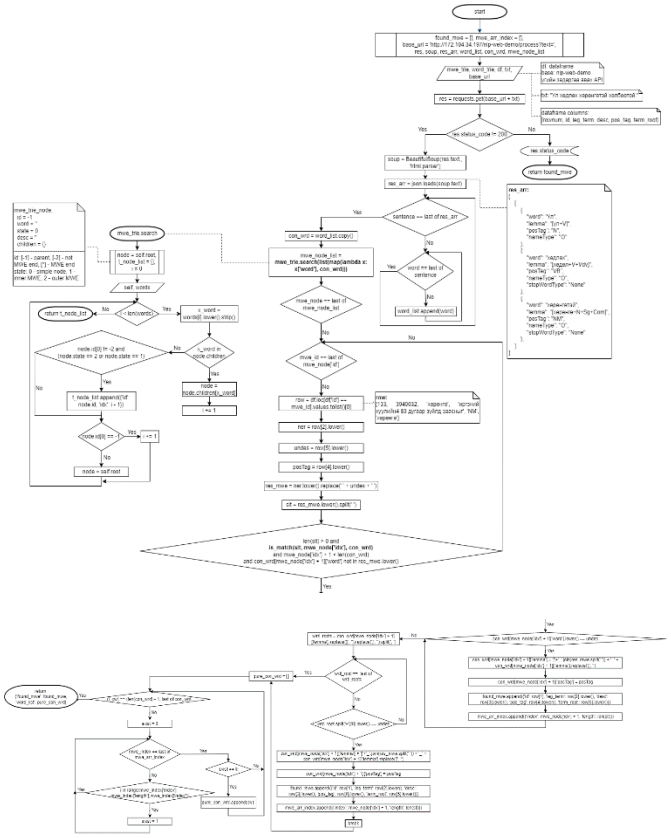
Зураг 1. Дата скрайпин дараалал

B. Хуулийн баримт бичгийн хуулийн нэр томьёо хэсгээс нэр томьёо ялгаж авах алгоритм хэрэгжүүлэх



Зураг 2. Нэр томьёо ялгаж авах алгоритм

C. Хадгалсан хуулийн нэр томьёог ашиглан текст файлаас хайлт хийх зохистой алгоритм хэрэгжүүлэх



Зураг 3. Хуулийн нэр томьёог хайх алгоритм

III. УР ДҮН БА ХЭЛЭЛЦҮҮЛЭГ

Хэд хэдэн үгээс тогтох хэлц буюу бүлэг үгсийг нийлмэл хэлц буюу multiword expression гэнэ. Энэхүү өгүүллээр Монгол хэлний нийлмэл хэлцийн таниурын хэрэгжүүлэлт болон түүний хэрэглээг хуулийн баримт бичгүүд дэх нэр томьёоны жишээн дээр үзүүллээ. Нийт 844 хуулийн баримт бичгээс хууль дотор тодорхойлсон 3272 ялгаатай нэр томьёог түүвэрлэн шинжиллээ. Эдгээрээс 10 ба түүнээс дээш урттай 15, хамгийн урт нь 15 үгтэй байна. Мөн нэг үгтэй 329 нэр томьёо байх бөгөөд дунджаар нэг нэр томьёо 3 үгийн урттай байна.

id	leg_term	desc
8260059	яллах тал	прокурорыг
3180025	ял эдэлж байгаа	эрүүгийн хуулийн 52 дугаар зүйлд заасан үнд
5530026	явуулын багшийн сургалт	боловсролын стандартыг баримтлан гэр бүл,
1980023	явган зорчигч	замаар явган явааа /зам дээр ажил үүрэг гүйц
3180029	ээлжит сонгууль	монгол улсын үндсэн хуулийн гучдугаар зүйл

	pos_tag	term_root	tindoc	len	FindocT	tindoc4
	N	тал	1	2	1	1
олон нэмэгдэл ял, тэнсэх, албадлагын арга хэмж	N	байгаа	22	3	22	30
иглэж, хүүхдэд сургуулийн өмнөх боловсрол эзэ	N	сургалт	1	3	1	1
яваагаас бусад/ хүн, жагсаалаар явааа болон хөгж	N	зорчигч	10	2	10	10
дахь хэсэгт зааснаар зургаан жил тутам явагдах	N	сонгууль	11	2	11	12

Зураг 4. Хуулийн нэр томьёонууд

Эдгээр нэр томьёог сан болгон хадгалж нийлмэл хэлцийн таниур хэрэгслийг python хэл дээр хэрэгжүүлж хуулийн баримтын сан дээр туршив. Үүнээс хамгийн олон нэр томьёо агуулсан хууль 278 нэр томьёо бүхий “Зөрчлийн тухай хууль” байна. Мөн нэг хуулийн баримт дунджаар 16 нэр томьёо агуулж байна.

No	Files	Tcount		Хууль	#Нэр томьёо
1	**MNCLW00239**	551		Зөрчлийн тухай	551
2	**MNCLW00234**	189		Зөвшөөрлийн тухай	189
3	**MNCLW00807**	184		Шүүхийн шийдвэр гүйцэтгэх тухай	184
4	**MNCLW00639**	178		Улсын тэмдэгтийн хураамжийн тухай	178
5	**MNCLW00572**	174		Татварын ерөнхий хууль	174
6	**MNCLW00825**	162		Эрүүгийн хууль /2015	162
7	**MNCLW00824**	160		Эрүүгийн хууль /2002	160

Зураг 5. Хуулийн баримт бичиг дэх нэр томьёоны тоо

Нэр томьёоны давтамжаар эрэмбэлбэл хамгийн их хэрэглэгдсэн нь ‘үйл ажиллагаа’ 329 , ‘арга хэмжээ’ 308, ‘баримт бичиг’ 217 баримтад тус тус олдож байна. Нэг нэр томьёо дунджаар 6 баримтад дурдагдаж байна.

No	Term_id	Fcnt	Нэр томьёо	#Баримт
1	**6770048**	330	үйл ажиллагаа	330
2	**6140040**	307	арга хэмжээ	307
3	**2770030**	304	хуулийн этгээд	304
4	**4940013**	243	монгол улсын олон улсын гэрээ	243
5	**938**	240	төрийн ордон	240
6	**7160014**	222	аж ахуйн нэгж	222
7	**350026**	211	баримт бичиг	211

Зураг 6. Нэр томьёоны давтамжаар эрэмбэлбэл

ДҮГНЭЛТ

Эхний хувилбарыг хэрэгжүүлж бодит хуулийн сан дээр туршлаа. Сайжруулалт нь хуулийн баримтаас бичвэр салгаж авах хэсэг дээр анхаарах.

НОМ ЗҮЙ

[1] Эрх зүйн мэдээллийн нэгдсэн систем *(https://legalinfo.mn/mn)*