

# Монгол хэлний нийлмэл хэлцийн таниур

\*Монгол хэлний нийлмэл хэлцийн таниурын хэрэгжүүлэлт

Э.Багабанди  
Мэдээлэл, компьютерийн ухааны  
тэнхим  
ХШУИС, МУИС  
Улаанбаатар, Монгол  
[19B1NUM0700@num.edu.mn](mailto:19B1NUM0700@num.edu.mn)

Ч.Алтангэрэл  
Мэдээлэл, компьютерийн ухааны  
тэнхим  
ХШУИС, МУИС  
Улаанбаатар, Монгол  
[altangerel@num.edu.mn](mailto:altangerel@num.edu.mn)

**Хураангуй** – Нийлмэл хэлц гэдэг нь хэд хэдэн үгнээс тогтон нэг утга санааг илэрхийлэх үгсийг хэлэх бөгөөд хуулийн нэр томьёо нь нийлмэл хэлцийн нэг тод жишээ юм. Монгол хэлэн дээр хараахан хөгжүүлэгдээгүй байгаа нийлмэл хэлцийн таниурыг хуулийн баримт бичиг дээр хэрэгжүүлснээр хуулийн нэр томьёог дурын баримт бичгээс таньж олох, түүнтэй холбоотой хуулийн заалтуудыг түргэн хугацаанд олж мэдэх, хуулийн давхардсан утга санаа бүхий нэр томьёог хэрэглээг багасгах, буруу ашигласан нэр томьёог олж илрүүлэх гэсэн олон давуу тал болон ач холбогдолтой юм. Тиймээс ч энэхүү судалгааны ажлаар монгол хэлний нийлмэл хэлцийн таниурыг хуулийн баримт бичгүүд дээр хэрэгжүүлэлт болон түүний үр дүнг танилцуулах билээ.

**Түлхүүр** – Монгол хэлний нийлмэл хэлц, хайлтын алгоритм, хуулийн нэр томьёо, хэрэгжүүлэлт, хэл боловсруулалт

## I. УДИРТГАЛ

Дэлхийн өндөр хөгжилтэй улс орны хиймэл оюун болон машин сургалт ашиглан хэл боловсруулах технологи нь манай улсаас нилээд хэдэн жилээр түрүүлж явж байгаа бөгөөд тухайн өндөр хөгжилтэй улс орнууд өөрсдийн хэл дээр ниймэл хэлц (Multi-word expression) – ийн таниурыг хэрэгжүүлсэн байдаг. Харин Монгол хэлэн дээр ниймэл хэлцийн таниур хараахан хэрэгжээгүй байдаг. 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 судалгааны 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 дээр хийгдсэн үн дүнг танилцуулна.

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

## II. АРГА ЗҮЙ

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

A. Эрх зүйн мэдээллийн нэгдсэн системээс бүх хуулийн нэр томьёог data скрайпинг(scraping) хийж авчрах дараалал

1. Эрх зүйн мэдээллийн нэгдсэн системийн “Хуулийн нэр томьёо” цэсэнд байрлах бүх хуулийн нэр томьёог татаж авчирна.
2. “ID, Хуулийн нэр томьёо, нэр томьёоны тодорхойлолт, үгийн аймгийн тэмдэглэгээ(нийлмэл хэлц гэх таних тэмдэглэгээ), үндсэн үг(нийлмэл хэлцийн хамгийн сүүлийн үг)” гэх загвар(цаашид нэр томьёоны загвар гэх) – аар TSV (Tab-separated values) файл болгож хадгална.

	leg_term	pos_tag	term_root
0	үл хөдлөх эд хөрөнгө	N	хөрөнгө
1	биет бус хөрөнгө	N	хөрөнгө
2	аж ахуйн нэгж	N	нэгж
3	гадаад улсад олсон орлого	N	орлого

Зураг 1. Нэр томьёоны загвар

3. Тухайн цэснээс хуулийн нэр томьёо болон хуулийн бүх Word файл буюу DOC/DOCX өргөтгөлтэй файлуудыг татаж авчирна.
4. DOC/DOCX өргөтгөлтэй файлууд текст файл руу хөрвүүлнэ.
5. “Хуулийн баримт бичгийн хуулийн нэр томьёоны хэсгээс нэр томьёог ялгаж авах” алгоритм ашиглан тухайн хуулийн баримт бичиг дээрээс бүх хуулийн нэр түүж аван нэр томьёоны загвараар мөн өөр нэг TSV файл үүсгэж хадгална.
6. Үүсгэсэн хоёр TSV файлыг нэгтгэж нэгдсэн нэг TSV файл үүсгэнэ.
7. Хадгалж авсан хуулийн нэр томьёонууд, мөн үгийг үндсээр задлаж өгөх монгол хэл боловсруулалтын үзүүлэн хуудасны API болон “дурын баримт бичгээс нийлмэл хэлц буюу хуулийн нэр томьёог хайх алгоритм” – ыг ашиглан хуулиуд дээр дурдагдсан нэр томьёонуудыг олж хадгална.

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

Оролт: Текст файл

1. Текст файлаас мөр унших

- Мөрийг сүүлийн мөр эсэхийг шалгах. Сүүлийн мөр бол 6 алхам, үгүй бол 3 алхам руу шилжинэ.
- Мөрөнд [.] эхэлж, [\"sгэж\|s] төгссөн загвартай үгс байвал тэдгээрийг олно.
- Тухайн загвартай үгс агуулсан мөрний “гэж” өмнөх хэсгийг “хуулийн томьёо”, харин хойш хэсгийг “нэр томьёоны тодорхойлолт” болгон нэр томьёоны загвараар хадгална.
- 2 алхам руу шилжинэ.
- Хадгалсан нэр томьёоны жагсаалтыг буцаана.

С. Дурын баримт бичгээс нийлмэл хэлц буюу хуулийн нэр томьёог хайх алгоритм хэрэгжүүлэх

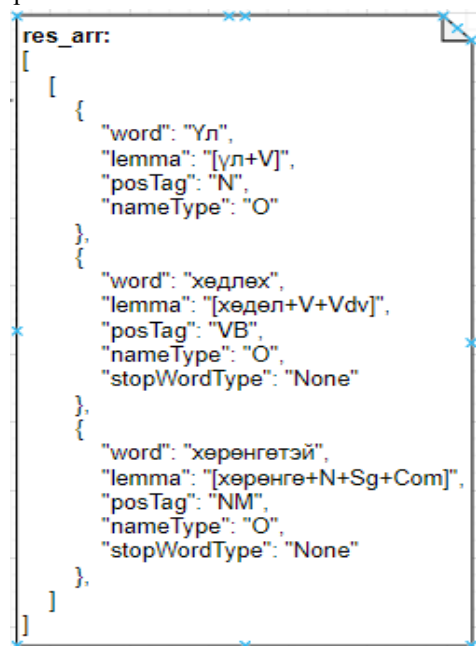
Trie (тэмдэг мөрүүдийг үр ашигтай хадгалах, сэргээн ашиглахад зориулагдсан мод/tree дээр суурилсан өгөгдлийн бүтэц) бүтэц рүү хуулийн нэр томьёог хадгалахдаа тухайн нэр томьёоны сүүлийн үгийг хасч, эхний үгсийг нэг нэг зангилаа болгон хадгална. Жишээ нь, “үл хөдлөх эд хөрөнгө” – ийг “үл хөдлөх эд” гэж хадгална.

хуулийн-[2770030, 7460013]	төслийг-[-2]	эцэслэн-[3500021]	
хуулийн-[2770030, 7460013]	этгээдийн-[7410020, 6550060, 7410016]	оршин-[-2]	байгаа-[1221]
хуулийн-[2770030, 7460013]	этгээдийн-[7410020, 6550060, 7410016]	улсын-[6350015]	
хуулийн-[2770030, 7460013]	этгээдийн-[7410020, 6550060, 7410016]	хувийн-[7410018]	
хуулийн-[2770030, 7460013]	этгээдийн-[7410020, 6550060, 7410016]	ээзмшлийн-[-2]	орон-[1223]

Зураг 2. Нэр томьёо хадгалах загвар

Оролт: 2 буюу түүнээс дээш хуулийн нэр томьёонуудыг хадгалсан Trie, 1 үгнээс бүтэх хуулийн нэр томьёонуудыг хадгалсан Trie, тэмдэг мөр буюу текст

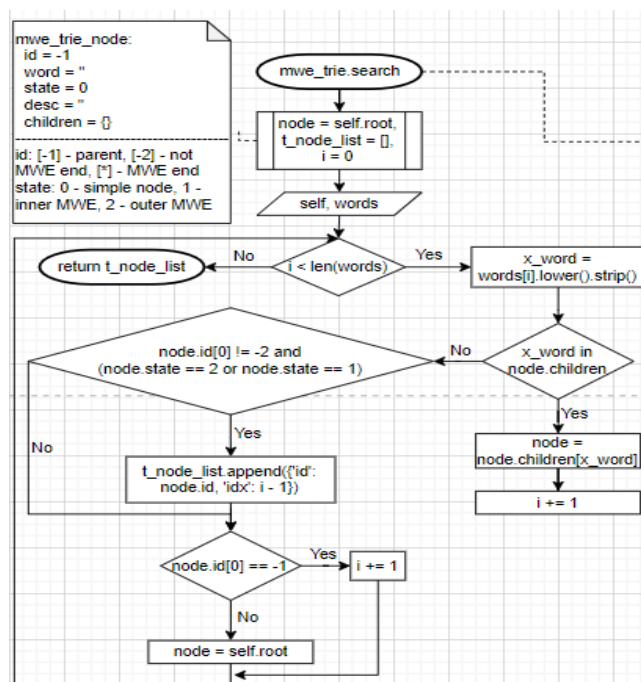
- Монгол хэл боловсруулалтын үзүүлэн хуудас ([Mongolian NLP Demo](#)) хуудас руу хүсэлт шидэж, оруулсан текст дэх үгсийг язгуураар нь задалж авчирна.



Зураг 3. Нэр томьёо хадгалах загвар

- Үгсийн задаргааны жагсаалтаас зөвхөн үгсийг ялгаж авна.

- Тухайн ялгаж авсан жагсаалтан дээр 2 буюу түүнээс урттай үгнээс бүтэх хуулийн нэр томьёог хадгалсан Trie – н хайлтыг ажиллуулна.



Зураг 4. Trie хайлтын алгоритм

- Trie – н хайлт тухайн жагсаалтанд агуулагдах хуулийн томьёоны ID болон тэмдэгт мөрний аль байрлал дээр олдсон эсэх мэдээлэлтэй жагсаалтыг буцаана.
- Жагсаалтын сүүлийн элемент эсэхийг шалгана. Сүүлийн элемент бол 10 алхам, үгүй бол 6 алхам руу шилжинэ.
- Нэр томьёоны загвараар хадгаллагдсан TSV файлаас тухайн нэр томьёоны сүүлийн үгийг олж авчирна.
- Олдсон байрлал дээрх үгийн ард талын үгийн язгуурыг олдсон нэр томьёоны сүүлийн үгийн язгууртай тэнцүү эсэхийг шалгана. Тийм бол 8 алхам, үгүй бол 5 алхам руу шилжинэ.
- Нэр томьёог жагсаалтад хадгалж авна.
- 5 алхам руу буцна.
- Тухайн ялгаж авсан жагсаалтан дээр нэг урттай үгнээс бүтэх хуулийн нэр томьёог хадгалсан Trie – н хайлтыг ажиллуулна.
- Trie – н хайлт тухайн жагсаалтанд агуулагдах хуулийн томьёоны ID болон тэмдэгт мөрний аль байрлал дээр олдсон эсэх мэдээлэлтэй жагсаалтыг буцаана.
- Жагсаалтын сүүлийн элемент эсэхийг шалгана. Сүүлийн элемент бол 17 алхам, үгүй бол 13 алхам руу шилжинэ.
- Нэр томьёоны загвараар хадгаллагдсан TSV файлаас тухайн нэр томьёоны сүүлийн үгийг олж авчирна.
- Олдсон байрлал дээрх үгийн язгуурыг олдсон нэр томьёоны сүүлийн үгийн язгууртай тэнцүү

эсэхийг шалгана. Тийм бол 15 алхам, үгүй бол 16 алхам руу шилжинэ.

15. Нэр томьёог жагсаалтад нэмж хадгална.
16. 12 алхам руу буцна.
17. Хадгалж авсан нэр томьёоны жагсаалтыг буцаана.

### III. ҮР 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 үзүүлээ. 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 яваа /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
олон нэмэгдэл ял, тэнсэх, албадлагын арга хэмж	N	байгаа	22	3	22
иглэж, хүүхдэд сургуулийн өмнөх боловсрол эзэ	N	сургалт	1	3	1
яваагаас бусад/ хүн, жагсаалаар яваа болон хөгж	N	зорчигч	10	2	10
? дахь хэсэгт зааснаар зургаан жил тутам явагдах	N	сонгууль	11	2	11

Зураг 5. 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 6. 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 7. Нэр томьёоны давтамжаар эрэмбэлбэл

### ДҮГНЭЛТ

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

### НОМ ЗҮЙ

- [1] Эрх зүйн мэдээллийн нэгдсэн систем 2023 (<https://legalinfo.mn/mn>)
- [2] Монгол хэл боловсруулалтын үзүүлэн хуудас. 2016, Machine Intelligent Laboratory, National University of Mongolia, (<http://172.104.34.197/nlp-web-demo/>)
- [3] Maabar, Maha (17 November 2014). "Trie Data Structure". CVR, University of Glasgow. Archived from the original on 27 January 2021. Retrieved 17 April 2022.