Монгол хэлний нийлмэл хэлцийн таниур

\*Монгол хэлний нийлмэл хэлцийн таниурын хэрэгжүүлэлт

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

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

Хураангуй – Орчин үеийн хиймэл оюун ухаан, машин сургалтын үе

Түлхүүр – Монгол хэлний нийлмэл хэлц, хэрэгжүүлэлт алгоритм

# 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.

# Арга зүй

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

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

1. Эрх зүйн мэдээллийн нэгдсэн систем(legalinfo.mn) – ийн “Хуулийн нэр томьёо” цэсэнд байрлах бүх хуулийн нэр томьёог татаж авчирна.
2. “ID, Хуулийн нэр томьёо, нэр томьёоны тодорхойлолт, үгийн аймгийн тэмдэглэгээ(ниймэл хэлц гэх таних тэмдэглэгээ), үндсэн үг(нийлмэл хэлцийн хамгийн сүүлийн үг)” гэх загвар(цаашид нэр томьёоны загвар гэх) – аар хадгална.
3. TSV (Tab-separated values) файл болгож хадгална.
4. Тухайн цэснээс хуулийн нэр томьёо болон хуулийн бүх DOCX өргөтгөлтэй файлуудыг татаж авчирна.
5. DOCX өргөтгөлтэй файлууд текст файл руу хөрвүүлнэ.
6. “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.
7. Үүсгэсэн хоёр TSV файлыг нэгтгэж нэгдсэн нэг TSV файл үүсгэнэ.
8. Хадгалж авсан хуулийн нэр томьёо, үгийг үндсээр задлаж өгөх API болон баримт бичгээс нийлмэл хэлц хайх алгоритмыг ашиглан хуулиуд дээр дурдагдсан нэр томьёонуудыг олж хадгална.

## Хуулийн баримт бичгийн хуулийн нэр томьёо хэсгээс нэр томьёо ялгаж авах алгоритм хэрэгжүүлэх

Оролт: Текст файл

1. Текст файлаас мөр унших
2. Мөрийг сүүлийн мөр эсэхийг шалгах. Сүүлийн мөр бол 6 алхам, үгүй бол 3 алхам руу шилжинэ.
3. Мөрөнд байгаа [**."**]эхэлсэн [**"\sгэж\s**] төгссөн загвартай үгсийг олно.
4. Тухайн өмнөх хэсгийн ”хуулийн томьёо”, харин хойш хэсгийн “нэр томьёоны тодорхойлолт” болгон нэр томьёоны загвараар хадгална.
5. 2 алхам руу шилжинэ.
6. Нэр томьёоны жагсаалтыг буцаана.

## Хадгалсан хуулийн нэр томьёог ашиглан текст файлаас хайлт хийх зохистой 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 хадгална. Жишээ нь, “үл хөдлөх эд хөрөнгө” – ийг “үл хөдлөх эд” гэж хадгална.

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

1. Монгол хэл боловсруулалтын үзүүлэн хуудас ([Mongolian NLP Demo](http://172.104.34.197/nlp-web-demo/)) хуудас руу хүсэлт шидэж, оруулсан тескт дэх үгсийн задаргааг авна.
2. Үгсийн задаргааны жагсаалтаас зөвхөн үгсийг ялгаж авна.
3. Тухайн ялгаж авсан жагсаалтан дээр 2 буюу түүнээс урттай үгнээс бүтэх хуулийн нэр томьёог хадгалсан Trie – н хайлтыг ажиллуулна.
4. Trie – н хайлт тухайн жагсаалтанд агуулагдах хуулийн томьёоны ID болон мөрөөс аль байрлал дээр олдсон эсэх мэдээлэлтэй жагсаалтыг буцаана.
5. ID жагсаалтын сүүлийн элемент эсэхийг шалгана. Сүүлийн элемент бол 10 алхам, үгүй бол 6 алхам руу шилжинэ.
6. Нэр томьёоны загвараар хадгаллагдсан TSV файлаас тухайн нэр томьёоны сүүлийн үгийг олж авчирна.
7. Олдсон байрлал дээрх үгийн ард талын үгийн язгуурыг олдсон нэр томьёоны сүүлийн үгийн язгууртай тэнцүү эсэхийг шалгана.
8. Нэр томьёог жагсаалтад хадгална.
9. 5 алхам руу буцна.
10. Тухайн ялгаж авсан жагсаалтан дээр нэг урттай үгнээс бүтэх хуулийн нэр томьёог хадгалсан Trie – н хайлтыг ажиллуулна.
11. Trie – н хайлт тухайн жагсаалтанд агуулагдах хуулийн томьёоны ID болон мөрөөс аль байрлал дээр олдсон эсэх мэдээлэлтэй жагсаалтыг буцаана.
12. ID жагсаалтын сүүлийн элемент эсэхийг шалгана. Сүүлийн элемент бол 11 алхам, үгүй бол 13 алхам руу шилжинэ.
13. Нэр томьёоны загвараар хадгаллагдсан TSV файлаас тухайн нэр томьёоны сүүлийн үгийг олж авчирна.
14. Олдсон байрлал дээрх үгий язгуурыг олдсон нэр томьёоны сүүлийн үгийн язгууртай тэнцүү эсэхийг шалгана.
15. Нэр томьёог жагсаалтад хадгална.
16. 12 алхам руу буцна.
17. Хадгалж авсан нэр томьёоны жагсаалтыг буцаана.

# 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 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.

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.

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.

##### 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 *(https://legalinfo.mn/mn)*