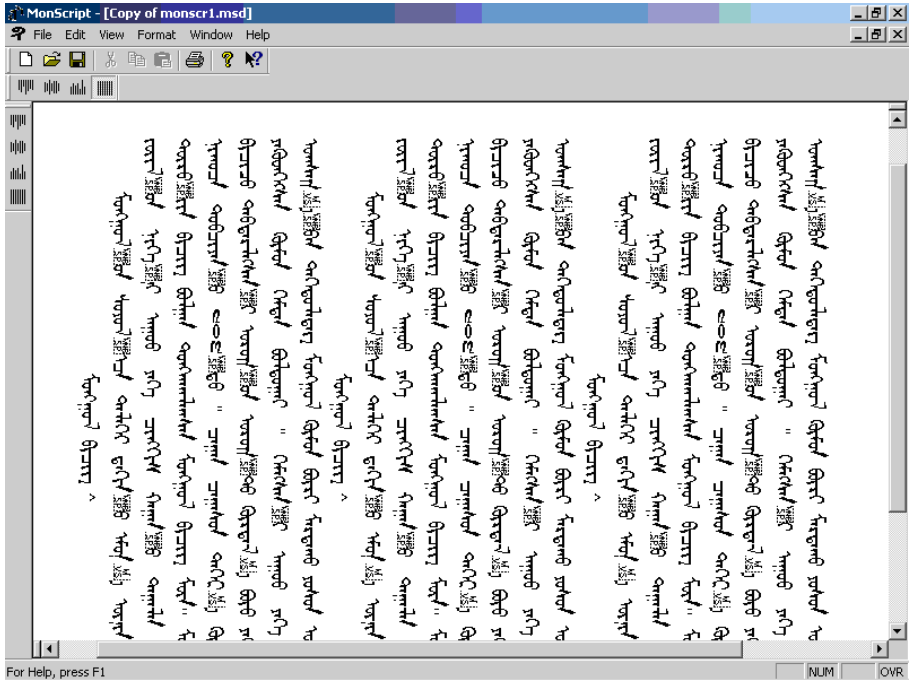


Текст боловсруулалтын үндэс

Лекц №2

Регуляр илэрхийлэл

2020 он

Регуляр илэрхийлэл

- Текстийн тэмдэгт мөрийг тодорхойлдог формал хэл
- Бид дараах үгсийн аль нэгийг хэрхэн олох вэ?
  - хэрэм
  - хэрэмнүүд
  - Хэрэм
  - Хэрэмнүүд

Регуляр илэрхийлэл: Тусгаарлалт (Disjunctions)

- [] хаалт доторх үсэг

Паттерн	Нийцэл(Match)
[xX]эрэм	хэрэм, Хэрэм
[1234567890]	дурын тоо

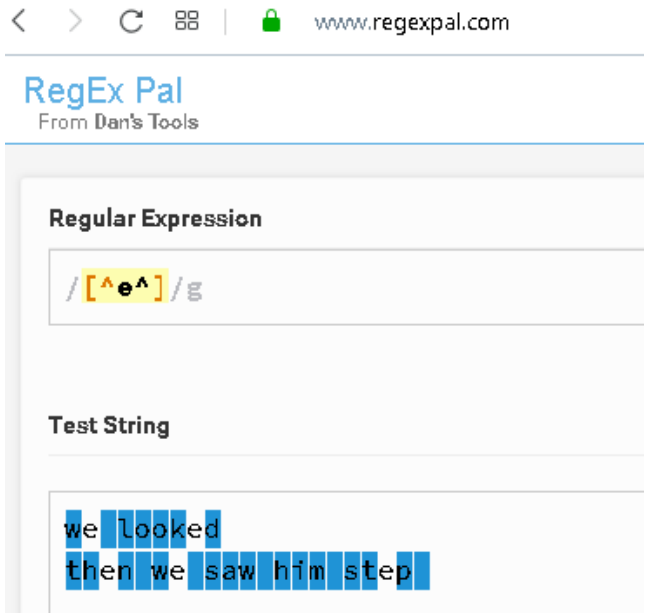
- Муж [A-Z]

Паттерн	Нийцэл	
[А-Я]	Том үсэг	Том үсэг
[а-я]	Жижиг үсэг	Жижиг үсэг
[0-9]	Нэг оронтой тоо	Гарчиг 1: Регуляр

Регуляр илэрхийлэл: Тусгаарлалт дахь үгүйсгэл

- Үгүйсгэл `[^Ss]`
  - Энэ нь хаалтан дахь үсгүүдээс бусад нь гэсэн үг

Паттерн	Нийцэл
`[^А-Я]`	Том үсэг биш
`[^Xx]`	X биш, мөн x биш
`[^e^]`	e болон ^ биш
`a^b`	a^b



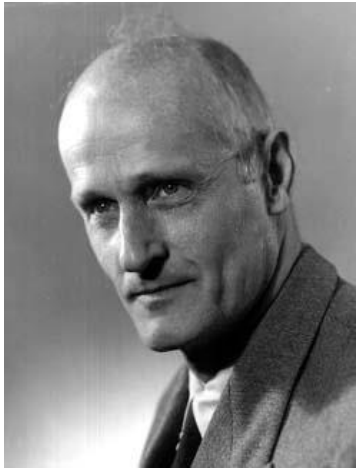
Регуляр илэрхийлэл: Бусад тусгаарлалтууд

- Эсвэл холбоос |

Паттерн	Нийцэл
`groundhog|woodchuck`	
`yours|mine`	yours mine
`a|b|c`	= [abc]
`[gG]roundhog|[Ww]oodchuck`	

Регуляр илэрхийлэл: ? * + .

Паттерн	Нийцэл	
colou?r	Өмнөх үсэг байхгүй байж болно	color colour
oo*h!	Өмнөх үсэг 0 эсвэл олон удаа давтагдаж болно	oh! ooh! oooh! ooooh!
o+h!	Өмнөх үсэг 1 эсвэл олон давтагдаж болно	oh! ooh! oooh! ooooh!
baa+		baa baaa baaaa baaaaa
beg.n	Ямар нэгэн	begin begun begun beg3n



Stephen C Kleene

Kleene *, Kleene +

Регуляр илэрхийлэл: ^ -эхлэх $ -төгсөх

Паттерн	Нийцэл
^[A-Z]	Palo Alto
^[^A-Za-z]	1 “Hello”
\.$	The end.
.$	The end? The end!

Жишээ

- Текст дэх “the” артиклийн бүх тохиолдол олох.

```
The
```

Буруу жишээ

```
[tT]he
```

`other` эсвэл `theology` буруу үр дүн буцаана

```
[^a-zA-Z][tT]he[^a-zA-Z]
```

Алдаа

- Хоёр төрлийн алдааг засах дээр суурилах ёстой
  - Нийцэх ёсгүй нийцсэн тэмдэгт мөрүүд (there, then, other)
    - False positives (Type I)
  - Нийцэх ёстой нийцээгүй тэмдэгт мөрүүд (The)
    - False negatives (Type II)

Алдаа

- ЭХ боловсруулалтад энэ төрлийн алдаануудтай байнга ажилладаг.
- Програмын алдааны түвшинг багасгахын тулд хоёр эсрэг оролдлогыг авч үздэг:
  - Зөв утгатай хэр ойр(accuracy) эсвэл хэр давтамжтай (precision) байгааг нэмэгдүүлэх нь (false positives хамгийн бага байлгахад туслана)
  - Хамрах хүрээ эсвэл recall -ыг нэмэгдүүлэх нь (false negatives хамгийн бага байлгахад тусална)

Дүгнэлт

- Регуляр илэрхийллийн үүрэг их
  - регуляр илэрхийллүүд нь текст боловсруулах текстүүдийн хувь эхний модел байдаг
- Хүнд бодлогуудын даалгаварын хувьд, машин сургалтын ангилагчуудыг ашиглана
  - Гэвч регуляр илэрхийлэл ангилагч дотор өргөн хэрэглэгддэг
  - Ерөнхий тохиолдлыг илрүүлэхэд хэрэгтэй