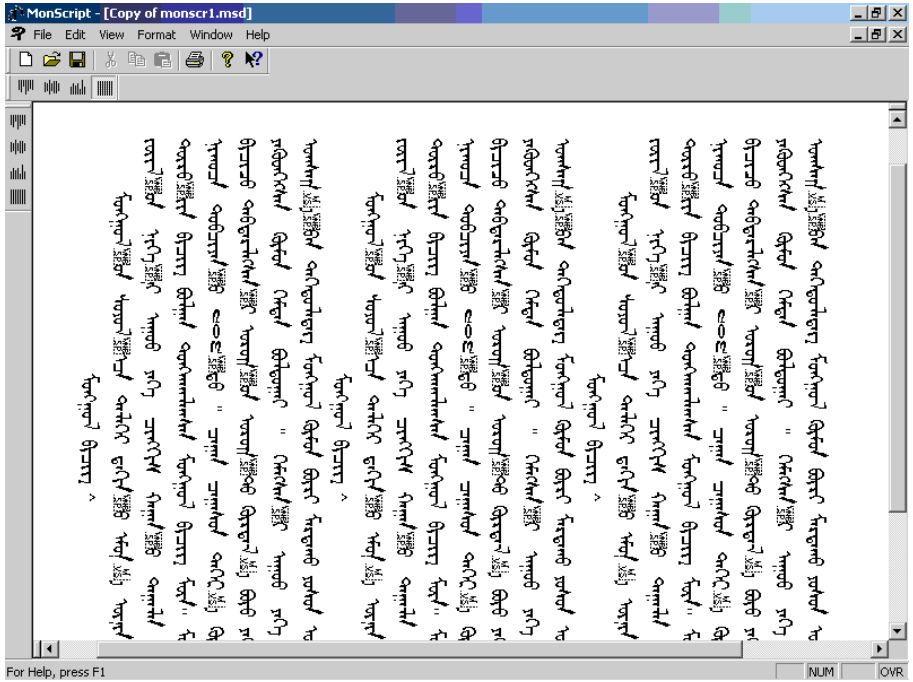


Текст боловсруулалтын үндэс

Лекц №3

Үгээр токенчлол буюу салгах

Word tokenization

2020 он

Текст нормчлол (normalization)

- ЭХБ –ын даалгавар бүрт текст нормчлол шаардлагатай байдаг:
  1. Текстээс үгийг сегментлэх/токенчлох
  2. Үг бүрийн хэлбэржүүлэлтийг нормчлох
  3. Текстээс өгүүлбэрүүдийг сегментлэх

Хэдэн үг байна вэ?

- Seuss’s cat in the hat is different from other cats!
  - **Лемма**: ижил үгийн үндэс(stem), үгийн аймаг тодорхойлох, эхний байдлаар үгийн утга тодорхойлох
    - cat болон cats = ижил лемма
  - **Үгийн хэлбэр**: бүрэн хувирсан үгийн бичиглэлийн хэлбэр
    - cat болон cats = ялгаатай үгийн хэлбэр

Хэдэн үг байна вэ?

they lay back on the San Francisco grass and looked at the stars and their

- **Төрөл зүйл**: үгийн сангийн нэг элемент.
- **Токен**: текст дэх төрөл зүйлийн тохиолдол.
- Хэд?
  - 15 токен (эсвэл 14)
  - 13 төрөл зүйл (эсвэл 12) (эсвэл 11?)

Хэдэн үг байна вэ?

N = токены тоо

V = үгийн сан = төрөл зүйлийн олонлог

$|V|$ *үгийн сангийн хэмжээ*

Church болон Gale (1990): $|V| > O(N^{½})$

	Токен = N	**Төрөл = |V|**
Утасны харилцааны мэссэж	2.4 сая	20 мян
Шекспир	884,000	31 мян
Google N-grams	1 трил	13 сая

UNIX систем дэх энгийн токенчлол

- (Inspired by Ken Church's UNIX for Poets.)
- Текст файл өгөхөд, гаралтанд үгэн токен болон түүний давтамж

```
tr -sc 'A-Za-z' '\n' < shakes.txt
      | sort
      | uniq -c
```

Үсгэн биш бүгдийг шинэ мөр болгох

Эрэмбэлэх үсгийн дараалал

Төрөл бүрээр нэгтгэж тоолох

```
1945 A
  72 AARON
  19 ABBESS
   5 ABBOT
... ...
```

```
  25 Aaron
   6 Abate
   1 Abates
   5 Abbess
   6 Abbey
   3 Abbot
... …
```

1-р алхам: токенчлох

```
tr -sc 'A-Za-z' '\n' < shakes.txt | head

THE
SONNETS
by
William
Shakespeare
From
fairest
creatures
We
...
```

2-р алхам: эрэмблэх

```
tr -sc 'A-Za-z' '\n' < shakes.txt | sort | head

A
A
A
A
A
A
A
A
A
A
...
```

3-р алхам: тоолох

- Том болон жижиг үсгийг нэгтгэх

```
tr ‘A-Z’ ‘a-z’ < shakes.txt | tr –sc ‘A-Za-z’ ‘\n’ | sort | uniq –c
```

- Давтамжаар эрэмблэх

```
tr ‘A-Z’ ‘a-z’ < shakes.txt | tr –sc ‘A-Za-z’ ‘\n’ | sort | uniq –c | sort –n –r
```

```
23243 the
22225 i
18618 and
16339 to
15687 of
12780 a
12163 you
10839 my
10005 in
8954  d
```

Токенчлоход гардаг асуудлууд

- Finland’s capital → Finland Finlands Finland’s ?
- what’re, I’m, isn’t → What are, I am, is not
- Hewlett-Packard → Hewlett Packard ?
- state-of-the-art → state of the art ?
- Lowercase → lower-case lowercase lower case ?
- San Francisco → 1 токен уу? 2 токен уу?

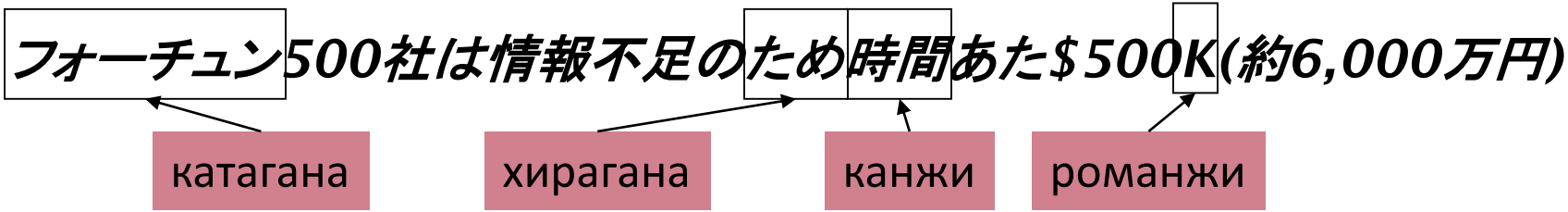
Токенчлох: хэлний хүндрэл

- Франц
  - ***L'ensemble*** → нэг токен уу эсвэл хоёр токен уу?
    - ***L*** ? ***L'*** ? ***Le*** ?
    - ***l'ensemble*** –ийг ***un ensemble*** **-**тэй нийцүүлэх үү?

- Герман хэлний нэр үгийн нийлэмж (combound) сегментлэгддэггүй
  - ***Lebensversicherungsgesellschaftsangestellter***
  - “Даатгалын компаний ажилтны амьдрал” гэсэн үг
  - Герман мэдээллийн хайлтанд ***нийлэмжийг салгагч*** шаардлагатай

Токенчлох: хэлний хүндрэл

- Хятад болон Япон хэлэнд үгийн хооронд зай байдаггүй:
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
  - Шарапова одоо амьдардаг -д АНУ хойд хэсэгт Флорида
- Мөн Япон хэлэнд хоорондоо холбоотой олон цагаан толгой байдаг
  - Он сар өдөр/тоо хэмжээ олон хэлбэртэй



Эцсийн хэрэглэгч асуултаа хираганагаар илэрхийлж болно!

Хятад хэлний үгийн токенчлол

- Мөн **үгийн сегментлэл** гэж хэлдэг.
- Хятад үгс тэмдэгүүдээс бүрдэнэ
  - Тэмдэг нь ерөнхийдөө 1 үе(syllable) болон 1 бүтээвэр(morpheme) байдаг.
  - Нэг үг дунджаар 2.4 тэмдгийн урттай байдаг.
- Стандарт суурь сегментлэх алгоритм:
  - Maximum Matching (Greedy гэж бас нэрлэдэг)

Maximum Matching
үг сегментлэх алгоритм

- Хятад үгсийн нэг жагсаалт болон нэг тэмдэгт мөр өгөгдөнө.

1) Тэмдэгт мөрийн эхнээс заагч эхэл
2) Заагчийн эхлэлээс тохирох толин дахь хамгийн урт үгийг ол
3) Олсон үгийн дараах руу заагчийг шилжүүл
4) 2 –р алхам руу оч

Max-match сегментлэлийн дүрслэл

- Thecatinthehat — the cat in the hat
- Thetabledownthere — the table down there / theta bled own there
- Англи хэлэнд ерөнхийдөө ажилладаггүй!

- Харин Хятад хэлэнд гайхалтай ажилладаг
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
- Орчин үед магадлалаар сегментлэх алгоритм илүү дээр байна.