Проекты

Кирилл Карпенко
GitHub: mssrchapelier

Чтобы ссылки стали активными при просмотре этого файла на GitHub, пожалуйста, загрузите файл.

① Классификация новостных заголовков
(амхарский язык, Эфиопия)

Google Colab

የሶማሊያው ሐሰን ሼኽ ሞሃሙድ ቃለ መሃላ ፈፀሙ
(Somalia's Hassan Sheikh Mohamud swears in)
→ **world news**

ኢትዮጵያ በአፍሪካ ዋንጫ የግጣርያ ጨዋታ ግብፅን 2 - 0 አሸነፈች
(Ethiopia defeats Egypt 2-0 in Africa Cup of Nations qualifiers)
→ **sports**

Технологии:
- Python
- NumPy, pandas
- SciPy
- matplotlib, seaborn, statannotations

② Внутритекстовые ссылки: конвертация формата

GitHub

```
... крупнейшие неразмеченные включают корпус Walta Information Center – WIC (описан в [4])... и корпус An Crúbadán [30] в 17 миллионов слов...

[4] Atelach Alemu Argaw and Lars Asker. An Amharic stemmer... In Proceedings ... – ACL 2007, pages 104–110, 2007.
```

→

```
... крупнейшие неразмеченные включают корпус Walta Information Center – WIC (описан в [Argaw and Asker 2007])... и корпус An Crúbadán [Scannell 2007] в 17 миллионов слов...

Argaw and Asker 2007 – Atelach Alemu Argaw and Lars Asker. An Amharic stemmer... In Proceedings ... – ACL 2007, pages 104–110, 2007.
```

Технологии: Python, regex

③ Морфологический парсинг
(язык тигре, Эфиопия/Эритрея)

GitHub

ለመጽአው →
```
la  - maS>    - aw
DEF - mS>:PRF - 3.M.PL
```

እትሕጹይ →
```
>t   - HSuy
CAUS - (HSy).PTCP.PASS.M.SG
```

Технологии: Java, regex

④ Затекстовые ссылки: сегментация

GitHub

```
Аркадьев П. М. О некоторых особенностях склонения в адыгских языках // Плунгян В. А. (отв. ред.). Язык. Константы. Переменные: Памяти Александра Евгеньевича Кибрика. СПб.: Алетейя, 2014. С. 552–563.
```

→

```
{
    "pagination": "С. 552–563",
    "city": "СПб.",
    "year": "2014",
    "collection-editors": "Плунгян В. А. (отв. ред.)",
    "collection-title": "Язык. Константы. Переменные: Памяти Александра Евгеньевича Кибрика",
    "publishers": "Алетейя",
    "article-title": "О некоторых особенностях склонения в адыгских языках",
    "authors": "Аркадьев П. М."
}
```

Технологии: Java, regex