Токенизация

[Разделение] [текста] [на] [слова]

Из курса Д. Скоринкина для ЦМГН

Классическая цепочка обработки текста:

- Сегментация на предложения
- Токенизация (разделение на слова)
- Морфологический анализ
- Синтаксический анализ
- Семантический анализ

Легко ли разбить текст на слова?

В одной из отдаленных наших губерний находилось имение Ивана Петровича Берестова. В молодости своей служил он в гвардии, вышел в отставку в начале 1797 года, уехал в свою деревню и с тех пор оттуда не выезжал. Он был женат на бедной дворянке, которая умерла в родах, в то время как он находился в отъезжем поле. Хозяйственные упражнения скоро его утешили. Он выстроил дом по собственному плану, завел у себя суконную фабрику, утроил доходы и стал почитать себя умнейшим человеком во всем околотке, в чем и не прекословили ему соседи, приезжавшие к нему гостить с своими семействами и собаками.

Легко:

В / одной / из / отдаленных / наших / губерний / находилось / имение / Ивана / Петровича / Берестова / В / молодости / своей / служил / он / в / гвардии / вышел / в / отставку / в / начале / 1797 / года / уехал / в / свою / деревню / и / с / тех / пор / оттуда / не / выезжал / Он / был / женат / на / бедной / дворянке / которая / умерла / в / родах / в / то / время / как / он / находился / в / отъезжем / поле / Хозяйственные / упражнения / скоро / его / утешили / Он / выстроил / дом / по / собственному / плану / завел / у / себя / суконную / фабрику / утроил / доходы / и / стал / почитать / себя / умнейшим / человеком / во / всем / околотке / в / чем / и / не / прекословили / ему / соседи / приезжавшие / к / нему / гостить / с / своими / семействами / и / собаками /

Пример посложнее:

Объявление

Продается ВАЗ-2109(1) 1997 г.в. Стоимость автомобиля — 300 000 руб. без торга. Пробег 50000км. Машина зверь, любимая ласточка, не подводила ни разу!!! Продаю, т.к. с деньгами край.

Тел +7 (956) 356 70 83 (Даниил Савельич)

адр. г. Москва, ул. Яблочкова, д. 25. кв. 7

В английском свои проблемы

Beatles, Don’t pass me by, припев:

Don't pass me by, don't make me cry,
don't make me blue.
'Cause you know, darling, I love only you.
You'll never know it hurt me so,
how I hate to see you go.
Don't pass me by, don't make me cry.

В английском свои проблемы

Beatles, Don’t pass me by, куплет:

I hear the clock a'ticking on the mantelshelf,
See the hands a'moving, but I'm by myself.
I wonder where you are tonight,
oh why I'm by myself.
I don't see you.
Does it mean you don't love me any more?

Сложности:

- Danya’s giving a great lecture
- Danya **is** giving a great lecture
- Danya’s given a great lecture today
- Danya **has** given a great lecture today
- Danya’s lecture was great
- ???
- Надо ли разделять? Является ли это задачей токенизатора?

Еще сложные случаи

The immediate roots of rock and roll lay in the rhythm and blues, then called "race music", and country music of the 1940s and 1950s.
Здесь я персона нон грата, потому ни должностей, ни нормальной работы мне не видать.
Поезд Москва–Казань. Расписание и маршрут движения. Цена ж/д билета от 958 руб.
в настоящее время в/ч расформирована
Эти события не помешали новгородцам выгнать зимой 1240/1241 годов Александра в Переяславль-Залесский

А еще...

Раньшеписалибезпробелов!
Человекунесложноэтопрочитать:)

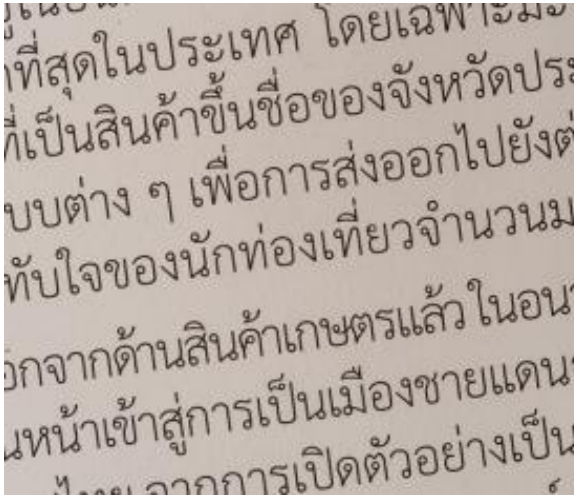
Пробелов не было во всех письменных языках

См. также:

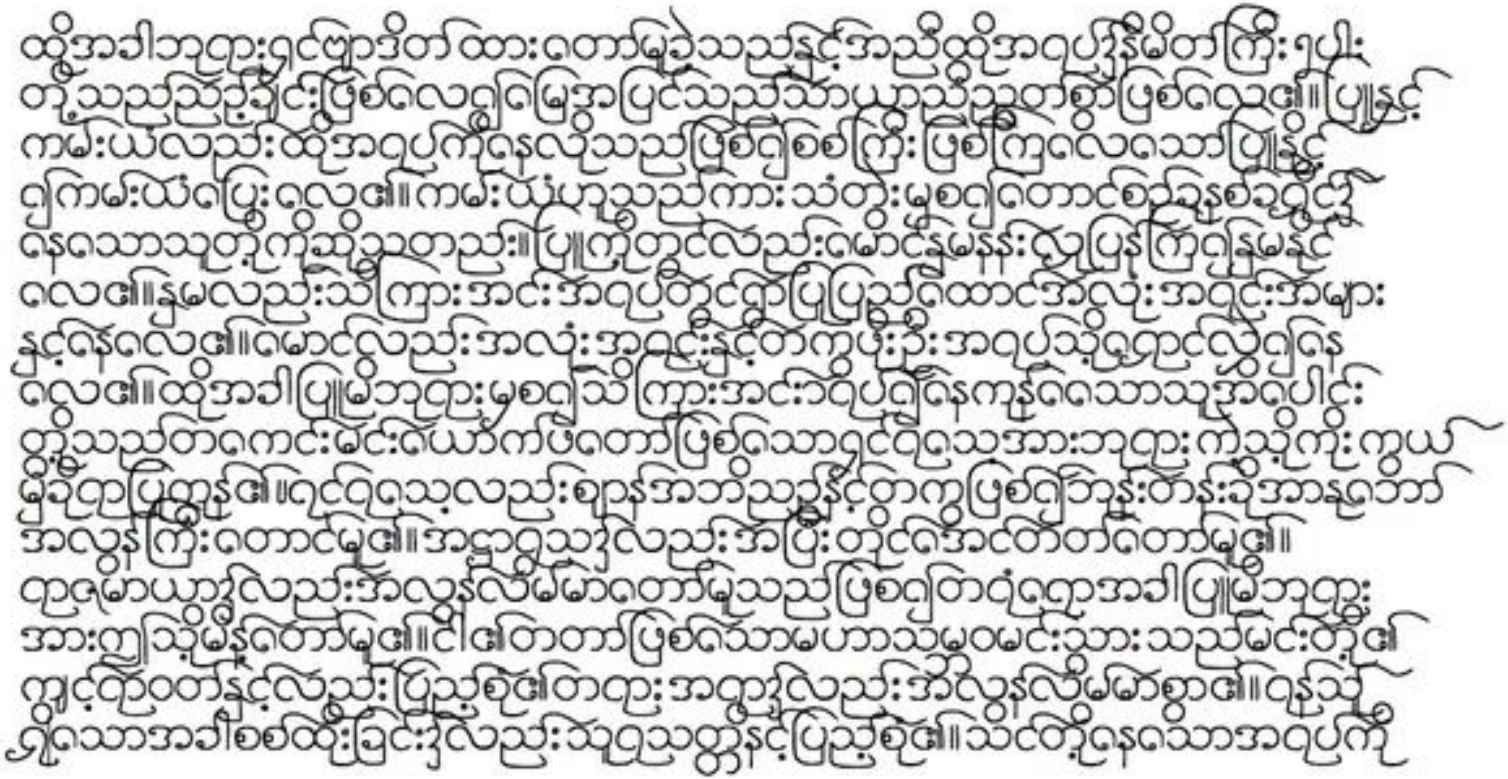
- ВНАЧАЛЕПРОБЕЛОВНЕБЫЛО (https://www.publish.ru/articles/199704_4041105)
- Шунейко А. «И надо оставлять пробелы…» // Наука и жизнь, 10, 2016
- Всё (или почти всё) о пробеле (https://habr.com/post/23250/)

А есть ли такие языки сегодня?

А есть ли такие языки сегодня?



Или, например, бирманский



Как токенизируют такие языки?

Самый простой и известный алгоритм:

MaxMatch

```
function MAXMATCH(sentence, dictionary) returns word sequence W

    if sentence is empty
        return empty list
    for i ← length(sentence) downto 1
        firstword = first i chars of sentence
        remainder = rest of sentence
        if InDictionary(firstword, dictionary)
            return list(firstword, MaxMatch(remainder,dictionary) )

    # no word was found, so make a one-character word
    firstword = first char of sentence
    remainder = rest of sentence
    return list(firstword, MaxMatch(remainder,dictionary) )
```

Figure 2.11 The MaxMatch algorithm for word segmentation.

MaxMatch works very well on Chinese; the following example shows an application to a simple Chinese sentence using a simple Chinese lexicon available from the Linguistic Data Consortium:

Input: 他特别喜欢北京烤鸭 "He especially likes Peking duck"
Output: 他 特别 喜欢 北京烤鸭
He especially likes Peking duck

Описание алгоритма MaxMatch:

Смотрим на предложение.

Если оно **пустое**: возвращаем любой условный символ конца предложения

Если **не пустое**: ищем **самое длинное слово в словаре**, которое **совпадает с началом** предложения

Если нашли: возвращаем **это слово** + любой разделитель слов + результат **применения maxmatch к остатку** предложения (без этого слова)

Если начало не совпало ни с одним словом в словаре: Возвращаем **первый символ** + любой разделитель слов + результат применения maxmatch к остатку предложения без первого символа