Предобработка текста (Preprocessing)

Корпус

Корпус — коллекция текстов, отобранных по некоторому принципу и специально обработанных. Корпус может служить как базой для поиска, так и материалом для создания лингвистических технологий.

Корпус

Корпус — коллекция текстов, отобранных по некоторому принципу и специально обработанных. Корпус может служить как базой для поиска, так и материалом для создания лингвистических технологий.

Национальный корпус русского языка
ruscorpora.ru

Особенности корпуса

Объем
Текстовое наполнение
Разметка (тип разметки, наличие снятой омонимии)
Выравнивание (для параллельных)
Доступность
Формат

Особенности корпуса

Объем
Текстовое наполнение
Разметка (тип разметки, наличие снятой омонимии)
Выравнивание (для параллельных)
Доступность
Формат

Устройство корпуса во многом зависит от задачи

Предобработка

Скачали много текстов, с чего начать?

Предобработка

Скачали много текстов, с чего начать?

Разбиение текста на предложения

Разбиение текста на токены

Нормализация

Разбиение на предложения

Сложности

- ! , ? могут выступать в разном значении/функции
- Точка “.” оказывается достаточно проблемным случаем

Разбиение на предложения

Сложности

- ! , ? могут выступать в разном значении/функции
- Точка “.” оказывается достаточно проблемным случаем
  - Граница предложения
  - Сокращения, например, Inc. or Dr.
  - В записи чисел, например, .02% или 4.3
  - В датах, например, 15.01.2019

Сколько котов?

Кот Пушистик отличался от других **котов** своим характером.

Сколько котов?

Кот Пушистик отличался от других **котов** своим характером.

Лемма: одна основа, одно и то же значение
кот и котов — одна лемма

Словоформа: грамматическая форма слова
кот и котов — разные словоформы

Сколько слов?

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

Слово: единица словаря (an element of vocabulary).

Token: минимальный сегмент

➢ *Сколько слов? Сколько токенов?*

Проблемы токенизации

Finland’s capital

Проблемы токенизации

Finland’s capital Finland Finlands Finland’s ?

Проблемы токенизации

Finland’s capital Finland Finlands Finland’s ?

what’re, I’m, isn’t

Проблемы токенизации

Finland’s capital	Finland Finlands Finland’s ?
what’re, I’m, isn’t	What are, I am, is not

Проблемы токенизации

Finland’s capital	Finland Finlands Finland’s ?
what’re, I’m, isn’t	What are, I am, is not
Hewlett-Packard	

Проблемы токенизации

Finland’s capital	Finland Finlands Finland’s ?
what’re, I’m, isn’t	What are, I am, is not
Hewlett-Packard	Hewlett Packard ?

Проблемы токенизации

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	

Проблемы токенизации

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 ?

Проблемы токенизации

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	

Проблемы токенизации

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 ?

Проблемы токенизации

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	

Проблемы токенизации

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	один токен или два?

Проблемы токенизации

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	один токен или два?
м.н.с, к. ф. н.	??

Проблемы токенизации

Французский

L'ensemble один токен или два?

L ? L’ ? Le ?

Хотим l’ensemble соотнести с un ensemble

Проблемы токенизации

Французский

L'ensemble один токен или два?

L ? L’ ? Le ?

Хотим l’ensemble соотнести с un ensemble

Составные существительные в немецком

Lebensversicherungsgesellschaftsangestellter

‘life insurance company employee’

Разбиение сложных слов

Проблемы токенизации

Отсутствие пробелов в китайском

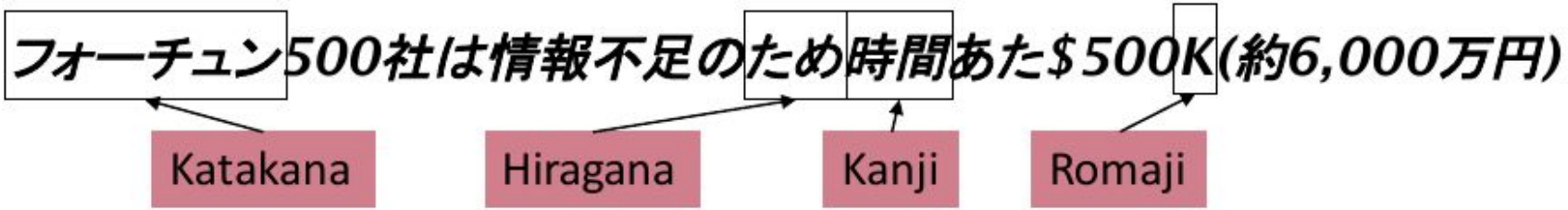
- 莎拉波娃现在居住在美国东南部的佛罗里达。
- 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
- Sharapova now lives in US southeastern Florida

Проблемы токенизации

Отсутствие пробелов в китайском

- 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
- Sharapova now lives in US southeastern Florida

Смешение разной письменности в японском при записи дат и чисел



Нормализация

Приведение единиц текста к единому формату

- Избавление от знаков пунктуации.

  Например, ***U.S.A. vs. USA***
- Приведение к нижнему регистру, но надо быть осторожными,

  Например, ***US* vs. *us***
- Лемматизация, реже стемминг (приведение к единой словарной форме).

  Например, *one black **cat** and two white **cats***

Регулярные выражения (Regular Expressions)

Регулярные выражения

Символы в квадратных скобках [...] находят один из перечисленных символов.
[aA]pple найдет apple и Apple,
[1238]kitty найдет 1kitty, 2kitty, 3kitty, 8kitty

Регулярные выражения

Символы в квадратных скобках [...] находят один из перечисленных символов.
[aA]pple найдет apple и Apple,
[1238]kitty найдет 1kitty, 2kitty, 3kitty, 8kitty

Символы в квадратных скобках через дефис задают диапазон [...-...].
Находят один символ из диапазона.
во[а-я] найдет вон, вот, воу, вор и т.д.,
[A-Z]at найдет Cat, Eat, Fat, Bat, Pat и т.д.
Можно комбинировать: [а-яА-Яbq]

Регулярные выражения

Символы в квадратных скобках [...] находят один из перечисленных символов.
[aA]pple найдет apple и Apple,
[1238]kitty найдет 1kitty, 2kitty, 3kitty, 8kitty

Символы в квадратных скобках через дефис задают диапазон [...-...].
Находят один символ из диапазона.
во[а-я] найдет вон, вот, воу, вор и т.д.,
[A-Z]at найдет Cat, Eat, Fat, Bat, Pat и т.д.
Можно комбинировать: [а-яА-Яbq]

Отрицание задается ^ внутри квадратных скобок
[^0-9a-zA-Z]kitty найдет !kitty, &kitty, *kitty, но не 8kitty или Akitty

Задание

Найти все последовательности union и Union

Найти все последовательности Was, was, War, war

Найти все упоминания 20-х годов 19го века (1920, 1921, 1923 и т.д)

Найти все упоминания 20-х годов 19го века (1920, 1921, 1923 и т.д), после которых нет пробела.

Регулярные выражения

Операторы

* — предыдущий символ, повторенный 0 и более раз

oo*h! найдет oh! ooh! oooh! ooooh!

+ — предыдущий символ, повторенный 1 и более раз

o+h! найдет oh! ooh! oooh! ooooh!

? — наличие или отсутствие предыдущего символа

colou?r найдет color и colour

. — задает один любой символ

beg.n найдет begin, began, begun, beg3n

Не забываем про неоднозначность

Экранирование метасимволов: \[, \], \?, \., *, \(, \)

Задание

1. Найдите буквенные последовательности символов более 1
2. Найдите последовательности ***o любой символ o,*** *например, p**olo,** s**o o**ver*
3. Найдите последовательности любой символ и точка
4. Найдите последовательности it и its

Задание

1. Найдите буквенные последовательности символов более 1
2. Найдите последовательности ***o любой символ o,*** *например, p**olo,** s**o o**ver*
3. Найдите последовательности любой символ и точка
4. Найдите последовательности it и its

\b обозначает границу

5. Сделайте так, чтобы предыдущее регулярное выражение находило местоимения it и its

6. Найдите слова, заканчивающиеся на ion

Регулярные выражения

Начало и конец строки

^[A-Z]	**P**alo Alto
^[^A-Za-z]	**“**Hello”
\.$	The end**.**
.$	The end**?** The end**!**

Регулярные выражения

\s - любой пробельный символ (пробел, новая строка, табуляция)

\S - не пробел

\w - слово (латиница!)

\W - не слово

\n - новая строка (на Windows \r\n)

\t - табуляция

\b - граница слова

\B - не граница слова

\d - цифра