Кому: С.Ю. Толдова, Е.А. Кузьменко

Копия: Компьютерные лингвисты, 3 курс

Тема: **Токенизация электронных писем**

От: Бибаева Мария, Картозия Инга, Мельник Анастасия

Тема: Input

- Корпус из **электронных писем** курсовой почты (linghse2014@gmail.com)
- **Объём корпуса:** 3096 писем, 195307 словоупотреблений
- **Языки:** русский, английский.

Token 149: 'например'
Type: cyr
Document: letter7

Token 184: 'obubnenkova@hse.ru'
Type: mail
Document: letter7

Token 156075: '#тыжлингвисте'
Type: cyr hashtag
Document: letter2565

Тема: Что мы считаем за токен?

- слова ЕЯ
- числа
- буквенно-цифровые комплексы:12-ти
- даты и время: множество форматов
- номера телефонов
- адреса: Старая Басманная, ул.
- почтовые адреса
- ссылки

- **Инициалы:** А.Б., МА, Н. Р.
- **Адреса:** ул. Старая Басманная, 21/4, каб. 304
- **Почтовые адреса:** linghse2014@gmail.com
- **Ссылки:** https://www.hse.ru/ba/ling/
- **Номера телефонов:** 8(495) 772-95-90*222-80

Тема: **Pre-text output**

Token 113: '№11-23-0111'
Type: date
Document: letter0

Token 376: '@user'
Type: lat @-tag
Document: letter0

Token 39: '«Диалог–2006».'
Type: name of something
Document: letter0

Token 352: 'google_color_link="CC6600'
Type: something
Document: letter0

Token 72: '33-летняя'
Type: cyr-num
Document: letter0

Точность, полнота и т.д. - пока не подсчитаны

Точность:

Полнота: