ПРЕПРОЦЕССИНГ

Квантитативный анализ текста

Кирилл Александрович Маслинский 14.02.2022 / 02

НИУ ВШЭ Санкт-Петербург

КАК СЧИТАТЬ СЛОВА

Чтобы оценить частотное распределение слов, нам необходимо посчитать количество употреблений (tokens) каждого слова (type) в тексте.

Qs:

- Что является токеном? (Что считать, а что не считать?)
- Какие токены считать одним и тем же словом?

Токены

Сколько токенов?

Ой какие фотки<smile006><smile006> A разве роды в 38недель не считаются нормой?

```
11? (разделим по пробелам)

Ой Какие

фотки<smile006><smile006>

А разве роды в З8недель не считаются нормой?
```

```
      11? (возьмем только слова)

      Ой какие фотки

      <smile006><smile006>< A</td>

      разве роды в 38 недель не

      считаются нормой ?
```

```
      13? (пунктуация тоже нужна)

      Ой какие фотки

      <smile006><smile006>< A</td>

      разве роды в 38 недель не

      считаются нормой ?
```

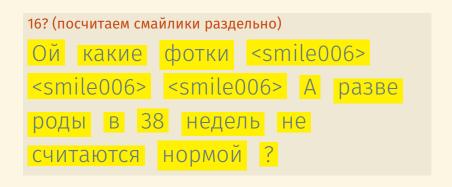
```
      14? (всё-таки исправим опечатку)

      Ой какие фотки

      <smile006><smile006>< A</td>

      разве роды в 38 недель не

      считаются нормой ?
```



N последовательно стоящих друг за другом слов.

Униграммы

Восторг внезапный ум пленил.

Биграммы

Восторг внезапный ум пленил <.>

Триграммы

N последовательно стоящих друг за другом слов.

Униграммы

Восторг внезапный ум пленил.

Биграммы

Восторг внезапный ум пленил <.>

Триграммы

N последовательно стоящих друг за другом слов.

Униграммы

Восторг внезапный ум пленил.

Биграммы

Восторг внезапный ум пленил <.>

Триграммы

N последовательно стоящих друг за другом слов.

Униграммы

Восторг внезапный ум пленил.

Биграммы

Восторг внезапный ум пленил <.>

Триграммы

N последовательно стоящих друг за другом слов.

Униграммы

Восторг внезапный ум пленил.

Биграммы

Восторг внезапный ум пленил <.>

Триграммы

Типы

How many types?

Кукушка кукушонку купила капюшон. Кукушонок в капюшоне смешон.

How many types?

Кукушка кукушонку купила капюшон. Кукушонок в капюшон-е смешон.

How many types?

Кукушка кукушон-ку купила капюшон. кукушон-ок в капюшоне смешон.

How many types?

КУКУШ-Ка КУКУШ-ОНКУ КУПИЛА

КАПЮШОН. КУКУШ-ОНОК В КАПЮШОНЕ

СМЕШОН.

How many types?

кукуш кукуш купи капюшон. кукуш в капюшон смеш.

МОРФОЛОГИЧЕСКИЙ АНАЛИЗ: ЛЕММАТИЗАЦИЯ

Сколько слов?

кукушка кукушонок купить капюшон. кукушонок в капюшон смешной.

лемматизация / lemmatization — приведение слова к начальной форме

Одно слово или разные?

Косил косой косой косой.

коса=S,жен,неод=твор,ед косая=S,жен,од=(род,ед|дат,ед|твор,ед|пр,ед) косой=S,муж,од=им,ед косой=A=(им,ед,полн,муж|род,ед,полн,жен| дат,ед,полн,жен|вин,ед,полн,муж,неод|твор,ед,полн,жен| пр,ед,полн,жен)

Омонимия языковых знаков

Одно слово или разные?

Косил косой косой косой.

косить=V,несов=прош,ед,изъяв,муж,пе косой=S,муж,од=им,ед косой=A=твор,ед,полн,жен коса=S,жен,неод=твор,ед

Морфологический анализ для русского

- Mystem (леммы, части речи, грамматические формы, снятие омонимии), прямо в R
- · udpipe (то же + синтаксические функции), прямо в R
- Stanza (то же + синтаксические функции), нейронная сеть: требуется python, медленный
- DeepPavlov (то же + синтаксические функции), нейронная сеть: требуется python, медленный

Терминология

- корпус здесь: исследуемая коллекция текстов token — словоупотребление, минимальный сегмент текста
- **словоформа / wordform** слово в тексте, измененное падеж, время и т.п.
- **лексема / lexeme** слово в словаре, совокупность всех форм
- **стемминг / stemming** урезание слова до основы **лемматизация / lemmatization** — приведение слова к начальной форме

Стоп-слова и частотные списки

Стоп-слова

Простейший способ уменьшить число лексических переменных — просто удалить наименее информативные слова:

- Статический список: без более бы был была были было быть в вам вас весь во вот все всего всех вы где да даже для ...
- Динамический список:
 - Слишком частотные (N самых частотных; частотность больше k)
 - Слишком редкие (частотность меньше к)
 - Слишком короткие (меньше М букв)
 - По документной частоте (присутствующие более чем в k% текстов или менее чем в k текстах)