

ПРЕПРОЦЕССИНГ

КВАНТИТАТИВНЫЙ АНАЛИЗ ТЕКСТА

Кирилл Александрович Маслинский

14.02.2022 / 02

НИУ ВШЭ Санкт-Петербург

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 фот-
ки<smile006><smile006><smile006> А
разве роды в 38недель не считаются
нормой?

11? (разделим по пробелам)

Ой какие

фотки<smile006><smile006><smile006>

А разве роды в 38недель не
считаются нормой?

11? (возьмем только слова)

Ой какие фотки

<smile006><smile006><smile006> А

разве роды в 38 недель не

считаются нормой ?

13? (пунктуация тоже нужна)

Ой какие фотки

<smile006><smile006><smile006> А

разве роды в 38 недель не

считаются нормой ?

14? (всё-таки исправим опечатку)

Ой какие фотки

<smile006><smile006><smile006> А

разве роды в 38 недель не

считаются нормой ?

16? (посчитаем смайлики отдельно)

Ой какие фотки <smile006>
<smile006> <smile006> А 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

<s> Восторг внезапный ум пленил <.>

N последовательно стоящих друг за другом слов.

Униграммы

Восторг внезапный ум пленил .

Биграммы

Восторг внезапный ум пленил <.>

Триграммы

<s> Восторг внезапный ум пленил <.>

N последовательно стоящих друг за другом слов.

Униграммы

Восторг **внезапный** ум пленил .

Биграммы

Восторг **внезапный** ум пленил <.>

Триграммы

<s> Восторг **внезапный** ум пленил <.>

N последовательно стоящих друг за другом слов.

Униграммы

Восторг внезапный **ум** пленил .

Биграммы

Восторг внезапный **ум пленил** <.>

Триграммы

<s> Восторг **внезапный ум пленил** <.>

N последовательно стоящих друг за другом слов.

Униграммы

Восторг внезапный ум **пленил** .

Биграммы

Восторг внезапный ум **пленил** <.>

Триграммы

<s> Восторг внезапный **ум пленил** <.>

Типы

How many types?

Кукушка кукушонку купила капюшон.
Кукушонок в капюшоне смешон.

стемминг / stemming — cut a word to its stem

How many types?

Кукушка кукушонку купила капюшон.

Кукушонок в капюшон-е смешон.

стемминг / stemming — cut a word to its stem

How many types?

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

стемминг / stemming — cut a word to its stem

How many types?

кукуш-ка кукуш-онку купила

капюшон. кукуш-онок в капюшоне
смешон.

стемминг / stemming — cut a word to its stem

How many types?

кукуш кукуш купи капюшон. кукуш в
капюшон смеш.

стемминг / stemming — cut a word to its stem

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)

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

- 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

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

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 k)
 - Слишком короткие (меньше M букв)
 - По документной частоте (присутствующие более чем в $k\%$ текстов или менее чем в k текстах)