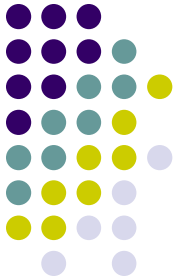


ГРАФЕМАТИЧЕСКИЙ АНАЛИЗ И СЕГМЕНТАЦИЯ ТЕКСТА

Большакова Елена Игоревна

СОДЕРЖАНИЕ

1. Функции и виды сегментации и графематического анализа и текста
2. Токенизация
3. Сегментация на предложения
4. Подходы к реализации сегментации:
   - ❖ на основе правил
   - ❖ на основе машинного обучения
5. Заключение

ЭТАПЫ АНАЛИЗА ТЕКСТА В МНОГОУРОВНЕВЫХ МОДЕЛЯХ

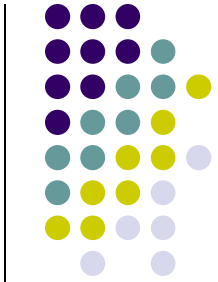
Уровни/Этапы анализа ~ Уровни языковой системы

1. Предобработка:
   Графематический анализ / Сегментация
2. Морфологический анализ: *лемматизация* (приведение словоформы к лемме, т.е. нормальной форме) + возможно, выявление ее морфопараметров
   или *стемминг*
3. Постморфологический анализ: разрешение морфологической омонимии

➢ 1-3 – начальные этапы анализа

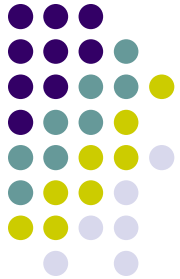
4. Синтаксический анализ – построение синтаксической структуры предложения
5. Семантический и дискурсивный анализ текста

ГРАФЕМАТИЧЕСКИЙ АНАЛИЗ



- Уровень: символы / знаки текста
  - *Графема* – минимальная неделимая единица текста (письменного): буква, знак препинания, спецзнак и др.: a, *A*, *a*, A, a, α, μ, ×, %
- Назначение графематич. анализа – посимвольная обработка текста, выделение и классификация нужных единиц : слов и их частей, предложений, абзацев, …
  - Сюда же относится предобработка текста, связанная с исключением незначащих элементов (знаков, графем): смайликов, HTML-тегов и др.
- В западной КЛ используется близкое понятие *сегментация* (*Segmentation*), т.е. разбиение текста на значимые части-сегменты
- Задачи сегментации зависят от конкретного ЕЯ: есть языки со слитным написанием слов, компаунды

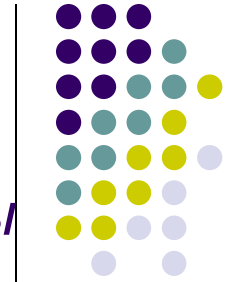
ВИДЫ СЕГМЕНТАЦИИ и ГРАФЕМАТИЧЕСКОГО АНАЛИЗА



- Токенизация (*tokenization: token* – кусочек)
  – обычно: выделение слов (псевдослов) в потоке знаков
- Разбиение/сегментация текста на предложения
- Морфологическая сегментация – разбиение слов (словоформ) на морфы (морфемы)
- Синтаксическая сегментация
- Выделение композиционных элементов текста, включая:
  - ✓абзацы и рубрики (списки)
  - ✓заголовки разделов и подразделов
  - ✓вставки, сноски, примечания, эпиграфы, грифы

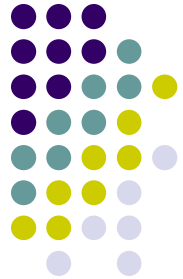
Установление иерархии этих элементов
(учет правил оформления документов)

СЛОВО В ЯЗЫКЕ И ТЕКСТЕ



- Важно: в тексте на русском языке – *словоформы*
- *Словоформа* – конкретная грамматическая форма: *стола, красного, забежал, eats*
- *Лексема* (единица языка) – совокупность всех словоформ слова: {*стол, стола, столу, столом, столы, столов, столам, столами* }
  по сути, семантический инвариант
- *Лемма* – нормальная (базовая, каноническая), словарная форма (*имя лексемы*) :
  - для существительных РЯ – ед. число, имен. падеж: *стол, пароход, ноутбук*
  - для глаголов – инфинитив: *писать*
  - для прилагательных РЯ:
    муж. род, ед. число, имен. падеж: *красный*

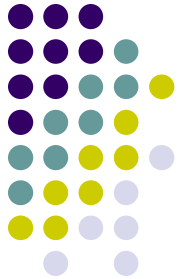
ТОКЕНИЗАЦИЯ



- Цель – выделение слов ЕЯ как носителей смысла
- В индоевропейских языках (включая русский) словоформы текста обычно выделяются пробелами, знаками препинания или др. разделителями
- В стандартном случае *Token* – это цепочка знаков текста между некоторыми разделителями, представляющая собой словоформу языка или *псевдослово*
- Внутри токена не м.б. пробелов, а пробел токеном не является, но токеном может быть любой непробельный символ, например, смайлик (если важен)
- На входе модуля токенизации:
  текст (последовательность символов)

  На выходе: этот текст, разбитый на токены +
  (желательно) классификация токенов

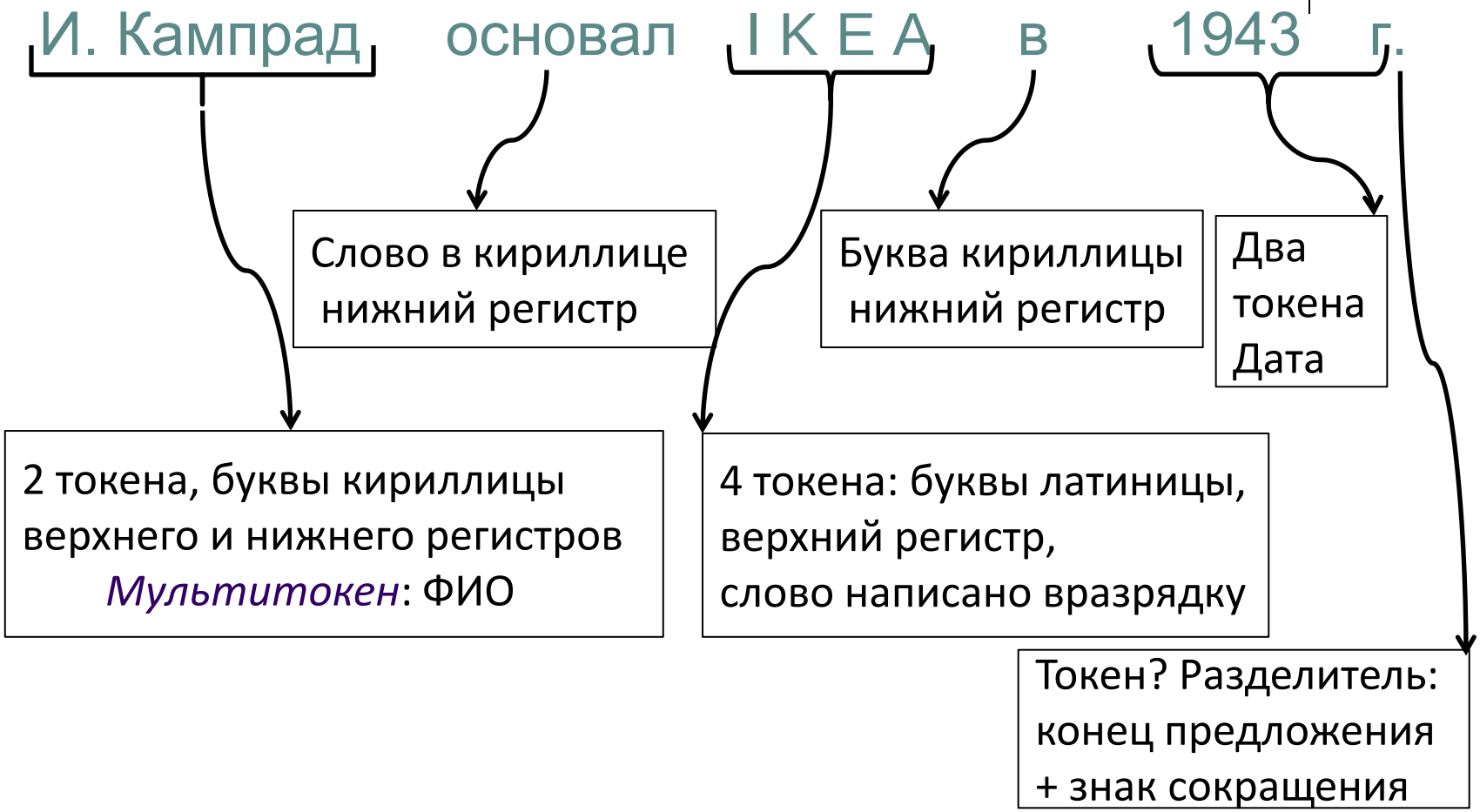
ВИДЫ ТОКЕНОВ



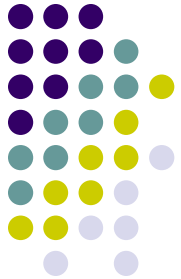
- Слова (словоформы) ЕЯ *деревом goes*
- Знаки препинания *? ; : . , ! « ’*
- Обозначения денежных единиц *$*
- Числа (разного вида) *786.9 1/4*
- Буквенно-цифровые комплексы *Boeing-747*
- Аббревиатуры: *ГОСТ*
- Даты (множество форматов) *23.01.2017*
- Обозначения времени *12:55*
- Номера телефонов *(495)555-77-99*
- Интернет-адреса , IP-адреса *http://yandex.ru*

TOKENIZATION: ПРИМЕР

Чаще всего есть разделитель – пробел



«НАИВНАЯ» ГРАФЕМАТИКА



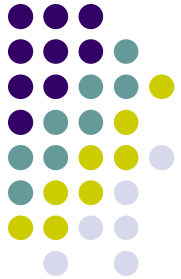
- Разбиение на токены (и предложения)
  – на основе пробелов, знаков препинания
    и м.б. других разделителей (тире, дефис)
  – с учетом вида символов: латиница, кириллица, цифры
- Средство разбиения: *RegExp* – регулярные выражения

Например, простые правила:

- Разделить входную строку по пробельным символам
- Отделить префиксные и постфиксные знаки пунктуации
  Пошли! *Исп. ¿Cómo te llamas?*
- Считать токенами подстроки, содержащие только буквы и цифры (*alphanumeric*)

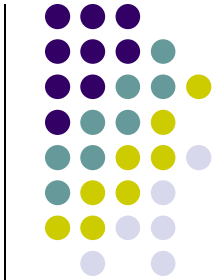
Для сложных текстов нередко требуется анализ контекста знаков, словарная информация

СЛОЖНЫЕ СЛУЧАИ



- Словоформы могут разделяться не только *and/or* пробелами: *больше/меньше героя-любовника*
- Слитное написание ряда слов : *Исп. Dímelo* (*Dí me lo*)
  четырестадвадцатьтретий Wörterbuch
- Разделители неоднозначны: *и т.д. г. Москва 15.00*
- Слова с небуквенными знаками: *куда-то Жан-Поль don't*
  - Слова с дефисами в РЯ: *веб-мастер царевна-лягушка Евровидение-2016 ха-ха*
  - Сокращения: *24-летний г-н Смит they're isn't boy's l'avion* (*le avion*)
- Ошибки: пропуск пробела /дефиса, лишний пробел:
  как бы как-бы что бы вебмастер
- Имена (ФИО) *В.И. Сидоров Иван В.Петров*
  Первичные токены в таких строках могут быть дополнительно преобразованы в *мультитокены*

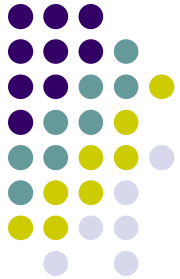
ПРОБЛЕМА: СОСТАВНЫЕ СЛОВОФОРМЫ



- Аналитические (составные) словоформы: степени прилагательных, формы глаголов:
  менее быстрый более ясно прочитал бы буду читать
- Составные числительные: *две третьих*
  триста двадцать восемь
- Сложные (составные) предлоги и союзы:
  потому что так как несмотря на
- Для обработки сложных случаев и соединения в мультитокены необходимо применение словарей:
  - сокращений *один млн 736 тыс*
  - имен, географических названий и др.
  - отдельных словоформ (числительных и др.)

  А в ряде случаев и морфологического анализа

СЕГМЕНТАЦИЯ НА ПРЕДЛОЖЕНИЯ



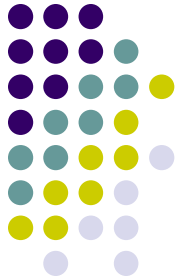
– На входе: текст (последовательность символов)
или последовательность токенов

– На выходе: текст, разбитый на предложения

Могут быть использованы:

- Маркеры конца предложения –
точка, вопросительный и восклицательный знаки
НО! неоднозначны из-за сокращений: *т.е. Dr. White*
и чисел: *2 марта в 11.40 он вышел из дома…*
- Маркер начала предложения – заглавная буква, **НО!**
тоже неоднозначный критерий, т.к. применяется в
именах и названиях: *Лена В.И. Сидоров*
- Маркеры цитат и прямой речи: *«Либо все люди должны быть счастливы, либо никто» (Р.Оуэн).*

ОПЯТЬ СЛОЖНЫЕ СЛУЧАИ



- Пропуски знаков-маркеров, в том числе ошибки
- Цитаты оформляются в разных ЕЯ по разному, причем кавычка в англ. языке не однозначна и служит для обозначения притяж. падежа и сокращений: *Ann's , it's*
- Особый случай – прямая речь:

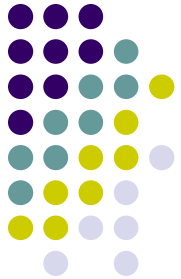
  Она сказала: «Был рассмотрен только первый вариант». Сколько здесь предложений?

  А потом позвонили зайчатки: - Нельзя ли прислать перчатки?

Для точной сегментации требуется анализ локального контекста символов обрабатываемого текста

Иванов А. смотрел на Петрова Б. Сидоров сидел в…

ПОДХОДЫ К СЕГМЕНТАЦИИ

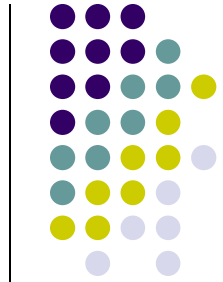


- Инженерный (*Rule-based*) подход использует:
  - лингвистические (эвристические) правила анализа знаков и их контекста (регистра букв и др.)
  - словари сокращений, имен, словосочетаний
  - регулярные выражения (*RegExp*) – встроены в современные языки программирования, что упрощает построение анализаторов

  Часто: более одного просмотра текста, дополнительная обработка токенов
- Машинное обучение по заранее размеченным текстам, это более эффективно для текстов с большим количеством токенов разного вида

❖ В обоих подходах нередко применяется одновременная сегментации на токены и предложения

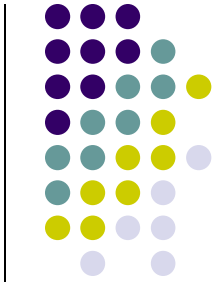
ПОДХОД НА ПРАВИЛАХ: ОБРАБОТКА ТОКЕНОВ



- Сборка устойчивых словосочетаний: *как бы*
  и слов, написанных в разрядку: *П о с т а н о в л е н и е …*
- Определение регистра букв и восстановление правильного (*truecase*) : *МОСКВА , москва → Москва*
- Различение тире и дефиса (– и -) *он - талант, сбор-ка*
- Приведение к единому формату дат и др. : *3 PM → 15.00*
- Перевод числительных в числа: *сто сорок пять* → 145
- Обработка сокращений слов и словосочетаний (разрывных и неразрывных): *г. Москва г-н Смит 1905 г. т.е. they're*
- Построение мультитокенов, в частности:
  - выделение полного имени (фамилия,имя,отчество), когда имя и отчество записано инициалами: *К.А. Смирнов*
  - Распознавание многословных имен: *ул. Кузнецкий мост*

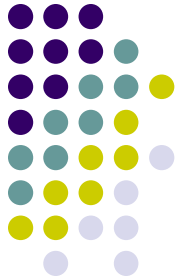
Применяются **словари** предлогов, имен, сокращений

ГРАФЕМАТИКА НА ПРАВИЛАХ В ПРОЕКТЕ ДИАЛИНГ-АОТ:



- Проект Диалинг-АОТ (*aot.ru)* лингвистического анализа русскоязычных текстов:
  - последовательная обработка текста: морфологический, синтаксический, семантич. анализ
  - программные модули на C++ с открытым кодом
  - инженерный (на правилах) подход к их построению
  - на сайте проекта есть текстовый онлайн-интерфейс
- Модуль графематического анализа строит таблицу токенов с *дескрипторами,* при этом выполняет
  - токенизацию
  - сегментацию на предложения
  - свертку устойчивых словосочетаний
  - анализ макроструктуры текста: выделение абзацев

МОДУЛЬ ГРАФЕМАТИКИ *aot.ru*



Вход модуля – текстовый файл (*плейн-текст*)

Последовательно, поэтапно анализируется поток символов и вычисляются 3 вида дескрипторов:

1. Токенизация и построение *основных дескрипторов*
2. Вычисление *контекстных дескрипторов*
3. Определение *макросинтаксических дескрипторов*

При этом распознаются:

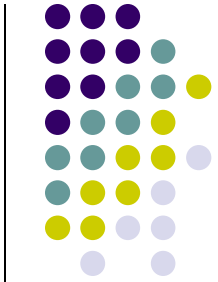
- слова и разделители
- предложения
- аббревиатуры, ФИО, даты и числа, электронные адреса, имена файлов
- тире и дефис
- устойчивые словосочетания
- абзацы, заголовки, перечисления (рубрики)

ГРАФЕМАТИКА *aot.ru*: ПРИМЕР

В гостинице на ул. Лесная в 1913 году останавливался Анатоль Франс; в 1917 г. – Джон Рид.

```
|                    0 0 BEG DOC
в                    0 1 RLE AA EXPR1 EXPR2 EXPR_NO16
гостинице            2 9 RLE aa
на                   12 2 RLE aa EXPR1 EXPR2 EXPR_NO277
ул                   15 2 RLE aa ABB1
.                    17 1 PUN SENT_END ABB2
Лесная               19 6 RLE Aa NAM?
в                    26 1 RLE aa EXPR1 EXPR2 EXPR_NO16
1913                 28 4 DC
году                 33 4 RLE aa
останавливался       38 14 RLE aa
Анатоль              53 7 RLE Aa NAM?
Франс                61 5 RLE Aa NAM?
;                    66 1 PUN
в                    68 1 RLE aa EXPR1 EXPR2 EXPR_NO16
1917                 70 4 DC
г                    75 1 RLE aa ABB1
.                    76 1 PUN
–                    78 1 PUN ABB2
Джон                 80 4 RLE Aa NAM?
Рид                  85 3 RLE Aa NAM?
.                    88 1 PUN CS? SENT_END
```

МАШИННОЕ ОБУЧЕНИЕ ДЛЯ СЕГМЕНТАЦИИ

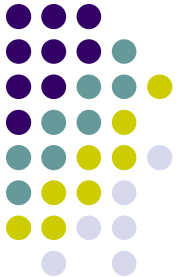


- Токенизация текста может решаться как задача классификации символов текста
- Простейший вариант: каждый символ текста необходимо отнести к одному из двух/трех классов:
  - последний символ в токене (L)
  - не последний, обычный символ в токене (R)
  - не токен (**O**)

ОН СЕЛ ВАВТОБУС – RL**O**RRL**O**LRRRRRRL

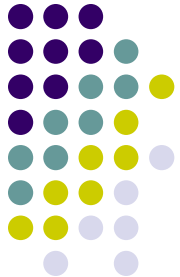
- Если одновременно – разбиение на предложения,то нужен еще один класс символов (S – конец предложения)
- Для обучения необходим *data set* с такой же разметкой символов текста, по которому собирается информация и строится вектор признаков (*features*) символов/классов (до 30-40 признаков каждого символа текста)

ПРИЗНАКИ СИМВОЛОВ ДЛЯ КЛАССИФИКАЦИИ



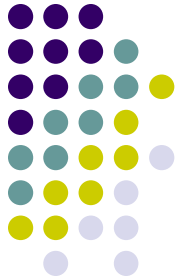
- Информация о текущем символе
  - Вид символа: кириллица/латиница/спецзнак
  - Вхождение в один из фиксированных подклассов спецзнаков (например, круглые скобки)
  - Регистр буквы и др.
- Аналогичная информация о символах левого/правого контекста (*контекстного окна*), т.к. сегментация на токены/предложения не может быть надежно решена независимой классификацией отдельных символов
- Возможен также учет информации о всем текущем токене (вхождение в словарь или др.)
- Вектора признаков символов используются в ходе машинного обучения, а для сегментации нового текста подаются на вход обученному маш. классификатору

СЕГМЕНТАЦИЯ КАК РАЗМЕТКА ПОСЛЕДОВАТЕЛЬНОСТИ



- Поскольку для сегментации на токены/предложения требуется учет контекстной информации, естественно рассматривать сегментацию как *задачу разметки последовательности* (*Sequence labeling*) символов
- Классическими методами для этой задачи являются:
  - Скрытые Марковские цепи (HMM)
  - Метод условных случайных полей (CRF), дает лучшие результаты, чем HMM
  - Нейронные сети RNN и BiRNN, но требуют большего обучающего множества
- Пример: сегментация в проекте *UDPipe*
- Задача сегментации на предложения в среднем решается с худшим качеством, чем токенизация

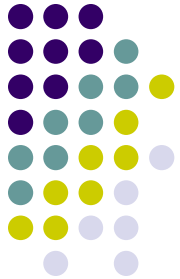
СЕГМЕНТАЦИЯ В ПРОЕКТЕ *UDPipe*



- Проект *UDPipe* (C++) :
  - последовательная обработка текста (*pipeline*): сегментация, морфологич. и синтаксич. анализ
  - мультиязыковая парадигма: для 32 языков, для которых есть данные (корпуса) в разметке *UD*
  - *UD* (*Universal Dependencies*) – универс. система лингвистических тегов из межд. проекта для ЕЯ
  - машинное обучение на нейронных сетях
- Модуль сегментации выполняет (*LSTM, GRU*)
  - Токенизацию, качество: 98.7 – 99.9%
    для чешского – 100% (но 100% и для *rule-based*)
  - Сегментацию на предложения: 71.2 – 98.7%

Три метки: конца токена, конца предложения, нет конца

МОРФОЛОГИЧЕСКАЯ СЕГМЕНТАЦИЯ



- Разбиение на слова в языках (восточных) со слитным написанием слов (*non-segmented languages*) и в европейских языках с большим числом компаундов (*сложносоставных* слов), например, для немецкого:

  Dampfschiffahrtsgesellschaft – пароходная компания

  Dampf / *schiff* / *ahrts* / *gesell* / *schaft*

  Rechtsschutzversicherungsgesellschaften – страховые компании, занимающиеся правовой защитой клиентов

- Разбиения на морфы/морфемы (или их группы) в языках с раздельным написанием слов, например, для русского:

  безвкусный – *без*:PREF/*вкус*:ROOT/*н*:SUFF/*ый*:END

  (английский: *taste*:ROOT/*less*:SUFF)

  душевность – *душ*:ROOT/*евн*:SUFF/*ост*:SUFF/*ь*:END

ОЦЕНКИ КАЧЕСТВА СЕГМЕНТАЦИИ

- Качество оценивается по количеству найденных границ токенов/предложений
- Ошибки: пропуск границы или лишние границы

- *Точность* – доля точно найденных границ среди всех найденных границ

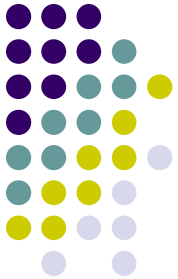
$$\text{Precision} = \frac{tp}{tp + fp}$$

- *Полнота* – доля верно найденных границ среди всех истинных границ

$$\text{Recall} = \frac{tp}{tp + fn}$$

эксперт / система	правильные (*positive* – *P*)	неправильные (*negative* – *N*)
правильные (*P*)	True P = TP	False P = FP
неправильные (*N*)	FN	TN

ОБЪЕДИНЕННАЯ ОЦЕНКА КАЧЕСТВА



- *F-мера* усредняет показатели полноты и точности

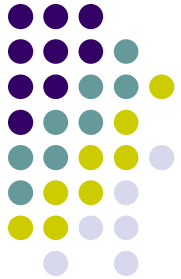
$$F = \frac{(\beta^2 + 1)\, P \times R}{\beta^2 R + P}$$

где β – коэффициент относительной важности

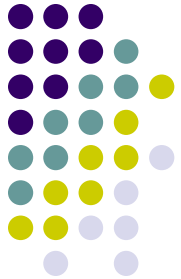
- В частности, чаще всего: $\beta = 1$, это *F1-мера* – среднее гармоничное точности и полноты

- Качество сегментации зависит от тематики, жанра, предметной области текстов (наличие разных названий, сокращений и др.)

СЕГМЕНТАЦИЯ: ВЫВОДЫ



- Токенизация и сегментация на предложения – довольно непростые задачи
- Сложность модуля токенизации зависит от дальнейшего приложения (прикладной задачи)
- От качества этого этапа зависит качество всех последующих, более сложных этапов анализа текста
- Требуется:
  - распознать и объединить символы текста так, чтобы выделенные единицы далее анализировать как целое
  - не потерять ничего существенного для дальнейшего анализа текста
- Выбор : Правила *Vs*. Машинное обучение?
  зависит от размеченных данных и разнообразия токенов
- При обоих подходах можно достичь высокого качества сегментации – до 99% и выше F-меры.



СПАСИБО ЗА ВНИМАНИЕ!

СЕГМЕНТАЦИЯ: ЭКСПЕРИМЕНТЫ

Запуск модуля токенизации и сегментации из библиотеки **nltk** (Питон) либо другой библиотеки либо **rutok** https://github.com/alesapin/rutok
и проанализировать результаты.

1. Выделить токены в предложениях:
   - *Свою карьеру летчика В.П. Чкалов начал в 1919г. слесарем-сборщиком самолетов в 4-м Канавинском авиационном парке в Нижнем Новгороде.*
   - *С 3 декабря 1931 года В.П.Чкалов испытывал самолеты-истребители 1930-х годов И-15 и И-16.*
2. Сколько предложений в данных фрагментах?
   - *Он сказал: - Поехали!*
   - *Одно из наиболее известных положений мерфологии гласит: «Всякая работа требует больше времени, чем кажется»*