Подготовка корпуса
Препроцессинг

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Первичная обработка текста

1. Препроцессинг
   - графическая нормализация
   - токенизация
   - сегментация на предложения
2. Дополнительная обработка
   - индекс
   - оффсеты
   - классификация токенов
3. Распознавание языка

Первичная обработка текста

1. Препроцессинг
   - графическая нормализация
   - токенизация
   - сегментация на предложения
2. Дополнительная обработка
   - индекс
   - оффсеты
   - классификация токенов
3. (Распознавание языков)

Графическая нормализация текста

Пример 1.

Л. Қалтаева жұмыспен қамту бағдарламасын «еңбек етіп табыс тауып, өзін лайықты, сенімді адам сезінгісі келетін» мүгедектігі бар адамдар үшін «жақсы трамплин» болып табылады деп санайды екен.
– 2013 жылғы 1 қаңтардағы жағдай бойынша Қазақстанда барлығы 609 мыңнан астам мүгедек бар, – деді Денсаулық сақтау министрі Салидат Қайырбекова.
18 жасқа дейінгі мүгедек балалар саны – 65 мың 844, оның ішінде 57 мың 627-сі – 16 жасқа дейінгі балалар.

Прмер 2.

Пример: niceweather =&gt nice weather
 Но, для workingrass сегментация опять же не будет найдена. Первое слово, которое заматчит наш алгоритм будет working, а не work и которое также поглотит первую букву в слове grass.

Может нужно скомбинировать каким-то образом оба алгоритма? Но, как тогда быть со строкой niceweatherwhenworkingrass? В общем пришли к брутфорсу.

Графическая нормализация текста

Пример 3.

Таскаю в Воспитательный
Своих незаконнорожденных детей...

<"ВИД ИМЕНИЯ ГУРЗУФ...">

{ Рис. 1 }

АТТЕСТАТ

Рукой А. Янова:

Сего февраля 24 дня осмотрена мною подушка черного атласа размером приблизительно 1/2 аршина в квадрате. Подушка с вышивкой по наложенным кретоновым цветам, среди коих помещается овал, также кретоновый, вырезан в овале сюжет буколического содержания в две фигуры (не считая собаки*): пастушка, разговаривающая с пастушком, на дальнопланной части кусты и деревья.

Все сие в общем кретоне вышивка шелком, черный фон атласа, красный шнур по швам подобраны с отменным вкусом и тщательностию. Особенно удачно вышли цветы. Вообще все хорошо, о чем свидетельствую своим подписом. Да! вообще все хорошо!!..... Да, хорошо...

_____

* или, может, козочки

Графическая нормализация текста

Пример 4.

risingvoices @risingvoices
Glad that our friends -)) (and partners for events like Tweet #MotherLanguage) from @IndigenousTweet & @livingtongues will be at #OxfordGL
Retweeted by Indigenous Tweets

Предварительная обработка текста

1. Удаление нетекстовых элементов (остатки HTML и других служебных «не текстовых» символов)
2. Исправление стандартных ошибок распознавания: кириллица vs. латиница
3. Стандартизация символов: тире, кавычки, пробелы
4. Артефакты конвертации в другой формат
5. Выделение и оформление нестандартных (нелексических) элементов, например:
    - элементов форматирования – жирность, курсивность, подчёркивание;
    - структурных элементов текста – заголовков, абзацев, примечаний;
    - различных элементов текста, не являющихся словами (числа, даты в цифровых форматах, буквенно-цифровые комплексы, и т.п.);
    - имен (имя, отчество), написанных инициалами;
    - иностранных лексем, записанных латиницей и т.д.
6. Сборка (например, слов, написанных в разрядку)

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Первичная обработка текста

1. Распознавание языка
2. Препроцессинг
   - графическая нормализация
   - токенизация
   - сегментация на предложения
3. Дополнительная обработка
   - индекс
   - оффсеты
   - классификация токенов
4. (Исправление ошибок)

ГРАФЕМАТИЧЕСКИЙ АНАЛИЗ

- Уровень обработки: символы текста
- Графема – единица текста (письменного), неделимый знак (буквы, знаки препинания и др.)
- Цель – выделение и классификация основных единиц текста: слов, предложений, абзацев

(близкое название – сегментация (англ. Segmentation) , т.е. разбиение текста на части)

Сегментация - более широкое понятие

Сегментация текста – процесс разбиения потока символов на отдельные единицы обработки (лингвистические объекты: слова, предложения, числа, пунктуационные знаки)

Токенизация

- ***Токены*** – единицы обработки соответствующие словам («псевдословам»)
- Token – кусочек, обычно цепочка знаков от пробела до пробела (компиляция ЯП: лексема)
- Цель – выделение минимальных лингвистически значимых элементов текста (токенов)
- Виды токенов:
  - Слова ЕЯ
  - Знаки препинания
  - Обозначения денежных единиц
  - Числа
  - Буквенно-цифровые комплексы
  - Даты (множество форматов)
  - Номера телефонов
  - IP -адреса и имена файлов

Токенизация

- Наивная токенизация – все разбиваем по пробелам
- Всегда ли пробелы имеют одну и ту же функцию:
  - ✓Both “Los Angeles” vs. “rock 'n' roll”
  - ✓100 000
- !!!!
- Обычно считается, что токенизация – очень простая, неинтересная и не очень значимая процедура
- НО: от качества токенизации может очень сильно зависеть качество выскоуровневых задач
- Ср. выделение именованных сущностей: U.S., ex-president, don’t

Токенизация

- Основания для уточнения правил токенизации:
  a) Языковая реальность
  b) Конечная NLP задача
  c) Архитектура обработки
- а) критерий слова -))),
  - ср. русск. *Петя-то пришел*

    Пойди-ка принеси
  - ср. *Time-frame*, *timeframe*, *time frame*
  - Интерпретация токена может зависеть от контекста:
  - 100 г. / г. Ростов - dr – doctor / drive

Токенизация. Компоненты

- регистр (обычно все приводим к одному регистру)
- знаки препинания и служебные символы
- обработка точки
- обработка дефиса
- обработка апострофа
- обработка буквенно-числовых комплексов
- обработка дат
- типизация токенов
- офсеты

Токенизация

- Аббревиатуры
- Готовые списки и словари акронимов (точка элемент сокращения)
- Точка отдельный токен
- Омонимия:
- `in' - `inches; `no' - `number, `bus' - `business; `sun' - `Sunday;
- ПО

Токенизация: аббревиатуры

- O.U.
- M.D.
- N.B.
- P.O.
- U.K.
- U.S.
- U.S.A.
- P.S.

- mr.
- mrs.
- .com
- dr.
- .sh
- .java
- st.
- .net

https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en

Токенизация

Обработка дефисов

- self-assessment,
- F-15,
- forty-two
- Los Angeles-based.
- !!! Зависит от задачи
- Частеречные разметчики (part-of-speech taggers)
- NER – скорее разные слова: *`Moscow-based'*

https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en

Токенизация

- Типы дефисов
  - переносы слов
  - лексические
    - Элементы словаря *Тянь-шань*
    - Стандартное написание некоторых префиксов co-, pre-, meta-, multi-, etc.
    - Модели: *forty-seven*
    - Окказионализмы: *end-of-line*
  - “Синтаксические”
    - `ed'-отглагольные прилагательные *case-based*, *computer-linked*, *hand-delivered*
    - *three-to-five-year*
    - *the New York-based co-operative was fine- tuning forty-two K-9-like models.*

Токенизация

Token	Type
New York-based	**Sentential**
co-operative	**Lexical**
fine-tuning	**End-of-Line**, but could also be considered a **Lexical** hyphen based on the author's stylistic preferences.
Forty-two	**Lexical**
K-9-like	**Lexical** and **Sentential**

Токенизация: обработка буквенно-цифровых комплексов

- Examples:Email addresses
- URLs
- Complex enumeration of items
- Telephone Numbers
- Dates
- Time
- Measures
- Vehicle Licence Numbers
- Paper and book citations
- etc

Телефонные номера

- 123-456-7890
- (123)-456-7890
- 123.456.7890
- (123) 456-7890
- etc

https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en

Токенизация

- Отдельные модули для распознавания и нормализации таких классов объектов (телефонных номеров, интернет адресов, дат)
- Date/Time Formats:
  8th-Feb
- 8-Feb-2013
- 02/08/13
- February 8th, 2013
- Feb 8th
- и т.п.

https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en

Tokenization Example

"I said, 'what're you? Crazy?'" said Sandowsky. "I can't afford to do that."

	Naïve по пробе-лам	Apache Open NLP (en-token.bin)	Stan-ford 2.0.3	Custom	Гипотет. (3С)
1		“	“	“	“
2	“I	i	i	i	i
3	said,	said	said	said	said
4		,	,	,	,
5	'what're	'what	`	'	'
6			what	what're	what
7		're	're		are
8	you?	You	you	You	you
9		?	?	?	?
10	crazy?'"	crazy	crazy	crazy	crazy
11		?	?	?	?
12		'	'	'	'

	Naïve по пробелам	Apache Open NLP	Stanford 2.0.3	Custom	Гипотетич. (3С)
13	Said	said	Said	said	said
14	sandowsky.	sandowsky	sandowsky	sandowsky	sandowsky
15		.	.	.	.
16		'	'	'	“
17	'i	i	i	i	i
18	can't	ca	ca	can't	can
19		n't	n't		not
20	afford	afford	afford	afford	afford
21	to	to	to	to	to
22	do	do	do	do	do
23	that.'	that	that	that	that
24		.	.	.	.
25		'	'	'	'

https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en

Токенизация: пунктуация внутри токенов

- u.s.a.,
- Ph.D.,
- AT&T,
- ma'am,
- cap'n,
- 01/02/06

- stanford.edu
- 7.1
- "$2,023.74"

https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en

Токенизация: задачи и архитектура системы

- Named Entity Extraction
- <node span="Rational Software Architect for WebSphere" type="**NP**">
  <node span="Rational Software Architect for WebSphere" type="**NNP**"/>
  </node>

Токенизация. «Клитики». Апострофы

The abbreviated forms of *be*:
1.’m in I’m
2.’re in you’re
3.’s in she’s
The abbreviated forms of *auxiliary verbs*:
1.’ll in they’ll
2.’ve in they’ve
3.’d in you’d
Note that clitics in English are ambiguous. The word "she's" can mean "she has" or "she is".
A tokenizer can also be used to expand clitic contractions that are marked by apostrophes, for example:
what're => what are
we're => we are https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en

Токенизация: этапы

- Шаг 1. Разбиение по пробелам, очистка от кавычек, скобок и др. служебных символов
- Шаг 2.
- Обработка сокращений и точек в сокращениях (в некоторых приложениях точка сохраняется как значимый символ аббревиатуры)
- Шаг 3.
- Дефисы
- Шаг 4.
- Обработка бувенно-числовых и числовых комплексов
- Шаг 5.
- Обработка дат

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Токенизация

- Языки с пробелами
  - Специальные случаи для обработки
- Беспробельные языки

Токенизация

- #habratopic => habra topic
- **geschwindigkeitsbegrenzung** —*ограничение скорости*
- **城市人的心爱宠物** — *любимое домашнее животное городских жителей*

Токенизация

Алгоритм 1. Minimum Matching

- *niceday*
- *niceweather*

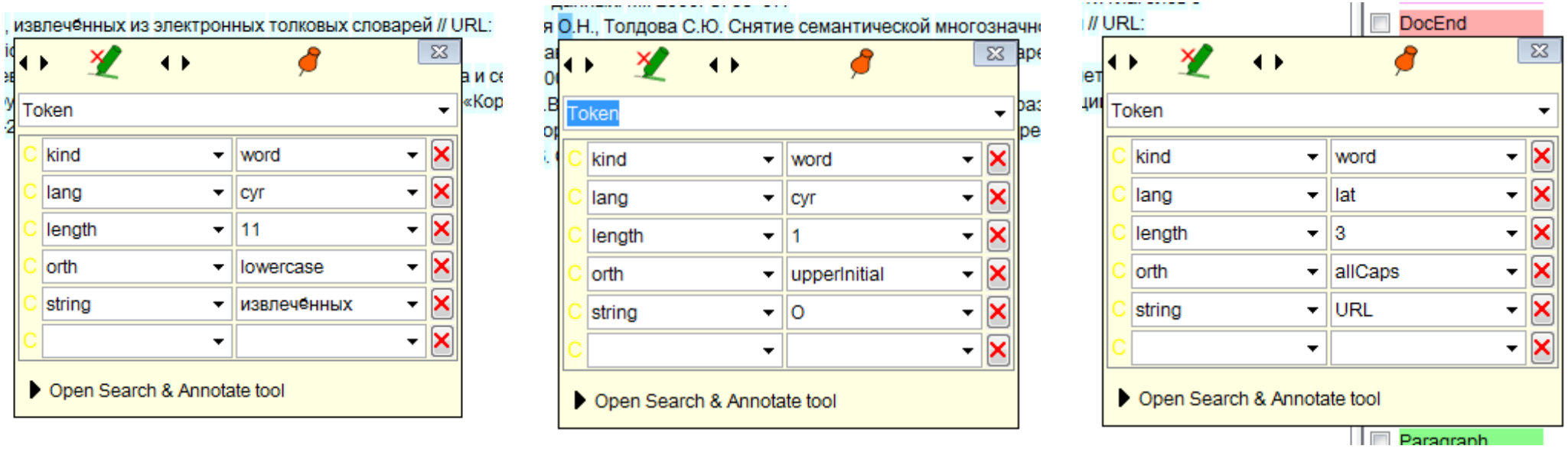
Алгоритм 2. Maximum Matching или Greedy

- *niceweather*
- *Workingrass*
- Алгоритм 3. Все варианты разбиения по словарю
- *expertsexchange* => (*expert sex change*, *experts exchange*)
- *dwarfstealorcore* -
  - «дворф крадет или ядро»?
  - «дворф крадет руду орков»

ГРАФЕМАТИКА aot.ru : ДЕСКРИПТОРЫ

- Основные графематические дескрипторы (19):
- OLE – русская лексема (последовательность букв кириллицы)
- OILE – иностранная лексема;
- OPun – знак препинания (. , : ; -);
- OOpn и OCls – открывающая и закрывающая скобки; Контекстные дескрипторы (19)
- OSent1 и OSent2 – признаки начала и конца предложения;
- OBulet – признак начала пункта перечисления;
- OPar – признак начала абзаца;
- OFIO1 и OFIO2 – признак начала и конца ФИО;
- Дескрипторы макросинтаксического анализа (5):
- CS_Heading – признак конца заголовка;
- CS_Parent – конца раздела, заканчивающегося знаком :

Типизация токенов. Пример
Описание токена в среде OntosMiner



Токенизация: адреса токенов

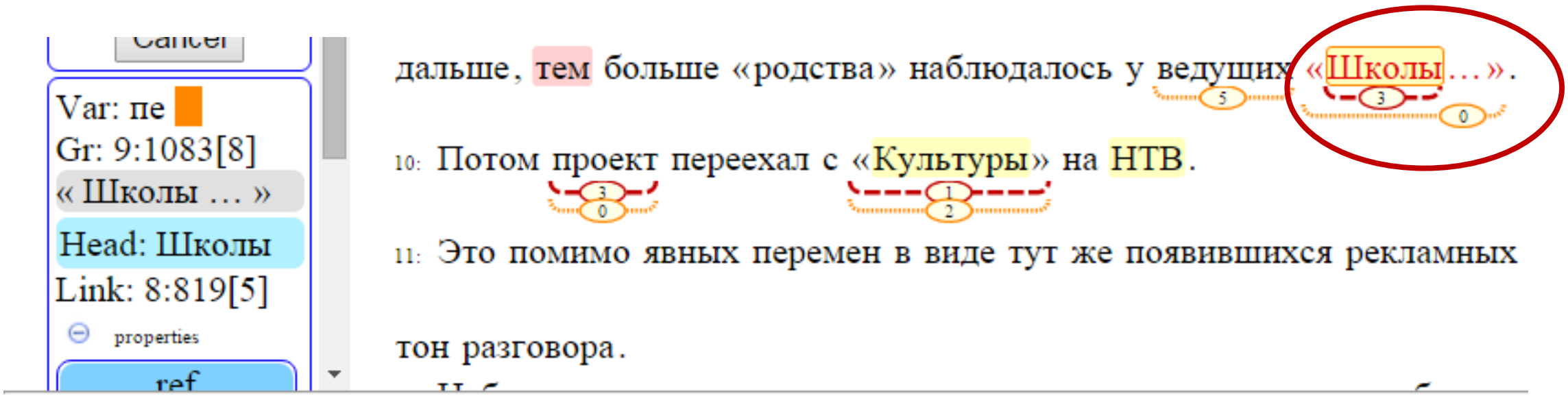
Sentence № 12835 ☐ Arrows ☑ Tables

Милиционер поднял голову , увидел пуле

SyntAutom

id	token	type	head
1	**Милиционер** ← *поднял*	subj:nom	2
2	**поднял**	root	
3	**голову** ← *поднял*	obj:acc	2
4	, ← *поднял*	conj	2
5	**увидел** ← *поднял*	coord	2
6	**пулевое** ← *отверстие*	adj	7

Токенизация: адреса токенов



Матвей Курзуков

Ref	Group	ref	str	type
8:788[17]	« Школа злословия »	def	noun	coref
8:819[5]	собой	def	refl	coref
9:1084[5]	Школы	def	noun	coref
10:1099[6]	проект	def	noun	coref
12:1335[9]	программы	def	noun	coref
13:1441[17]	« Школы злословия »	def	noun	coref
17:2064[17]	« Школа злословия »	def	noun	coref

Саша Печеный

Ref	Group	ref	str	type
3:13[17]	« Школа злословия »	def	noun	coref
4:96[9]	программы	def	noun	coref
5:187[8]	« Школа … »	def	noun	coref
7:734[9]	программы	def	noun	coref
8:788[17]	« Школа злословия »	def	noun	coref
8:819[5]	собой	def	refl	coref
9:1083[8]	« Школы … »	def	noun	coref

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Токенизация

- Шаг 1. Разбиение по пробелам, очистка от кавычек, скобок и др. служебных символов
- Шаг 2. Обработка сокращений и точек в сокращениях (в некоторых приложениях точка сохраняется как значимый символ аббревиатуры)
- Шаг 3. Дефисы
- Шаг 4. Обработка бувенно-числовых и числовых комплексов
- Шаг 5. Обработка дат
- Типицация токенов
- Адреса токенов

Форматы представления текстовых корпусов

```
<document>
  <docID>040404-27793</docID>
  <docURL> URL документа в Веб в base
64</docURL>
  <subject encoding="base64"> тема новости
в base64 </subject>
  <agency>название новостного агенства в
base64</agency>
  <timestamp>
     <date>20040402</date>
     <daytime>50493</daytime>
  </timestamp>
  <content encoding="base64">
      содержимое в base64
  </content>
</document>
```

Форматы представления текстовых корпусов

```
<resultlist>

 <entitylist srcRef="1-27793">
   <entity class="person" offset="432" length="18" id="1">Эдуарда Шеварднадзе</entity>
   <entity class="other" offset="198" length="13" id="2">революция роз</entity>
 </entitylist>
 <entitylist srcRef="2-45913">
   <entity class="organization" offset="1" length="19" id="1">большая восьмерка</entity>
   <entity class="place-name" offset="56" length="9" id="2">Шотландия</entity>
   <entity class="organization" offset="320" length="3">G8</entity>
   <entity class="organization" offset="548" id="1">большой восьмерке</entity>
 </entitylist>

</resultlist>
```

Форматы представления текстовых корпусов. CoNLL

ID FORM LEMMA PLEMMA POS PPOS FEAT PFEAT HEAD PHEAD DEPREL PDEPREL

ID (index in sentence, starting at 1)
FORM (word form itself)
LEMMA (word's lemma or stem)
POS (part of speech)
FEAT (list of morphological features separated by |)
HEAD (index of syntactic parent, 0 for ROOT)
DEPREL (syntactic relationship between HEAD and this word)

http://universaldependencies.org/docs/format.html

Форматы представления текстовых корпусов. CoNLL

```
1   Då    då    ADV    AB                    _
2   var   vara  VERB   VB.PRET.ACT           Tense=Past|Voice=Act
3   han   han   PRON   PN.UTR.SIN.DEF.NOM
Case=Nom|Definite=Def|Gender=Com|Number=Sing
4   elva  elva  NUM    RG.NOM               Case=Nom|NumType=Card
5   år    år    NOUN   NN.NEU.PLU.IND.NOM
Case=Nom|Definite=Ind|Gender=Neut|Number=Plur
6   .     .     PUNCT  DL.MAD                _
```

http://universaldependencies.org/docs/format.html

Особые случаи
Сегментация в текстах социальных сетей

- @SentimentSymp: can't wait for the Nov 9 #Sentiment talks! YAAAAAAY!!! >:-D http://sentimentsymposium.com/.

http://sentiment.christopherpotts.net/index.html

Сегментация текстов социальных сетей

- http://sentimhttp://sentiment.christopherpotts.net/tokenizing/ent.christopherpotts.net/code-data/happyfuntokenizing.py

Особые случаи
Сегментация в текстах социальных сетей

- Все HTML и XML теги были распознаны и собраны.
- HTML коды отдельных символов - < < переведены в символы Unicode.
- @SentimentSymp: can't wait for the Nov 9 #Sentiment talks! YAAAAAAY!!! >:-D http://sentimentsymposium.com/.

http://sentiment.christopherpotts.net/index.html

Сегментация текстов социальных сетей
Whitespace tokenizer

Простой split: (а) приведет все к нижнему регистру; (б) разобьет текст по пробелам, (в) табуляции или новой строке:

- @sentimentsymp:
- can't
- Wait
- For
- The
- Nov
- 9
- #sentiment
- talks!
- yaaaaaay!!!
- >:-d
- http://sentimentsymposium.com/.

http://sentiment.christopherpotts.net/index.html

Сегментация текстов социальных сетей

Наблюдения:

- имя пользователя никак не разбивается
- The URL не разбивается, но ссылка не будет работать, так как на конце осталась точка
- эмотикон не разбился, но поменялся
- Некоторые токены не соответствуют интересующим нас языковым единицам: “talks!”
- Имя пользователя не будет идентифицировано правильно @sentimentsymp: vs. @sentimentsymp

Сегментация текстов социальных сетей
Стандарт Penn Treebank

- **Treebank-style**

@	!
SentimentSymp	YAAAAAAY
:	!
Ca	!
n't	!
Wait	>
For	;
The	:
Nov	-D
9	http
#	:
Sentiment talks	//sentimentsymposium.com/

http://sentiment.christopherpotts.net/index.html

Сегментация текстов социальных сетей
Penn Treebank

- «Неудобства» стандарта Penn Treebank для определения тональности:
  - Can’t – может быть хорошим «признаком» оценки
  - Все токены с пунктуационными знаками внутри разбиты на части: URL, специальные символы Twitter, телефоны, даты, адреса ... , эмотиконы
  - такой стандарт удобен для совмещения работы разных модулей обработки, но не очень удобен, если мы хотим анализировать тексты типа твитов (слишком много дополнительных правил придется вводить на других этапах обработки)

http://sentiment.christopherpotts.net/index.html

Сегментация текстов социальных сетей
Sentiment-aware tokenizer

- **Эмотиконы:**
- [<>]? # optional hat/brow
- [:;=8] # eyes
- [\-o*\']? # optional nose [\)\]\(\[dDpP/\:\}\{@\|\\] # mouth
- | #### reverse orientation [\)\]\(\[dDpP/\:\}\{@\|\\] # mouth
- [\-o*\']? # optional nose
- [:;=8] # eyes
- [<>]? # optional hat/brow
- Эмотиконы – наиболее надежные признаки позитивной/негативной оценки в текстах социальных сетей.
- Эти регулярные выражения «ловят» 96% эмотиконов в Твиттере (всего 36% разных видов эмотиконов, но остальные слишком редкие)

http://sentiment.christopherpotts.net/index.html

Сегментация текстов социальных сетей
Twitter mark-up

Полезно также распознавать ники пользователей и темы твитов

- Usernames:
- @+[\w_]+
- Hashtags (topics):
- \#+[\w_]+[\w\'_\-]*[\w_]+

http://sentiment.christopherpotts.net/index.html

Сегментация текстов социальных сетей
Informative HTML

- Для определения тональности (sentiment analysis) важно форматирование
- HTML теги для выделения strong, b, em, и i – открывающий или закрывающий тег можно оставлять как токен
- Можно перевести их в верхний регистр, чтобы дальше обрабатывать также, как фрагменты, набранные в верхнем регистре для передачи эмоций
- really bad idea ->
- REALLY BAD IDEA
- Иногда имеет смысл учитывать специальную разметку:
- <span class="rating>2 of 5
- <span class="title>The Good, the Bad, and the Ugly

http://sentiment.christopherpotts.net/index.html

Сегментация текстов социальных сетей
Mask curses

- На некоторых сайтах ненормативная лексика «маскируется»: ****, s***t или первая буква+многоточие
- некоторые передают ругательства с помощью комбинации небуквенных символов: $#!@)

http://sentiment.christopherpotts.net/index.html

Сегментация текстов социальных сетей
Additional puctuation

- На этапе токенизации оставляем знаки препинания;
- Вначале пытаемся идентифицировать «внутрисловные» пунктуационные знаки, тогда остальные можно уже обрабатывать как разделители.
- Некоторые соображения:
  - эмотиконы, важные символы разметки Твиттера и HTML, «маски» для ненормативной лексики;
  - последовательности из букв, цифр, апострофов, дефисов и нижнего подчеркивания – слова;
  - Последовательности состоящие только из цифр, запятых и точек – токены; можно еще в токен включать знаки валюты и проценты;
  - Последовательности из более одной точки можно заменить на многоточие ...

Сегментация текстов социальных сетей
Additional punctuation

- Остальное может рассматриваться как отдельные токены:
- Вопросительные и восклицательные знаки, знаки доллара без цифр (имеет смысле оставлять отдельные токены в последовательности восклицательных/вопросительных знаков, как в !!!)
- Их можно отфильтровать на более поздних этапах

Сегментация текстов социальных сетей
Capitalization

- Слова в верхнем регистре стоит оставлять в верхнем регистре

http://sentiment.christopherpotts.net/index.html

Сегментация текстов социальных сетей
Lengthening

- Редупликация символов, как в *оооочень* значима для определения тональности
- Стоит превращать любую цепочку в последовательность не более трех символов

Сегментация текстов социальных сетей
Multi-word expressions

- Even in English, whitespace is only a rough approximation of token-hood in the relevant sense:
- Named entities
- Phone numbers
- Dates
- Idioms like out of this world
- Multi-word expressions like absolutely amazing
- The basic strategy is to tokenize these greedily, first, and then proceed to substrings, so that, for example, November 9 is treated as a single token, whereas an isolated occurrence of November is tokenized on its own.
- If one starts including n-grams like really good as tokens, it is hard to know where to stop. For large enough collections, bigram or even trigram features might be included (in which case you can tokenize without paying attention to these phrases). For smaller collections, some of the mark-up strategies discussed later on can approximate such information (and often prove more powerful).

http://sentiment.christopherpotts.net/index.html

Сегментация текстов социальных сетей
Putting the pieces together

- The tokenizer that I use for sentiment seeks to isolate as much sentiment information as possible, and it also identifies and normalizes dates, URLs, phone numbers, and various kinds of digital address. These steps help to keep the vocabulary as small as possible, and they provide chances to identify sentiment in areas that would be overlooked by simpler tokenization strategies (July 4th, September 11).

http://sentiment.christopherpotts.net/index.html

Сегментация текстов социальных сетей
Putting the pieces together

- @sentimentsymp
- :
- can't
- Wait
- For
- The
- Nov_09
- #sentiment
- Talks
- :
- YAAAAY
- !

The social-media mark-up is all left intact, the date is normalized, and YAAAAAAY has been put into a canonical elongated form. http://sentiment.christopherpotts.net/index.html

СЕГМЕНТАЦИЯ НА ПРЕДЛОЖЕНИЯ

- Используются:
- Маркеры конца предложения – точка, вопросительный и восклицательный знаки но! – неоднозначны (сокращения слов, инициалы, сокращения в конце предложения: Dr.White)
- Маркер начала предложения – заглавная буква, также неоднозначен (цитаты и др.) ..пр. Мишка прыгал по полу… сказал он: «Я вижу лес..
- Цитаты (прямая речь), оформляются в разных языках кавычками /апострофами (в англ. языке - также для обозначения притяж. падежа и сокращений : Ann’s , it’s)
- Требуется анализ локального контекста маркеров
- Точность сегментации зависит от тематики и жанра текстов, количества сокращений, имен собственных
- Применение машинного обучния (статистика по корпусам)

СРЕДСТВА ГРАФЕМ. АНАЛИЗА

- Таким образом, при сегментации нужны компоненты:
  - Словарь сокращений
  - Словарики графических знаков
  - Словарь устойчивых оборотов (обычно более 500)
  - Эвристические правила анализа контекстов, более чем один просмотр текста
  - Языково-зависимые компоненты !
  - (также зависит от тематики текстов, причины - различная роль знаков препинания и др.)
  - Достигаемая точность – до 99, полнота – 60-80 %
  - Восточные языки (non-segmented languages) :
  - слитное написание слов ⇒ применяются:
  - статистические методы сегментации, морфословари, грамматические правила (японский) ,
  - также европ. языки с большим сложносоставных слов,
  - например, немецкий: Worterbuch

ТЕХНОЛОГИИ РЕАЛИЗАЦИИ ГРАФЕМАТИЧЕСКОГО АНАЛИЗА

- Формальный аппарат на базе теории формальных языков и грамматик
- Простейший графематический анализ – анализ регулярных языков (Тип 3 по Хомскому)
- Более сложный граф. анализ – учет локального контекста, словари
- Средства описания регулярных языков
  - Регулярные выражения
  - Регулярные (автоматные) грамматики
  - Конечные автоматы

РЕГУЛЯРНЫЕ ВЫРАЖЕНИЯ

- Регул. выражение описывает структуру цепочки языка. Синтаксис регулярных выражений α * – повторение цепочки α нуль и более раз, например:
- b* – последовательность из произв. количества букв b (в том числе – нулевого, т.е. пустая цепочка);
- α + – повторение цепочки α один и более раз, например:
- b+ – непустая последовательность из букв b;
- α ? – опциональная цепочка α (входит 0 или 1 раз), например:
- ab? – цепочка a или ab;
- ( ) – группировка, например:
- (ab)+ – последовательность из повторений строки ab;
- α | β – альтернатива, например:
- b|a+ – буква b или последовательность букв a;
- Пример регулярного выражения для двоичного числа с точкой:
- (0|1)+ . (0|1)+

Регулярные выражения

- Примеры текстов см. файл

- Упражнение:
- напишите 5 разных шаблонов для извлечения дат
- Напишите шаблон для извлечения токенов, которые в конце имеют окончания одного из склонений творительного падежа существительных (нужно, чтобы шаблон перечислял все возможные окончания творительного падежа)

Сегментация на предложения: признаки для автоматической классификации

- Символ-разделитель
- Что стоит слева
  - Знак препинания
  - цифры
  - одна буква в верхнем регистре
  - одна буква в нижнем регистре
  - сокращение из списка
  - …
- Что стоит справа
  - большая буква – граница предложения
  - НО: г. Москва
  - …

Сегментация на предложения: признаки для автоматической классификации

- Символ-разделитель
- «псевдослово» справа
- «псевдослово» слева
- количество символов в слове
- …

- Чтение
- О.Урюпина. Автоматическое разбиение текста на предложения для русского языка http://www.dialog-21.ru/digests/dialog2008/materials/html/83.htm

ЛИТЕРАТУРА

- Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.
- Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008.
- Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.
- Oxford Handbook on Computational Linguistics. R. Mitkov (Ed.). Oxford University Press, 2005 , p. 201-218.
- http://sentiment.christopherpotts.net/tokenizing/