ТЕХНОСФЕРА

Лекция 1
Лингвистика в поиске
Основы обработки текстов

Владимир Гулин

13 сентября 2016 г.

Что мы знаем на текущий момент?

- Как качать интернет
- Что класть в индекс
- Фильтрация нежелательного контента (антиспам/антипорн)
- Исправление опечаток и саджесты
- Модель булева поиска
- Как формировать сниппеты
- Что еще?

Структура курса

1. **Лингвистика в поиске. Основы обработки текстов.**
2. **Коллокации, n-граммы, скрытые марковские цепи** ^{H}
3. **Текстовое ранжирование. Базовые модели** ^{H}
4. **Оценка качества поиска.** ^{H}
5. **Оценка качества поиска. Интерливинг.**
6. **Ссылочное ранжирование.** ^{H}
7. **Поведенческое ранжирование.**
8. **Машинное обучение в ранжировании 1** ^{H}
9. **Машинное обучение в ранжировании 2**
10. **Текстовое ранжирование. Хитрые модели**
11. **Разбор ключевых ДЗ**
12. **Мультимедия поиск**
13. **Технологии поиска по лицам** ^{H}
14. **Контентные рекомендательные системы**

Лингвистика в поиске

Этапы ранжирования

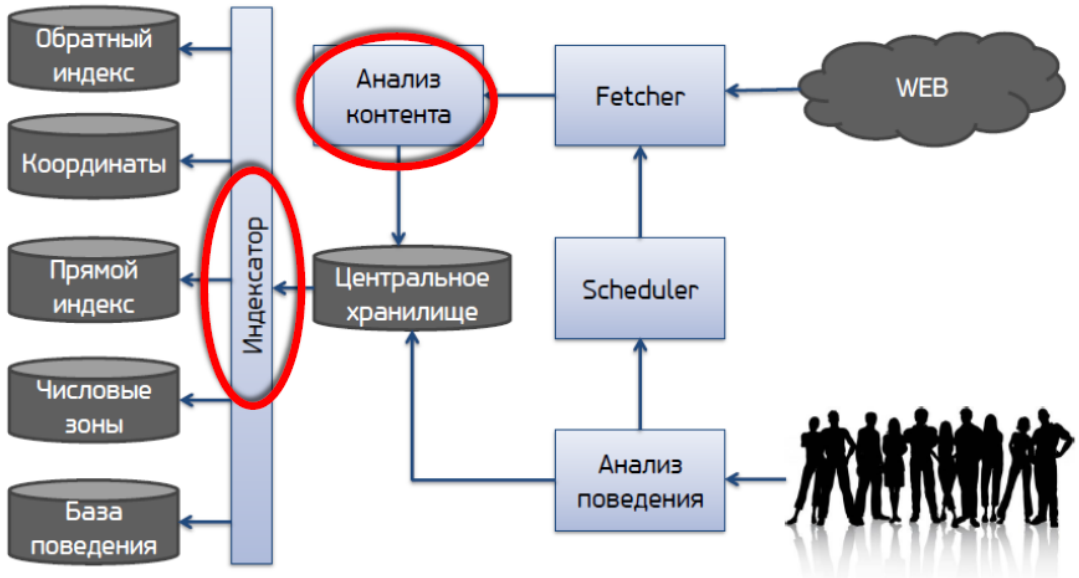
Терминология

- *Токен* - экземпляр последовательности символов в документе, объединенных в семантическую единицу для обработки
- *Термин* - “нормализованный” токен (регистр, морфология, исправление ошибок и т.п.)

Нормализация

- Необходимо “нормализовывать” термины как в индексируем тексте, так и в запросе
- Например: Желательно считать одинаковыми термины U.S.A. и USA
- Обычно термины объединяются в классы эквивалентности
- Можно постпуть наоборот, расширять:
  - window $\rightarrow$ window, windows
  - windows $\rightarrow$ Windows, windows
  - Windows (нет расширения)
- Такой подход более гибкий, но более ресурсоемкий

Обработка документа



Документы

Делаем неявное предположение:

- Мы знаем, что такое документ
- Каждый документ доступен для автоматического разбора

Вопрос:

- Какие тут есть проблемы?

Лингвистика при обработке документов

- Определение формата документа (pdf, word, html и т.д.)
- Определение кодировки документа
- Определение языка документа
- Токенизация и сегментация
- Нормализация и лемматизация
- Выделение объектов и зон
- Вычисление текстовых факторов

Нормализация

Нормализация зависит от языка документа

- PETER WILL NICHT MIT. $\rightarrow$ MIT = mit
- He got his PhD from MIT. $\rightarrow$ MIT $\neq$ mit

Нормализация

Ударения и диакритика

- résumé vs. resume
- Умуляуты: Universität vs. Universitaet (заменяем на специальную последовательность «ae» или даже «æ»)

- Самый важный вопрос: как пользователи предпочитают писать запросы с этими словами?

Нормализация

Классы эквивалентности

- Soundex
  - фонетическая эквивалентность, Muller = Mueller
- Тезаурус
  - семантическая эквивалентность, car = automobile

Нормализация

Регистр

- Понизить регистр всех букв.
- Возможны исключения, например, для капитализированных слов внутри предложения.
  - MIT и mit
  - Fed и fed
  - КОТ и кот (Калининградская областная таможня)
- NB: немецкий → существительные с большой буквы
- Часто лучше понижать всё, потому что пользователи не заботятся о капитализации в запросах.

Токенизация

Проблемы токенизации

- Hewlett-Packard
- State-of-the-art
- co-education
- San Francisco
- York University vs. New York University

Проблемы токенизации

Числа

- 3/20/91
- 20/3/91
- Mar 20, 1991
- B-52
- 100.2.86.144
- (800) 234-2333
- 800.234.2333

Обработка запроса

Обработка запроса

Запросы задают не по-русски

- Распознавание языка
- Исправление опечаток
- Токенизация
- Нормализация и лемматизация
- Кореференция (расширение запроса)
- Переформулировки запросов
- Сегментация запроса
- Извлечение объектов

Проблемы токенизации

В китайском нет пробелов

李克强说，当前国际和地区形势发生复杂深刻变化，中越都处于发展的关键阶段，双方要从战略高度和长远角度出发，在发展中越关系十六字方针和“四好”精神指 引下，坚定不移推进中越友好。中方愿同越方保持高层战略沟通，加强治国理政经验交流，坚持经济优先、民生优先，深化务实合作，推动中越全面战略合作伙伴关 系迈上新台阶。

(c) news.xinhuanet.com

Другие случаи отсутствия пробелов

Компаунды в датском, немецком, шведском

- Computerlinguistik $\rightarrow$ Computer + Linguistik
- Lebensversicherungsgesellschaftsangestellter $\rightarrow$ leben + versicherung + gesellschaft + angestellter
- Kallistuksenvaimennusjarjestelma - система, предотвращающая крен (в погрузчиках).

Льезоны в романских языках

- em os $\rightarrow$ nos
- por a $\rightarrow$ pela

Эскимосы: tusaatsiarunnanngittualuujunga (Я не очень хорошо слышу)

Японский

- 4 разных алфавита.
- ローマ字 Romaji
- 平仮名 Hiragana
- 片仮名 Katakana
- 漢字 Kanji

Запрос может быть задан в любом из них

Named Entity Recognition

Извлечение объектов (группа слов/токенов в запросе, которые означают одно понятие)

- ФИО
- Телефоны
- Адреса
- Даты
- Названия песен, фильмов, книг и т.д.

Кодировки

- ASCII (ISO 646) - 7-битный стандарт
- ISO 8859
  - 8859-1 (ISO Latin-1)
  - ISO 8859-5
- Русские кодировки
  - CP1251 (windows)
  - 866 (dos)
  - KOI8-R (unix)
- Unicode
  - UTF-8
  - UTF-16
  - UTF-32

ASCII

American Standard Code for Information Interchange (1967)

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

Кодировки

koi8-r

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	─ 128	│ 129	┌ 130	┐ 131	└ 132	┘ 133	├ 134	┤ 135	┬ 136	┴ 137	┼ 138	▀ 139	▄ 140	█ 141	▌ 142	▐ 143
9	░ 144	▒ 145	▓ 146	⌠ 147	■ 148	∙ 149	√ 150	≈ 151	≤ 152	≥ 153	nbsp 154	⌡ 155	° 156	² 157	· 158	÷ 159
A	═ 160	║ 161	╒ 162	ё 163	╓ 164	╔ 165	╕ 166	╖ 167	╗ 168	╘ 169	╙ 170	╚ 171	╛ 172	╜ 173	╝ 174	╞ 175
B	╟ 176	╠ 177	╡ 178	Ё 179	╢ 180	╣ 181	╤ 182	╥ 183	╦ 184	╧ 185	╨ 186	╩ 187	╪ 188	╫ 189	╬ 190	© 191
C	ю 192	а 193	б 194	ц 195	д 196	е 197	ф 198	г 199	х 200	и 201	й 202	к 203	л 204	м 205	н 206	о 207
D	п 208	я 209	р 210	с 211	т 212	у 213	ж 214	в 215	ь 216	ы 217	з 218	ш 219	э 220	щ 221	ч 222	ъ 223
E	Ю 224	А 225	Б 226	Ц 227	Д 228	Е 229	Ф 230	Г 231	Х 232	И 233	Й 234	К 235	Л 236	М 237	Н 238	О 239
F	П 240	Я 241	Р 242	С 243	Т 244	У 245	Ж 246	В 247	Ь 248	Ы 249	З 250	Ш 251	Э 252	Щ 253	Ч 254	Ъ 255

Кодировки

cp866

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	А 128	Б 129	В 130	Г 131	Д 132	Е 133	Ж 134	З 135	И 136	Й 137	К 138	Л 139	М 140	Н 141	О 142	П 143
9	Р 144	С 145	Т 146	У 147	Ф 148	Х 149	Ц 150	Ч 151	Ш 152	Щ 153	Ъ 154	Ы 155	Ь 156	Э 157	Ю 158	Я 159
A	а 160	б 161	в 162	г 163	д 164	е 165	ж 166	з 167	и 168	й 169	к 170	л 171	м 172	н 173	о 174	п 175
B	▒ 176	▓ 177	▓ 178	│ 179	┤ 180	╡ 181	╢ 182	╖ 183	╕ 184	╣ 185	║ 186	╗ 187	╝ 188	╜ 189	╛ 190	┐ 191
C	└ 192	┴ 193	┬ 194	├ 195	─ 196	┼ 197	╞ 198	╟ 199	╚ 200	╔ 201	╩ 202	╦ 203	╠ 204	═ 205	╬ 206	╧ 207
D	╨ 208	╤ 209	╥ 210	╙ 211	╘ 212	╒ 213	╓ 214	╫ 215	╪ 216	┘ 217	┌ 218	▉ 219	▄ 220	▌ 221	▐ 222	▀ 223
E	р 224	с 225	т 226	у 227	ф 228	х 229	ц 230	ч 231	ш 232	щ 233	ъ 234	ы 235	ь 236	э 237	ю 238	я 239
F	Ё 240	ё 241	Є 242	є 243	Ї 244	ї 245	Ў 246	ў 247	° 248	∙ 249	· 250	√ 251	№ 252	¤ 253	■ 254	nbsp 255

Кодировки

cp1251

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	Ђ 128	Ѓ 129	‚ 130	ѓ 131	„ 132	… 133	† 134	‡ 135	€ 136	‰ 137	Љ 138	‹ 139	Њ 140	Ќ 141	Ћ 142	Џ 143
9	ђ 144	‘ 145	’ 146	“ 147	” 148	• 149	– 150	— 151	™ 152	™ 153	љ 154	› 155	њ 156	ќ 157	ћ 158	џ 159
A	nbsp 160	Ў 161	ў 162	Ј 163	¤ 164	Ґ 165	¦ 166	§ 167	Ё 168	© 169	Є 170	« 171	¬ 172	shy 173	® 174	Ї 175
B	° 176	± 177	І 178	і 179	ґ 180	µ 181	¶ 182	· 183	ё 184	№ 185	є 186	» 187	ј 188	Ѕ 189	ѕ 190	ї 191
C	А 192	Б 193	В 194	Г 195	Д 196	Е 197	Ж 198	З 199	И 200	Й 201	К 202	Л 203	М 204	Н 205	О 206	П 207
D	Р 208	С 209	Т 210	У 211	Ф 212	Х 213	Ц 214	Ч 215	Ш 216	Щ 217	Ъ 218	Ы 219	Ь 220	Э 221	Ю 222	Я 223
E	а 224	б 225	в 226	г 227	д 228	е 229	ж 230	з 231	и 232	й 233	к 234	л 235	м 236	н 237	о 238	п 239
F	р 240	с 241	т 242	у 243	ф 244	х 245	ц 246	ч 247	ш 248	щ 249	ъ 250	ы 251	ь 252	э 253	ю 254	я 255

Кодовое пространство Unicode

- обозначения: U+xxxx, U+xxxxx, U+xxxxxx
- пространтсво разделено на 17 плоскостей по 2^{16} символов
- первые 128 символов совпадают с ASCII
- плоскость 0 (base multilingual plane) содержит основные символы
- остальные плоскости содержат символы редких письменностей
- 2048 кодов U+DC00 - U+DFFF заняты под “суррогатные пары”

Всего символов в Unicode

$$17 \cdot 2^{16} - 2048 = 1112064$$

Кодировки

UTF-8

- Нужен для передачи Unicode по однобайтовым каналам связи
- Начало аналогично первой половине таблицы ASCII
- Обладает свойством самосинхронизации
- Мультибайтная кодировка

0xxxxxxx
110xxxxx 10xxxxxx
1110xxxx 10xxxxxx 10xxxxxx
11110xxx 10xxxxxx 10xxxxxx 10xxxxxx

Кодировки

UTF-16

- Символы из base multilingual plane кодируются как есть
- Символы из других плоскостей кодируются парой слов с помощью “суррогатных пар” U+D800 - U+DFFF
- Важен порядок байт в слове!
- BOM (byte order mark)
- big-endian (от старших разрядов к младшим): U+FEFF (UTF-16BE)
- little-endian (от старших разрядов к младшим): U+FFFE (UTF-16LE)

110110xx xxxxxxxx 110111xx xxxxxxxx

Объединение и дублирование символов

Для многих символов в Unicode есть отдельные коды. Это имеет место по историческим причинам, так как они присутствуют в национальных кодировках.

Новые составные символы в Unicode не добавляются, их нужно “конструировать” из нескольких кодов:

Кодировки

Классы эквивалентности в Unicode

Cyrillic extensions

0400 Ѐ CYRILLIC CAPITAL LETTER IE WITH GRAVE
≡ 0415 Е 0300 ◌̀
0401 Ё CYRILLIC CAPITAL LETTER IO
≡ 0415 Е 0308 ◌̈
0402 Ђ CYRILLIC CAPITAL LETTER DJE
0403 Ѓ CYRILLIC CAPITAL LETTER GJE
≡ 0413 Г 0301 ◌́
0404 Є CYRILLIC CAPITAL LETTER UKRAINIAN IE
0405 Ѕ CYRILLIC CAPITAL LETTER DZE
0406 І CYRILLIC CAPITAL LETTER BYELORUSSIAN-UKRAINIAN I
→ 0049 I latin capital letter i
→ 0456 і cyrillic small letter byelorussian-ukrainian i
→ 04C0 Ӏ cyrillic letter palochka
0407 Ї CYRILLIC CAPITAL LETTER YI
≡ 0406 І 0308 ◌̈
0408 Ј CYRILLIC CAPITAL LETTER JE
0409 Љ CYRILLIC CAPITAL LETTER LJE
040A Њ CYRILLIC CAPITAL LETTER NJE
040B Ћ CYRILLIC CAPITAL LETTER TSHE
040C Ќ CYRILLIC CAPITAL LETTER KJE
≡ 041A К 0301 ◌́
040D Ѝ CYRILLIC CAPITAL LETTER I WITH GRAVE
≡ 0418 И 0300 ◌̀
040E Ў CYRILLIC CAPITAL LETTER SHORT U
≡ 0423 У 0306 ◌̆
040F Џ CYRILLIC CAPITAL LETTER DZHE

Basic Russian alphabet

0410 А CYRILLIC CAPITAL LETTER A
0411 Б CYRILLIC CAPITAL LETTER BE
→ 0183 ƃ latin small letter b with topbar
0412 В CYRILLIC CAPITAL LETTER VE
0413 Г CYRILLIC CAPITAL LETTER GHE
0414 Д CYRILLIC CAPITAL LETTER DE
0415 Е CYRILLIC CAPITAL LETTER IE
0416 Ж CYRILLIC CAPITAL LETTER ZHE
0417 З CYRILLIC CAPITAL LETTER ZE
0418 И CYRILLIC CAPITAL LETTER I
0419 Й CYRILLIC CAPITAL LETTER SHORT I
≡ 0418 И 0306 ◌̆
041A К CYRILLIC CAPITAL LETTER KA
041B Л CYRILLIC CAPITAL LETTER EL
041C М CYRILLIC CAPITAL LETTER EM
041D Н CYRILLIC CAPITAL LETTER EN
041E О CYRILLIC CAPITAL LETTER O
041F П CYRILLIC CAPITAL LETTER PE
0420 Р CYRILLIC CAPITAL LETTER ER
0421 С CYRILLIC CAPITAL LETTER ES
0422 Т CYRILLIC CAPITAL LETTER TE
0423 У CYRILLIC CAPITAL LETTER U
0424 Ф CYRILLIC CAPITAL LETTER EF
0425 Х CYRILLIC CAPITAL LETTER HA
0426 Ц CYRILLIC CAPITAL LETTER TSE
0427 Ч CYRILLIC CAPITAL LETTER CHE
0428 Ш CYRILLIC CAPITAL LETTER SHA
0429 Щ CYRILLIC CAPITAL LETTER SHCHA
042A Ъ CYRILLIC CAPITAL LETTER HARD SIGN
042B Ы CYRILLIC CAPITAL LETTER YERU
042C Ь CYRILLIC CAPITAL LETTER SOFT SIGN
042D Э CYRILLIC CAPITAL LETTER E
042E Ю CYRILLIC CAPITAL LETTER YU
042F Я CYRILLIC CAPITAL LETTER YA
0430 а CYRILLIC SMALL LETTER A
0431 б CYRILLIC SMALL LETTER BE
0432 в CYRILLIC SMALL LETTER VE
0433 г CYRILLIC SMALL LETTER GHE
0434 д CYRILLIC SMALL LETTER DE
0435 е CYRILLIC SMALL LETTER IE
0436 ж CYRILLIC SMALL LETTER ZHE
0437 з CYRILLIC SMALL LETTER ZE
0438 и CYRILLIC SMALL LETTER I
0439 й CYRILLIC SMALL LETTER SHORT I
≡ 0438 и 0306 ◌̆
043A к CYRILLIC SMALL LETTER KA
043B л CYRILLIC SMALL LETTER EL
043C м CYRILLIC SMALL LETTER EM
043D н CYRILLIC SMALL LETTER EN
043E о CYRILLIC SMALL LETTER O
043F п CYRILLIC SMALL LETTER PE
0440 р CYRILLIC SMALL LETTER ER
0441 с CYRILLIC SMALL LETTER ES
0442 т CYRILLIC SMALL LETTER TE
0443 у CYRILLIC SMALL LETTER U
0444 ф CYRILLIC SMALL LETTER EF
0445 х CYRILLIC SMALL LETTER HA
0446 ц CYRILLIC SMALL LETTER TSE
0447 ч CYRILLIC SMALL LETTER CHE
0448 ш CYRILLIC SMALL LETTER SHA
0449 щ CYRILLIC SMALL LETTER SHCHA
044A ъ CYRILLIC SMALL LETTER HARD SIGN
044B ы CYRILLIC SMALL LETTER YERU
→ A651 ъы cyrillic small letter yeru with back yer
044C ь CYRILLIC SMALL LETTER SOFT SIGN
→ 0185 ƅ latin small letter tone six
→ A64F ъ cyrillic small letter neutral yer

Полезные иструменты

- CLD (Compact Language Detector) – C++, Python
  - http://code.google.com/p/chromium-compact-language-detector/
- LanguageDetection – Java
  - http://code.google.com/p/language-detection/
- Apache Tika
  - http://tika.apache.org/
- Видеолекция (Яндекс, RuSSIR 2012)
  http://videolectures.net/russir2012_grigoriev_language/

Определение языка

Подходы

- Графематический
- N-граммный
- Лексический

Графематический подход

Система письменности

- Кириллица
- Латиница
- ...

Алфавит

- Русский А..Я
- Украинский - не используются Ё, Ъ, Ы, Э, но есть Г’, І с точками и т.д.
- Казахский

N-граммный подход

Russian		Ukrainian		English		French	
^п	1.91	^по	0.84	^п	1.97	^на	0.85	^t	3.17	^th	2.00	es	2.31	es$	1.77
^с	1.71	^пр	0.68	^в	1.75	на$	0.73	th	2.48	the	1.62	le	1.97	^de	0.98
^в	1.68	^на	0.66	^н	1.68	^по	0.72	^a	2.41	he$	1.44	^d	1.84	le$	0.82
^н	1.55	^и$	0.61	на	1.45	^пр	0.63	he	2.24	ed$	0.78	^l	1.74	de$	0.76
ст	1.43	^в$	0.60	^з	1.40	^за	0.59	in	1.94	nd$	0.73	on	1.70	^le	0.72
то	1.29	^не	0.56	^с	1.25	^не	0.56	er	1.60	ing	0.73	re	1.48	re$	0.68
но	1.23	ть$	0.48	ро	1.13	ого	0.54	an	1.54	^an	0.72	^c	1.46	nt$	0.58

- Ранговый
- Марковский

Пословный подход

- ???
  - án került vagy től majd új ami ő kategória ben szerint amikor hogy amerikai két ezt mint alatt magyar itt második már ….
- ???
  - cel cod său cu cea l după ro va județul această în către sunt pe toate astfel ani prin ca departamentul din timpul într ….
- ???
  - ayrıca iklimi gibi tarafından olu kültür birlikte ula yol tarihinde veya iyi sonra türk bulunan kar çalı göre oldu ….

Пословный подход

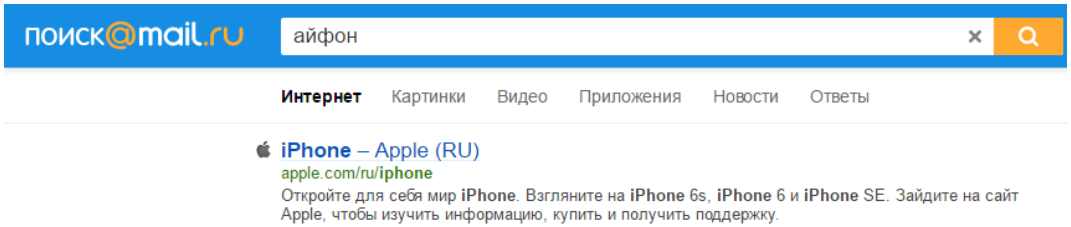
- Hungarian
  - án került vagy től majd új ami ő kategória ben szerint amikor hogy amerikai két ezt mint alatt magyar itt második már ….
- Romanian
  - cel cod său cu cea l după ro va județul această în către sunt pe toate astfel ani prin ca departamentul din timpul într ….
- Turkish
  - ayrıca iklimi gibi tarafından olu kültür birlikte ula yol tarihinde veya iyi sonra türk bulunan kar çalı göre oldu ….

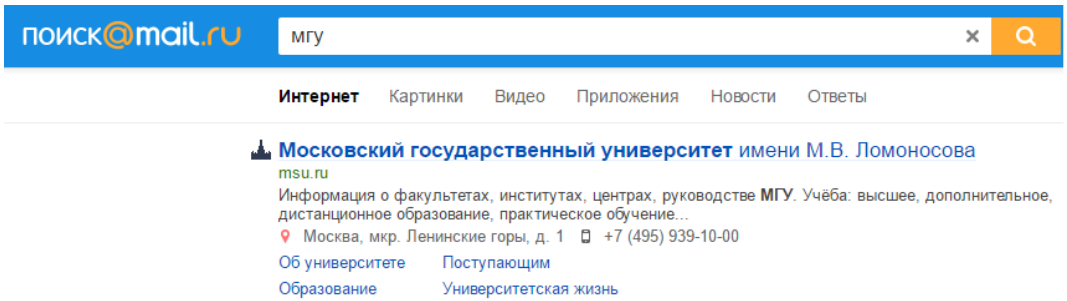
Кореференция: синонимы

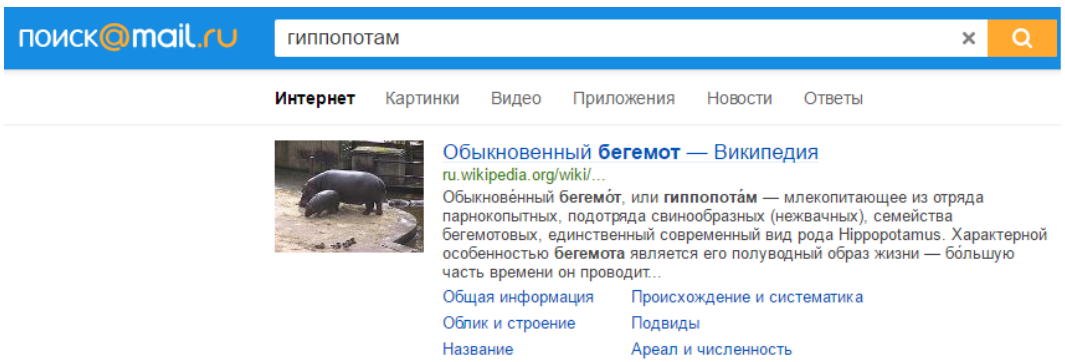
Различные способы названия одного и того же объекта

- Синонимы: [“ШАВЕРМА”, “ШАУРМА”]
- Аббревиатуры: [“БМП”, “БОЕВАЯ МАШИНА ПЕХОТЫ”]
- Транслитерация: [“PLAZMA”, “ПЛАЗМА”]
- Грамматические замены: [“ПОЗДРАВЛЕНИЕ”, “ПОЗДРАВИТЬ”]
- Переводы: [“ВОЗДУШНАЯ ТЮРЬМА”, “CON AIR”]
- Джойны: [“АУДИО КОДЕКИ”, “АУДИОКОДЕКИ”]

Примеры


поиск@mail.ru
айфон
Интернет
Картинки
Видео
Приложения
Новости
Ответы
iPhone – Apple (RU)
apple.com/ru/iphone
Откройте для себя мир iPhone. Взгляните на iPhone 6s, iPhone 6 и iPhone SE. Зайдите на сайт Apple, чтобы изучить информацию, купить и получить поддержку.


поиск@mail.ru
мгу
Интернет
Картинки
Видео
Приложения
Новости
Ответы
Московский государственный университет имени М.В. Ломоносова
msu.ru
Информация о факультетах, институтах, центрах, руководстве МГУ. Учёба: высшее, дополнительное, дистанционное образование, практическое обучение...
Москва, мкр. Ленинские горы, д. 1
+7 (495) 939-10-00
Об университете
Поступающим
Образование
Университетская жизнь


поиск@mail.ru
гиппопотам
Интернет
Картинки
Видео
Приложения
Новости
Ответы
Обыкновенный бегемот — Википедия
ru.wikipedia.org/wiki/...
Обыкнове́нный бегемо́т, или гиппопота́м — млекопитающее из отряда парнокопытных, подотряда свинообразных (нежвачных), семейства бегемотовых, единственный современный вид рода Hippopotamus. Характерной особенностью бегемота является его полуводный образ жизни — бо́льшую часть времени он проводит...
Общая информация
Происхождение и систематика
Облик и строение
Подвиды
Название
Ареал и численность

Источники расширений

- Словари синонимов
- Энциклопедические сайты
  - Википедия
  - Тематические сайты (kinopoisk)
- Скобочные написания в документах
- Логи запросов с кликами
- Переформулировки запросов
- Грамматические преобразования
- Языковые модели и дистрибутивная семантика

Дистрибутивная семантика

Дистрибутивная гипотеза

- Значение лингвистической единицы складывается только из ее употребления, использования.
- В мозге хранится сумма всех тех контекстов, в рамках которых мы слышали или видели то или иное слово.
- Это и есть его смысл. Без знания типичных соседей никакой семантики нет.

Вывод:

- Слова с похожими типичными контекстами имеют схожее значение

Дистрибутивная семантика

Счетные модели

- Совмместная встречаемость
- Косинусная близость

Predictive models

- word2vec

Дистрибутивная семантика

Пример модели

- `http://ling.go.mail.ru/dsm/ru/`

Ближайшие по контекстам слова для *дефицит* (с косинусной близостью)

1. нехватка 0.65354729
2. профицит 0.62004328
3. госдолг 0.61431211
4. несбалансированность 0.59282905
5. дефицитность 0.58463109
6. дисбаланс 0.57497036
7. недопоступление 0.57044709

Ближайшие по контекстам слова для *сингл* (с косинусной близостью)

1. альбом 0.91364485
2. лонгплей 0.88309568
3. пластинка 0.85516018
4. кавер 0.84436756
5. студийник 0.83440500

Стемминг

Усечение окончаний

- Отсекаем самое длинное возможное окончание от слова, надеемся, что это не сильно ухудшит результат по сравнению с лемматизацией
- Для каждого языка свои таблицы окончаний
- Для некоторых языков другие аналогичные методы
- Результат выходит забавный :)
  - Сочи и сочиться приводятся к одной форме

Стемминг

Примеры

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Стемминг

Преимущества и недостатки

- Усечение окончаний улучшает одни запросы и ухудшает другие
- Дает в большинстве случаев увеличение полноты
- При этом часто ухудшает точность
- Алгоритм Портера определяет следующий класс эквивалентности
  - operate operating operates operation operative operatives operational
- Ясно что в запросах станет хуже
  - operational AND research
  - operating AND system
  - operative AND densistry

Лемматизация

Цель: привести все разные формы одного слова к начальной (каноничной).

Примеры:

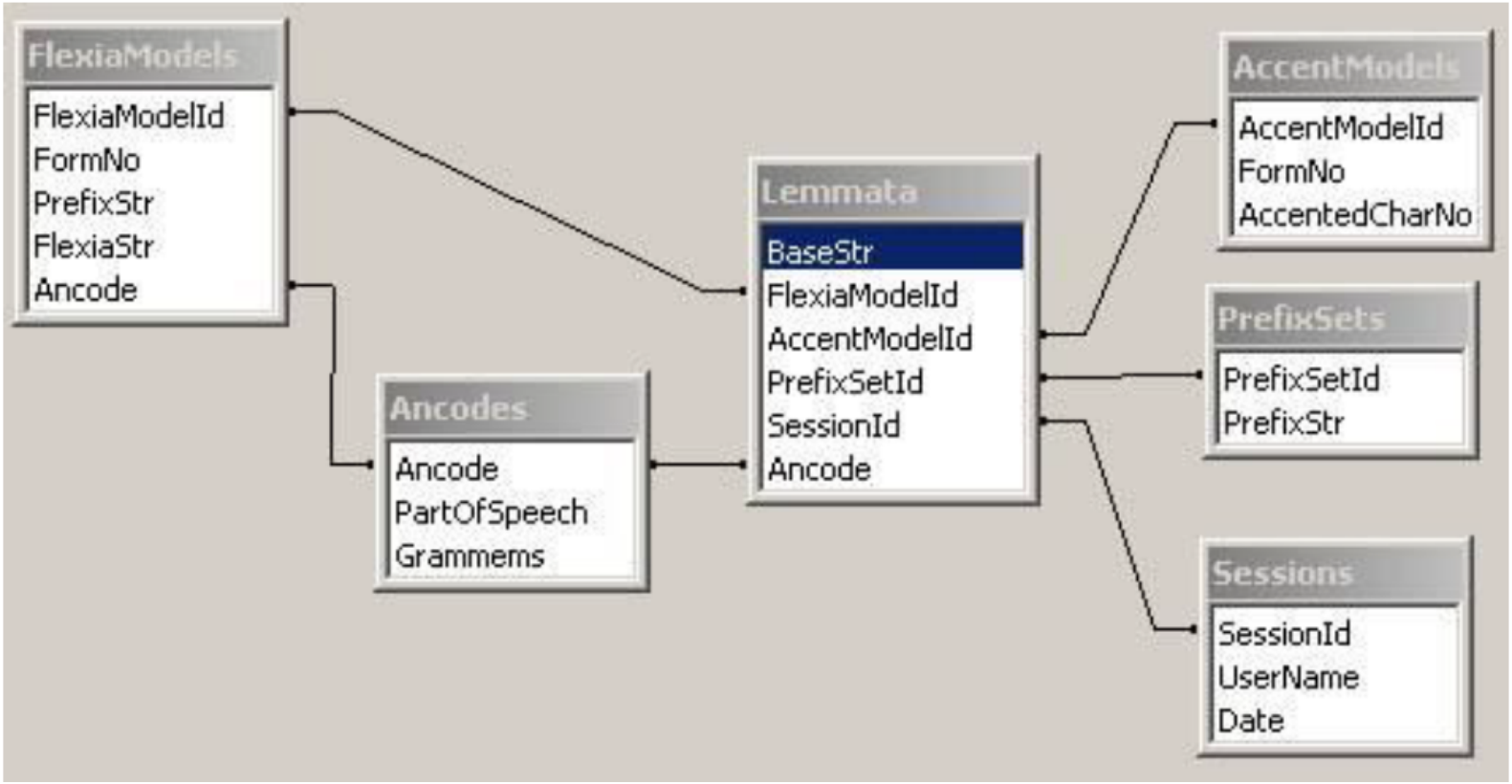
- am, are, is $\rightarrow$ be
- car, cars, car's, cars' $\rightarrow$ car
- the boy's cars are different colors $\rightarrow$ the boy car be different color

Лемматизация заключается в поиске начальной формы (леммы) в словаре

Лемматизация

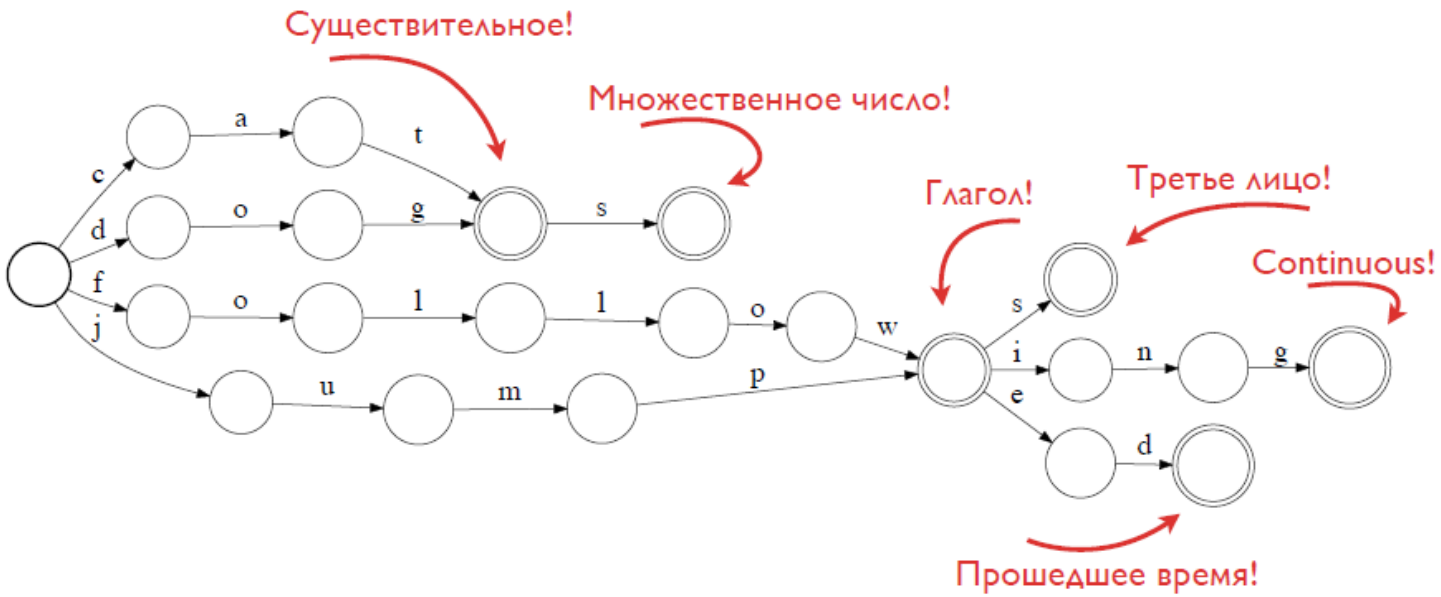
- Слово = машинная основа + парадигма
- Парадигма
  - Парадигм а
  - Парадигм ы
  - Парадигм е
  - Парадигм у
  - Парадигм ой
  - Парадигм е

Лемматизация



- (c) Сокирко А.В., `http://aot.ru/docs/sokirko/Dialog2004.htm`

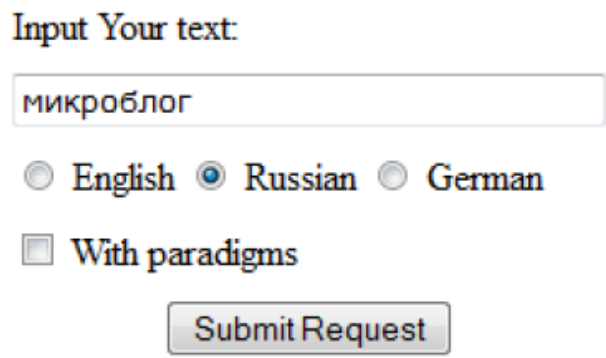
Лемматизация



Лемматизация

- Что делать со словами, которых нет в словаре?
- Ищем похожие!

Предиктор



Found	Dict ID	Lemma	Grammems	
-	но,	МИКРОБЛОГ	С мр,вн,им,ед,	АНАЛОГ
-		МИКРОБЛГИЙ	КР_ПРИЛ но,од,мр,ед,	НЕДОЛГИЙ

Откуда взять морфологию?

Английская

- Стеммер Портера (Porter)
- Стеммер Ловинса (Lovins)

Русская

- aot.ru
- keva.ru (СтемКа)
- MyStem (`http://company.yandex.ru/technologies/mystem/`)
- PyMorphy

Стеммер можно сделать самостоятельно

- Snowball - фреймворк для алгоритмов стемминга

Какие проблемы?

Языки сильно отличаются

- Изолирующие/Аналитические
- Синтетические
  - Флективные
  - Агглютинативные
- Полисинтетические

Флективные языки

Индо-европейские

- Русский
- Испанский
- Немецкий

Агглютинативные языки

Тюркские языки

Turkish	English
ev	(the) house
evler	(the) houses
evin	your (sing.) house
eviniz	your (pl./formal) house
evim	my house
evimde	at my house
evlerinizin	of your houses
evlerinizden	from your houses

Полисинтетические языки

Полисинтетические языки — языки, в которых все члены предложения или некоторые компоненты словосочетания соединяются в единое целое без формальных показателей у каждого из них.

- Чукотско-камчатские
- Эскимосско-алеутские

Тымэйӈылевтпыгтыркын

(*t-ə-mejŋ-ə-levt-pəγt-ə-rkən*)

У меня сильно болит голова.

Омонимия

Разные по смыслу слова имеют одинаковое написание

Примеры:

белки бегали по лесу и ели орехи
(Лемма: белка, сущ. жен. род)

белки различаются по степени растворимости в воде
(Лемма: белок, сущ. муж. род)

- *Словоформа* - конкретная морфологическая разновидность слова

белка, белку, белкой, белке

Неоднозначность

Английский

- Leg
- Chair

Русский

- Лук
- Очки
- Лист

Снятие омонимии

Rule-based

```
//~ TN 26.1.a
//~ TC Если К - омоним с прилагательным или наречием
//~ TA И справа - enough
//~ TA И если омоним - переходный глагол, и справа от
enough не существительное
else if ((IsAdj(k) || IsAdverb(k))
    && CheckQuantitativeParticular(k + 1, QP_ENOUGH)
    && !CheckPrepParticular(k + 2, PP_OF)
    && !(IsTransitiveVerb(k) && IsNoun(k + 2)))
{
    //~ TD  Тогда не глагол
    NOT_VERB(k);
}
```

Снятие омонимии

Rule-based

```
//~ TN 28.
//~ TC Если k - омоним глагол/существительное и k не прилагательное и не глагол 'need',
//~ TA И слева есть ИГ
//~ TA И за ИГ следует глагол (не омоним)
//~ TA который не управляет инфинитивом без to
//~ TA не управляет придаточным
//~ TA и не является строго непереходным
//~ TA и справа с пропуском наречий to + infinitive
//~ TA        Или K - соч. союз или предлог
//~ TA        Если K переходный И справа не начинается ИГ И справа не DO
//+ The Princess of Wales has opened a new Aids CENTRE in south-east London.
else if (IsVerb(k) && IsNoun(k) && !IsAdj(k)
            && !CheckVerbSemantic(k, VS_NEED)  //~ не глагол 'need'
            && (nTmp = SearchNGAtLeftBeg(LEFT_K, SNGLM_NORMAL))
            && (nTmp2 = SkipAdvOmon(nTmp, 5, SAO_LEFT))
            && (IsVerb(nTmp2)
                  && IsOnePartOfSpeech(nTmp2)
                  && !IsParticipleI(nTmp2)
                  && !IsVerbObjBareInfControl(nTmp2)
                  && CheckVGClauseGovernment(GetPrizn(nTmp2), VCG_NOT_SET)
                  && !IsVGStrictlyIntransitive(GetPrizn(nTmp2)))
            && (nTmp = SkipAdvHomo(k + 1))
            && (IsTo(nTmp) && IsInfinitive(nTmp + 1)
                        && !(IsVerbInfControl(k) && !IsNounInfControl(k))
                  || IsCoConj(k) || IsPrep(k)
                  || IsTransitiveVerb(k) && !NGCheck(NGM_BEGIN, NGO_ALL, k + 1) && !IsDo(nTmp)))
{
      //~ TD Тогда не глагол
      NOT_VERB(k);
}
```

Статистическое снятие омонимии

Скрытые марковские модели

Задача: Приписать наиболее вероятным образом каждому слову w_k в предложении тег t_k

По формула Байеса $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ получаем

$$P(t_1 \dots t_m | w_1 \dots w_m) = \frac{P(w_1 \dots w_m | t_1 \dots t_m) P(t_1 \dots t_m)}{P(w_1 \dots w_m)}$$

Нужно найти для данного предложения набор тегов, который делает эту вероятность максимальной.

Статистическое снятие омонимии

Скрытые марковские модели

Предположение: Распределение тегов подчиняется марковскому свойству:

$$P(w_1 \dots w_m | t_1 \dots t_m) P(t_1 \dots t_m) \approx \prod_{k=1}^{m} P(w_k | t_k) P(t_k | t_{k-(n-1)} \dots t_{k-1})$$

Вероятности могут быть оценены как частоты по большому размеченному корпусу (`http://ruscorpora.ru`)

Для поиска оптимальных тегов используется динамический **алгоритм Витерби**.

Вопросы