ИССЛЕДОВАНИЕ ГОРОДСКОЙ ДИАЛЕКТНОЙ ЛЕКСИКИ С ПОМОЩЬЮ ПОИСКОВЫХ СИСТЕМ

А.В. Богданов

Задача

Показать, какие средства на сегодняшний день предоставляет сеть Интернет для диалектологических и прочих лингвистических исследований.

Проблемы, встающие перед исследователем

- География пользователя
- Усреднение данных с поправкой на количество пользователей и жителей данного региона
- Шум в поисковой выдаче

Пример исследования: географическое распределение слова «плойка»

Город	Кол-во сайтов	Интернет-индекс	(Кол-во сайтов)/ (Интернет-индекс)
Новосибирск	51	568	0,0898
Ростов-на-Дону	18	202	0,0891
Киев	41	476	0,0861
Челябинск	30	362	0,0829
Томск	12	185	0,0649
Екатеринбург	39	662	0,0589
Пермь	15	356	0,0421
Оренбург	6	149	0,0403
Вологда	4	151	0,0265
Саратов	5	245	0,0204
Мурманск	4	211	0,019
Тверь	3	162	0,0185
Владивосток	6	349	0,0172
Ярославль	4	254	0,0157
Ставрополь	3	224	0,0134
Краснодар	5	376	0,0133
Самара	5	480	0,0104

Пример исследования: портрет аудитории «блогосервисов»

Взято из блога sheldon-j.livejournal.com

Запрос	Сервис	Коэффициент популярности сервиса	Найдено записей	Частота
«прив»	Livejournal	594	2937	4,94444444
	Liveinternet	301	15464	51,3754153
	blogs.mail.ru	68	2317	34,0735294
«пасиб»	Livejournal	594	14420	24,2760943
	Liveinternet	301	16306	54,1727575
	blogs.mail.ru	68	1079	15,8676471
«кажеться»	Livejournal	594	13306	22,4006734
	Liveinternet	301	17612	58,5116279
	blogs.mail.ru	68	2963	43,5735294
«пробывал»	Livejournal	594	3333	5,61111111
	Liveinternet	301	2949	9,79734219
	blogs.mail.ru	68	516	7,58823529

Ключевые моменты метода исследования

- Подбор поискового запроса, однозначно задающего некоторую характеристику
- Вычисление индекса (коэффициента) популярности источника
- Возможность сужения области поиска

Заключение и перспективы

- На сегодняшний день сеть Интернет предоставляет достаточно разнообразных средств для проведения подобных исследований
- Подобные исследования могут проводиться автоматически и являться основой для разметки ресурсов / текстов в сети Интернет