Яндекс

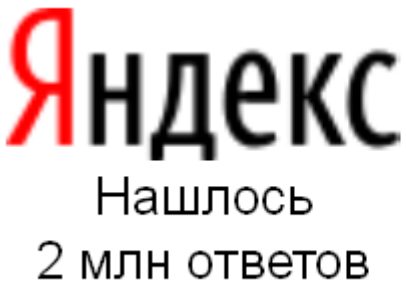
TO FIND OUT OR TO BUY? PRODUCT REVIEW VS. WEB SHOP CLASSIFIER

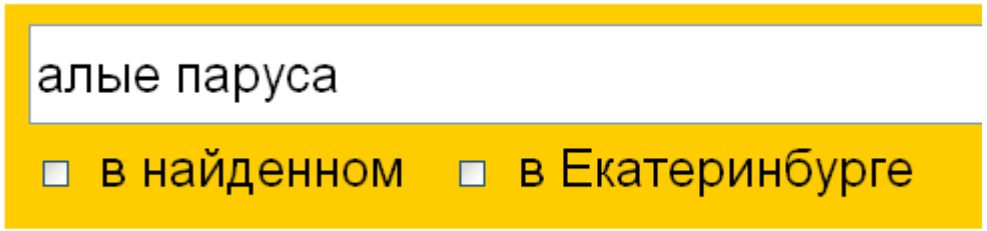
Pavel Braslavski

Yuri Kiselev

Dialog 2011

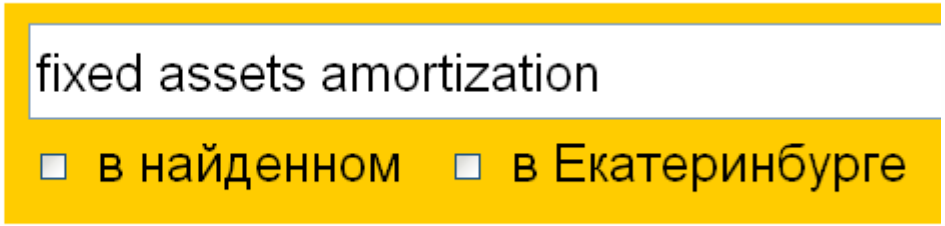
What does user actually mean?





- a novel by Alexander Grin?
- its screen version of the 1961?
- a residential complex in Moscow?

Who is behind the query?

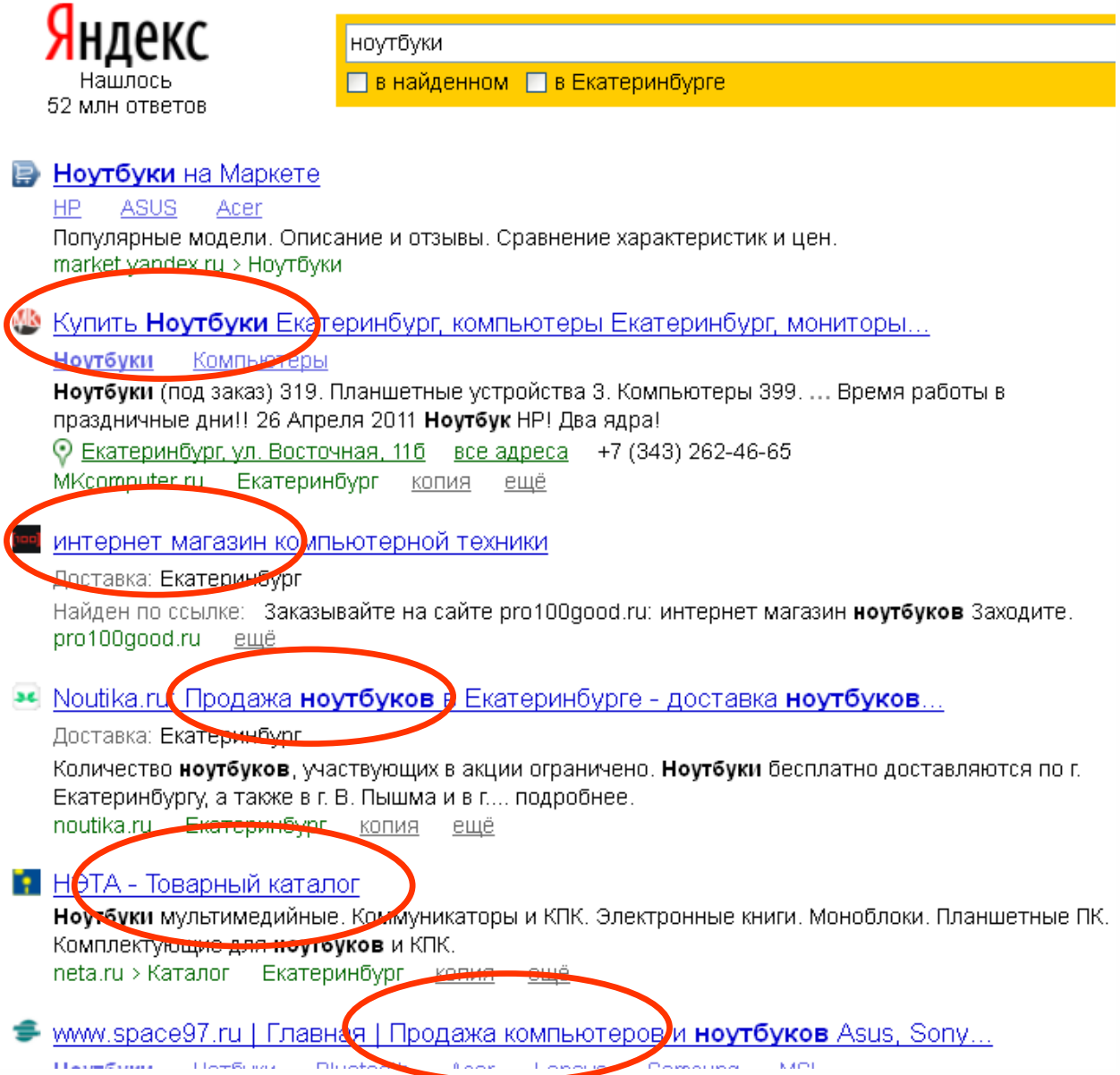


- experienced accountant
- college student

The diversity of search results

When it is impossible to disambiguate the query, we can try to organize the result list the way it reflects the different intents of the query.

Skew of a search engine result list



Product queries

query: [*samsung g400*]

query: [*home air conditioner*]

query: [*asus motherboard*]

query: [*lg monitor*]

query: [*gps navigation systems*]

Main intents

- comparison, studying the properties of goods
- purchase
- technical documentation
- software for devices
- etc.

Web documents classifier

Two types of documents:

1) online surveys and **reviews**;

2) pages of **online shops** you can immediately make an order on.

Online shop

EKBSHOP
интернет-магазин
электроники

Контакты
Тел.: (343) 219-17-88, 8-908-920-8888
ICQ: 176272

Ваша корзина
Товаров: 0 шт.
Всего к оплате: 0 руб.

О магазине | Доставка и Оплата | Гарантия | Статус заказа | Контакты

ПОИСК: Выберите категорию от до

- Телекоммуникационное оборудование
- Сотовые телефоны и коммуникаторы
  - Apple iPhone
  - Коммуникаторы ASUS
  - Коммуникаторы Glofish Eten
  - Коммуникаторы HTC
  - Сотовые телефоны LG
  - Сотовые телефоны Motorola
  - Сотовые телефоны Nokia
  - Сотовые телефоны Philips
  - Сотовые телефоны Samsung
  - Сотовые телефоны Sony Ericsson
  - Сотовые телефоны Toshiba
  - Bluetooth USB-адаптеры
  - Bluetooth гарнитуры
  - Проводные гарнитуры
  - Зарядные устройства для телефонов
  - Аккумуляторы для телефона
- Аудио и Видео
- Автомобильные устройства
- Видеокамеры
- Мини-АТС и факсы
- Карты Памяти

htc smart mobility

Главная » Сотовые телефоны и коммуникаторы » Коммуникаторы HTC

Коммуникаторы HTC купить в Екатеринбурге!

Новейшие **коммуникаторы HTC** от именитого производителя, зарекомендовавший себя исключительно с лучшей стороны. Разработчики заменили привычный джойстик на так называемый джог-болл - миниатюрный вариант трекболла. HTC оснащены не только всеми беспроводными модулями (от телефонного GSM/GPRS/EDGE до Bluetooth и Wi-Fi), но и GPS-приёмником.

1 2 >> | показать все

Коммуникатор HTC A3333 Wildfire grey

- Экран TFT 3,2", 16777216 цв.
- 64-голосная полифония
- MP3-мелодии
- Камера 5 Мрх
- Bluetooth, USB, Wi-Fi
- Размеры 107 x 60 x 13 мм

9 500 руб.

В корзину

Коммуникатор HTC A3333 Wildfire red

- Экран TFT 3,2", 16777216 цв.
- 64-голосная полифония
- MP3-мелодии
- Камера 5 Мрх
- Bluetooth, USB, Wi-Fi
- Размеры 107 x 60 x 13 мм

9 500 руб.

В корзину

Коммуникатор HTC A3333 Wildfire white

- Экран TFT 3,2", 16777216 цв.
- 64-голосная полифония
- MP3-мелодии
- Камера 5 Мрх
- Bluetooth, USB, Wi-Fi
- Размеры 107 x 60 x 13 мм

9 500 руб.

В корзину

Review document

- Intel Centrino 802.11a/b/g/n wireless
- Bluetooth 2.1+EDR
- 90W Smart AC adapter with HP Fast Charge
- 8-cell (62 WHr) Li-Ion battery
- Weight: 6.05 lb
- Dimensions (w x d x h): 14.72 x 9.86 x 1.34 inches
- MSRP: **$1,499.00** (starting price: $1,099.00)

Build and Design

In recent years the HP EliteBooks have distinguished themselves in the business world thanks to an exterior design featuring brushed-metal cladding and an interior chassis made of durable magnesium alloy. In short, the EliteBooks look cool and are built tough. The latest generation of HP EliteBooks takes that heritage a few steps further with what HP calls its "FORGE" design philosophy. FORGE is actually an acronym for the words, "Form, Optimization, Richness, Green and Enduring." If you want to overlook that marketing fluff for a moment, what HP is trying to say is that these notebooks are stylish, offer excellent performance, deliver a premium feel, provide efficient power management for long battery life, and are very well built to survive the rigors of business use.

Those might sound like bold claims but, based on what we've seen from the previous generation of EliteBooks, those marketing claims have a very real basis in fact.

One of the first things you'll notice when you pick up the new 15-inch EliteBook 8560p is that the notebook feels like it's made of very thick chunks of aluminum. You won't feel the "flex" that you normally see in cheaper plastic laptops. The

Classifier

To compose a three-class classifier, we built two binary classifiers: “*shop – other*”, “*review – other*”.

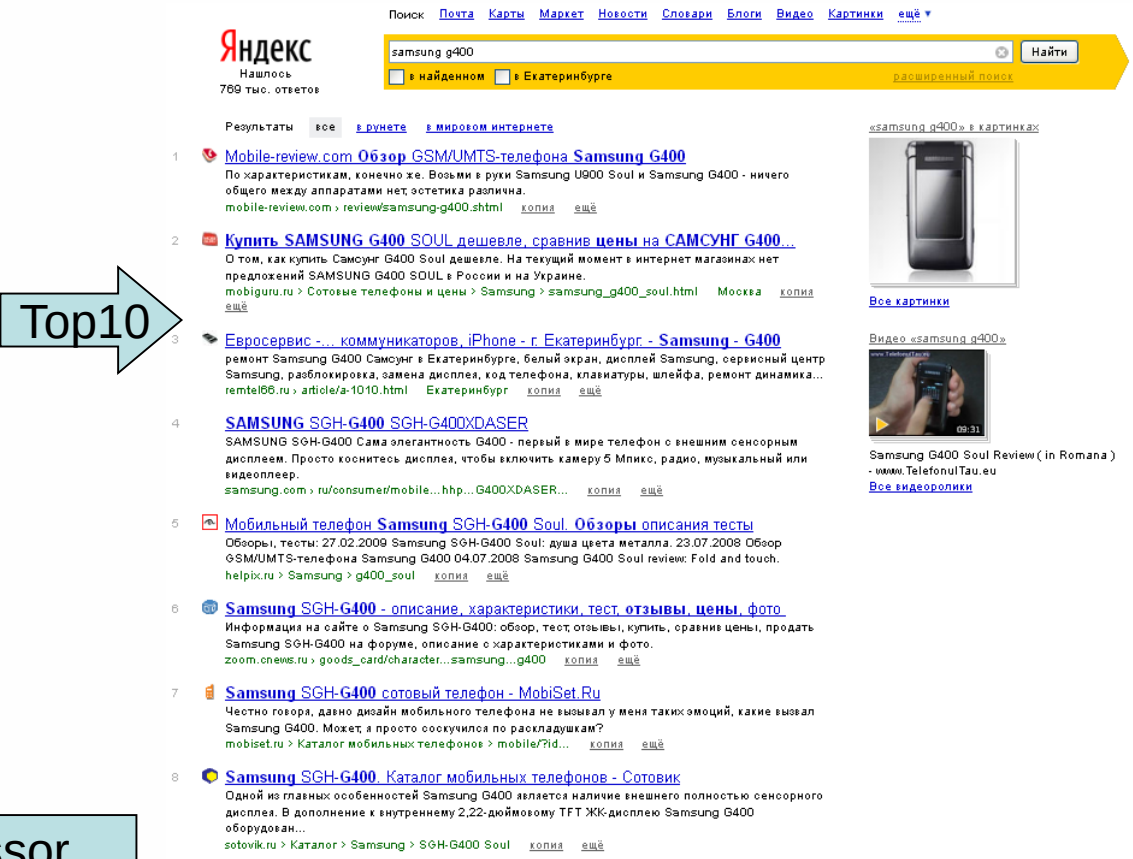
Classifier:

- high performance → light-weight features
- LIBSVM

Learning sample

Query list	
1	samsung g400
2	monitor LG
3	ручная газонокосилка
	…
110	кофеварка Delonghi

Top10



Assessor

Labeled set

Shop / Review / Misc

Learning sample – shop classifier

Class	# of pages
Shop	301
Review	87
Misc	591
Total	979

Learning sample – review classifier

Class	# of pages
Review	150
Misc	150
Long docs	50
Total	350

Test sample

Class	# of pages
Shop	431
Review	101
Misc	557
Total	1089

The test sample was obtained the same way as the shop training sample.

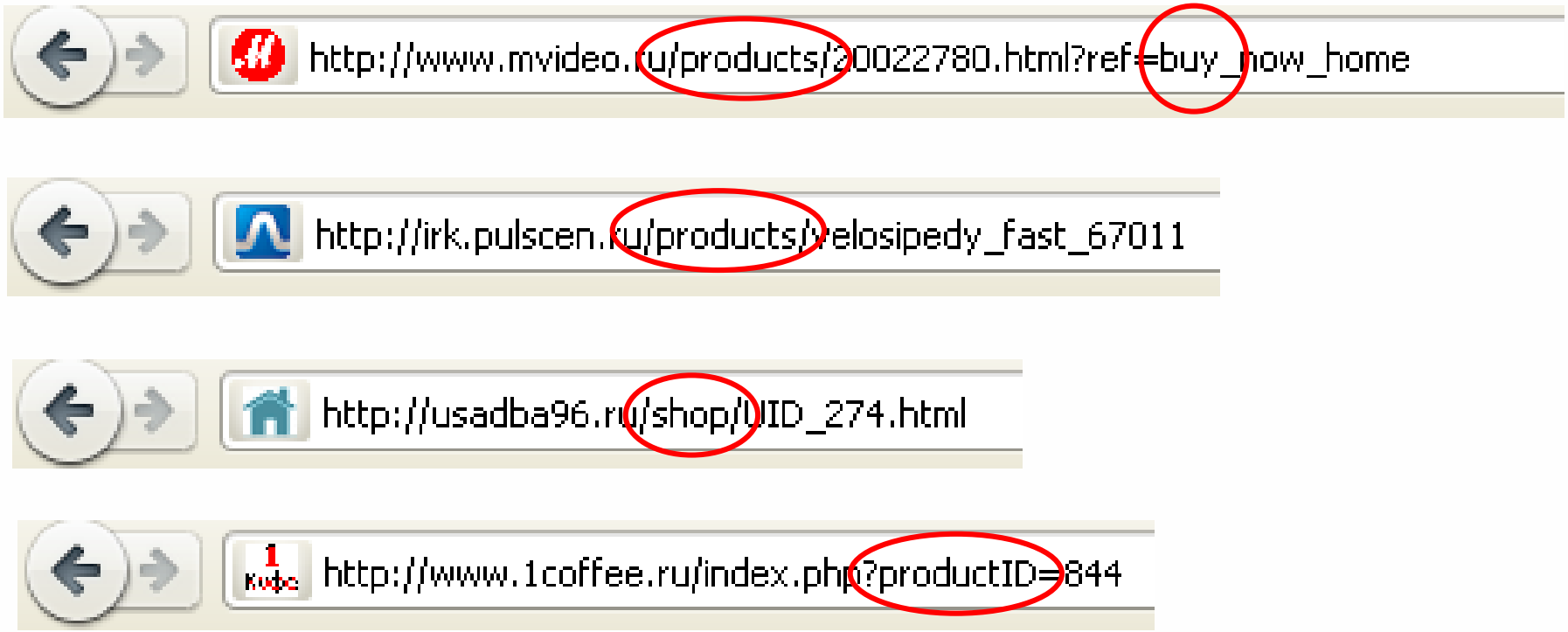
Classification features

- ***Term features***. The most contrast terms for shop documents are магазин, рубль, каталог, цена, прайс, and корзина (*shop*, *ruble*, *catalog*, *price*, and *basket*).
- ***HTML features***. The possibility to make an order (“buy” button detection)

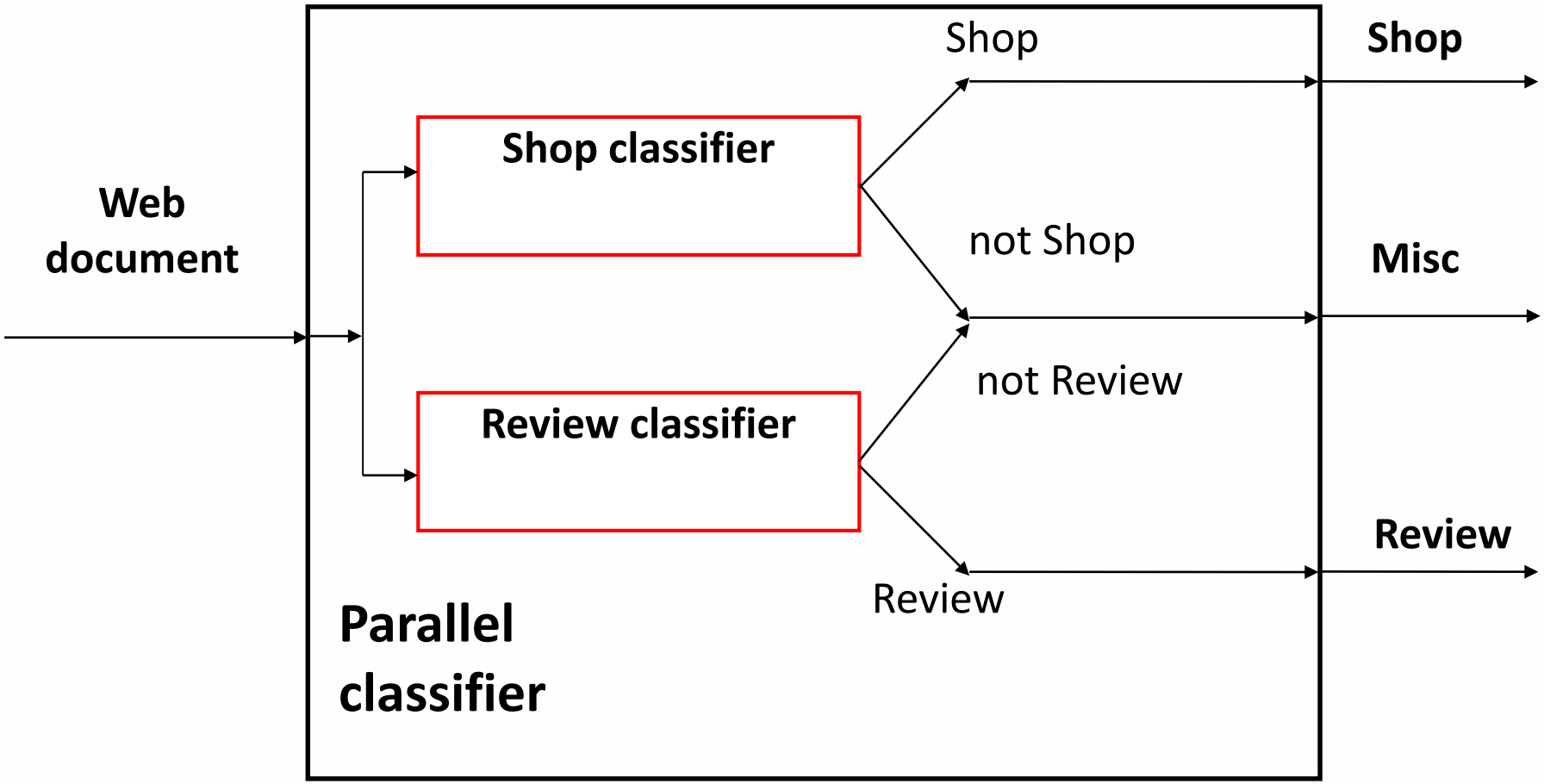
Classification features - 2

- ***Lexical features***. *Dictionary of trademarks* and brands (Sony, Samsung, Asus) presented in Yandex.Market service + *appraising adjectives* (стильный, ужасный, прекрасный, отвратительный – *stylish*, *awful*, *wonderful*, *terrible*)
- **Textual features**. The length of document in words and characters, distribution of sentence lengths, and etc.

URL features



Parallel classifier



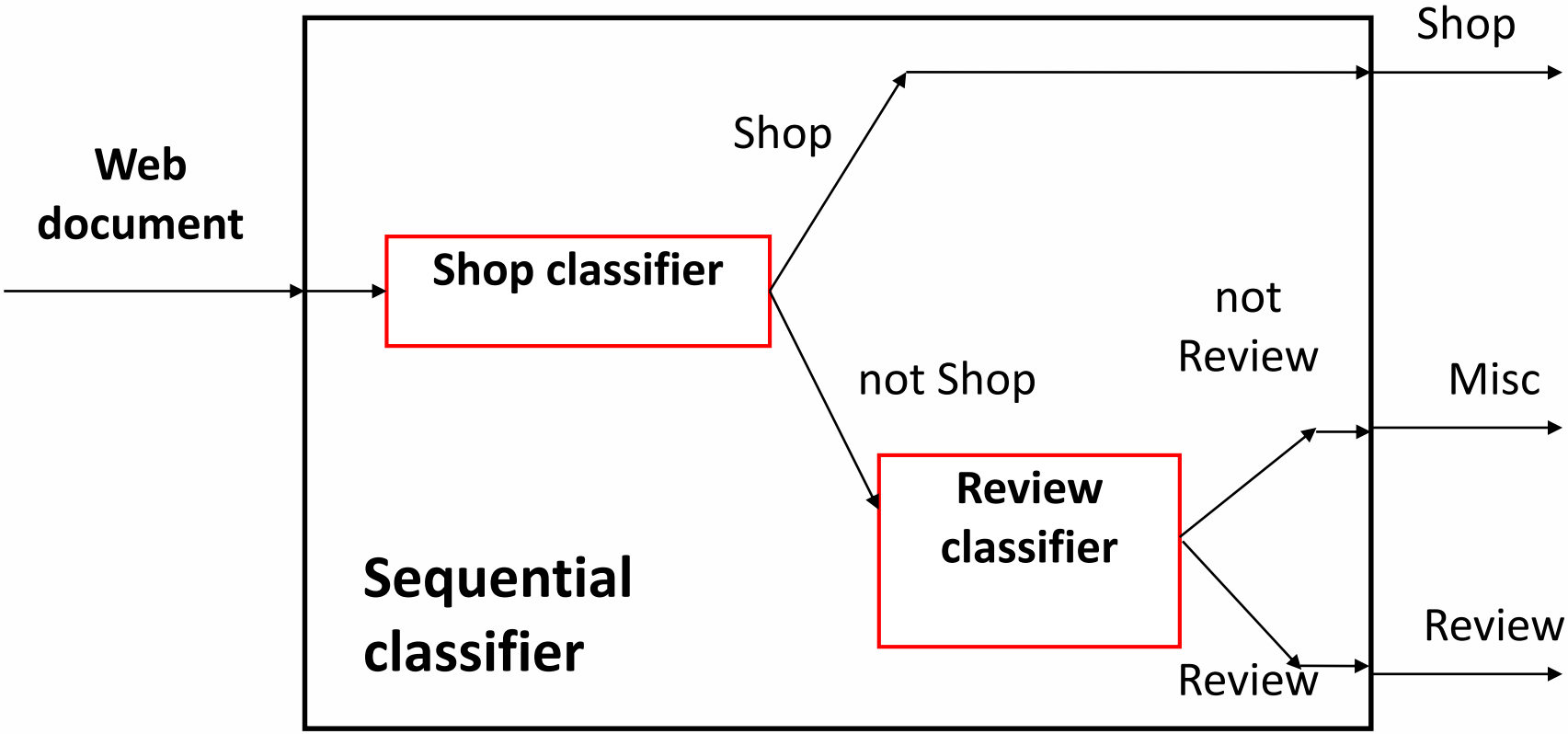
Shop classification results

Set of features	Precision	Recall
Terms only	0.918	0.809
HTML features only	0.894	0.491
Term + HTML features	0.934	0.800
Term + lexical features	0.910	0.807
Term + URL features	0.876	**0.856**
All features	**0.937**	0.837

Review classification results

Set of features	Precision	Recall
Terms only	0.644	0.861
Term + URL features	0.643	0.841
Term + lexical features	0.625	0.861
Term + textual features	**0.681**	**0.891**

Sequential classifier



Confusion matrix of the three-class classifier

	Shop	Review	Misc	**Recall**
Shop	361	3	67	**0.84**
Review	1	90	10	**0.89**
Misc	23	23	511	
Precision	**0.94**	**0.79**		

Conclusion and future work

- Light-weight features
- Small learning samples
- Acceptable quality of classification
- High performance

We want to improve the review classifier:

- Offline classification
- Add more sophisticated features

Pavel Braslavski (pb@yandex-team.ru)

Yuri Kiselev (yurikiselev@yandex-team.ru)