An Introduction to Social Mining

Vladimir Gorovoy* and Yana Volkovich†

†@yvolkovich
Barcelona Media, Information, Technology & Society Group
Barcelona, Spain

* @vgorovoy
Yandex, Yandex.Uslugi
Saint Petersburg, Russia

August, 15-19 2011

RuSSIR
Russian Summer School in Information Retrieval

Outline

RuSSIR
Russian Summer School in Information Retrieval

Vladimir Gorovoy

- Head of Yandex.Uslugi, Yandex
- Dipl. Eng. Degree in Mathematics and Computer Science from the Saint Petersburg State University

Yana Volkovich

- Research Scientist in Information, Technology and Society Group, Barcelona Media Innovation Center
- Ph.D. in Applied Mathematics from the University of Twente
- Dipl. Eng. Degree in Mathematics and Computer Science from the Saint Petersburg State University

An Introduction to Social Media

Outline

day 1 An introduction to Social Media; Social Market; Practical task announcement;

day 2 Yandex.Market;

day 3 Social graph mining; Recommended deadline for the practical task;

day 4 Twitter, Foursquare, etc.; Results for the practical task;

day 5 New research directions; Presentations by the practical task winners.

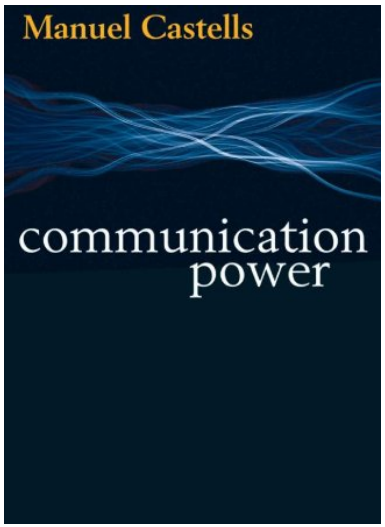
RuSSIR
Russian Summer School in Information Retrieval

- **What is Social Media?**

- ***"Social media is like teen sex. Everyone wants to do it. No one actually knows how."***
  (Avinash Kaushik, Google's analytics evangelist)

- **Social Media** is not only about **Social Networks**

- **Social Media** is a media for social interaction using highly accessible and scalable communication techniques.
- **Social Media** is the use of web-based and mobile technologies to turn communications into interactive dialog.



- “In contrast to **one-to-many communication** structure of traditional mass-media, **social media** allows the emergence of **many-to-many communication**, and gives a rise to mass self-communication” [Castells, 2009]

- **What are the goals /purposes of Social Media?**

social communication emails, mobiles, forums, chats;
social networking facebook, google+;
social blogging/microblogging twitter, livejournal, blogger;
social sharing flickr, vimeo, youtube;
social news digg, slashdot, cnn ireport;
social bookmarking delicious, citeulike;
social knowledge, wikis wikipedia, tripadvisor;
social shopping groupon, amazon, ebay;
social apps & games foursquare, farmville;
etc.

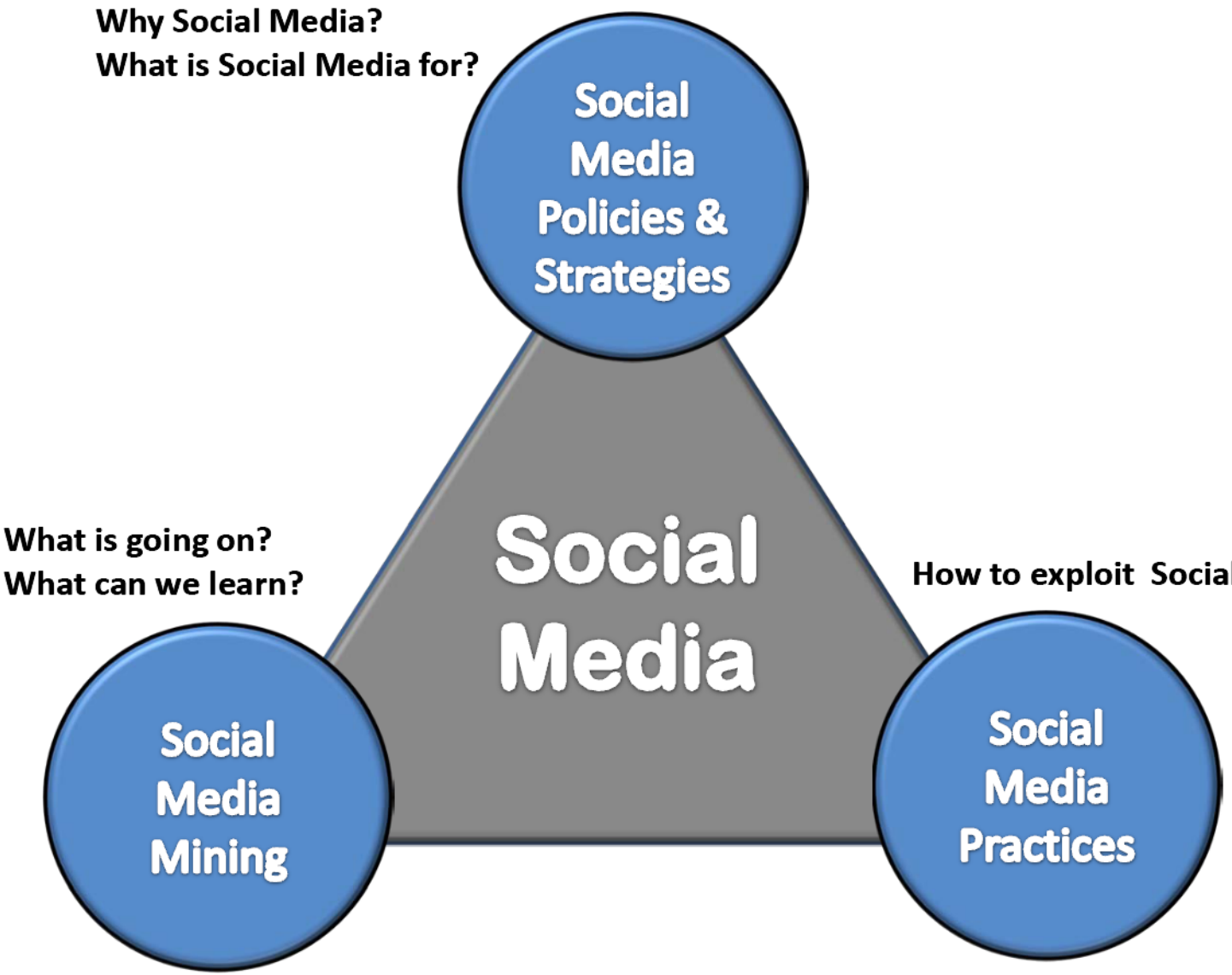
too much data!

- **What could we do?**

to ask right questions

Introduction

Questions



Opinion Mining

Opinion mining

History

- 1993: cartoon by Peter Steiner published by The New Yorker on July 5, 1993

"On the Internet, nobody knows you're a dog."

Opinion mining

History

- 2011:

Opinion mining

Introduction

- people search for and are affected by online opinions;
- Consumer reviews are significantly more (1̃2 times) trusted than descriptions that come from manufacturers. (eMarketer, Feb. 2010)
- 90% of consumers online trust recommendations from people they know; 70% trust opinions of unknown users. (Econsultancy, Jul. 2009)

RuSSIR
Russian Summer School in Information Retrieval

Opinion mining

Introduction (cont.)

- People express their opinions via
  - voting;
  - pressing *like* or *+1*;
  - rating;
  - commenting;
  - sharing;
  - etc.

Opinion mining

Introduction (cont.)

- People evaluate/reflect on
  - items;
  - real events;
  - other people;
  - items created by others.

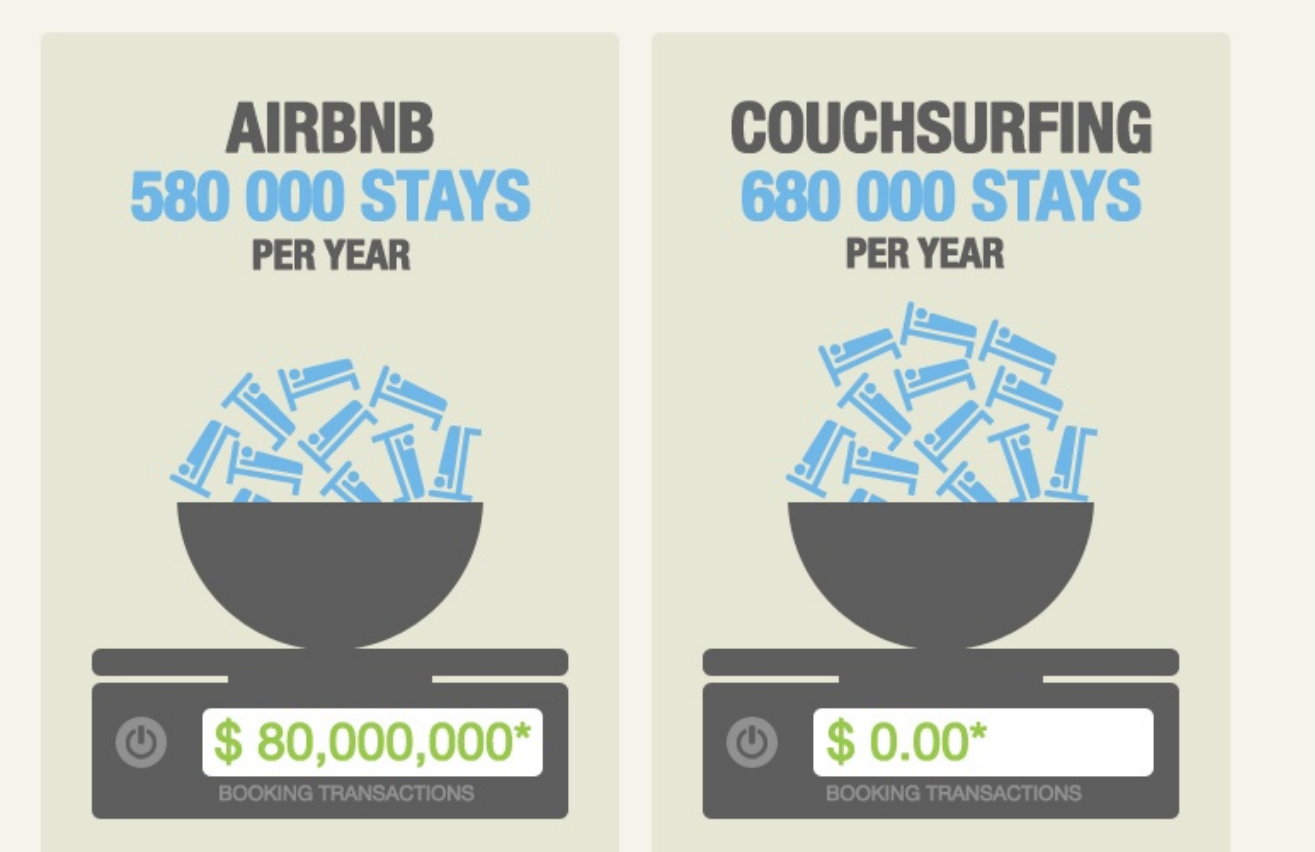
Opinion mining

Examples

Opinion mining

CouchSurfing

- CouchSurfing is a hospitality exchange network and website with 3 million members in 246 countries and territories;



Opinion mining

CouchSurfing (cont.)

- survey study by [Adamic et al., 2011].
- different level of participation:
  - some prefer to host as it allows them to meet people without leaving home.
  - some use the site mainly for travel. (One interviewed participant had been couchsurfing nonstop for a year.)

Opinion mining

Rating people

- Discomfort in leaving negative references: Negative ratings are seldom given publicly in part because the individual being rated can reciprocate. [Adamic et al., 2011].
- Textual references (and their number) are far more important

	Very Important	Important	Neutral	Unimportant	Very Unimportant
how many vouches they have	8.6%	29.3%	37.1%	15.8%	6.9%
whether they are verified	9.1%	23.0%	36.4%	18.9%	11.6%
number of references received	23.0%	57.0%	13.7%	4.4%	1.1%
text of references received	47.6%	40.8%	8.2%	2.1%	0.8%
number of friends	4.4%	23.6%	37.7%	25.1%	8.0%
friends' friendship level	3.2%	19.0%	35.6%	25.0%	13.5%

RuSSIR
Russian Summer School in Information Retrieval

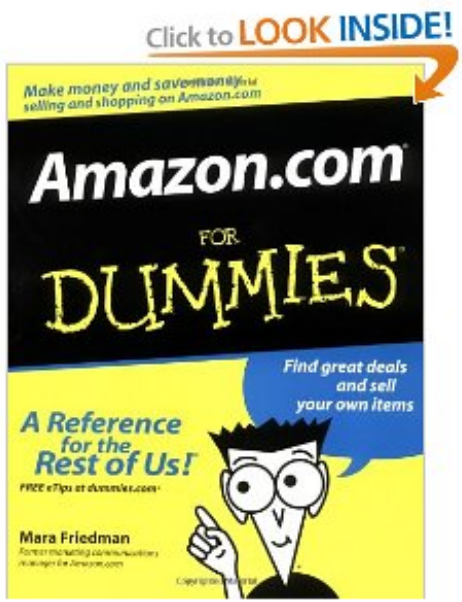
Opinion mining

Rating items and rating ratings

opinion: What does A think about this item? [Pang and Lee, 2008]

meta-opinion: What do other users think about A's opinion about this item? [Danescu et al., 2009]

- Amazon.com for Meta-Opinion Analysis (Danescu et al. [2009])



Share your own customer images

Search inside this book

Amazon.com For Dummies [Paperback]

Mara Friedman (Author)

(17 customer reviews) | Like (0)

Available from these sellers.

17 new from $5.01 26 used from $1.81

Formats	Amazon Price	New from	Used from
Kindle Edition	$17.47	--	--
Paperback	--	$5.01	$1.81

Opinion mining

Rating reviews: Question

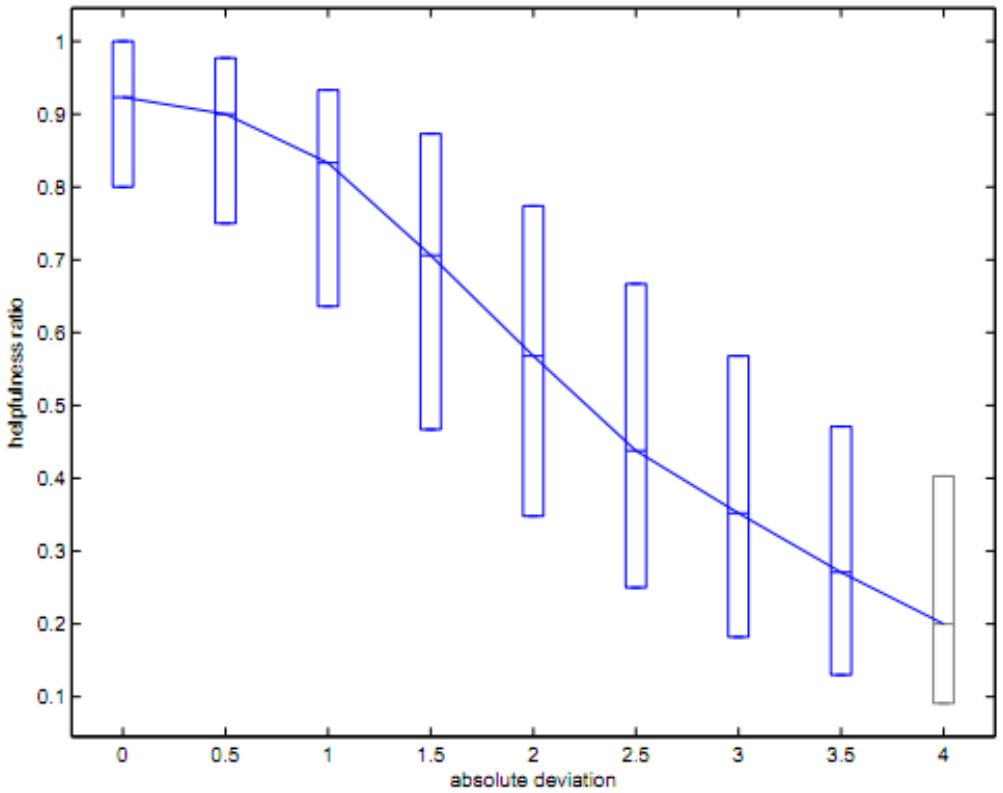
- A product has a average star rating of $\star\star\star$.
- Aim is to write a helpful review for the product.
- Which would be your star rating choice if you can only alter the star rating of the review?

- Social Psychology Hypotheses:
  - **Conformity** star rating is closer to the average star rating for the product;
  - **Brilliant but cruel** star rating is below to the average star rating for the product;
  - **Individual bias** star rating reflects the evaluators' personal opinion about the product.

Opinion mining

Rating reviews (cont.)

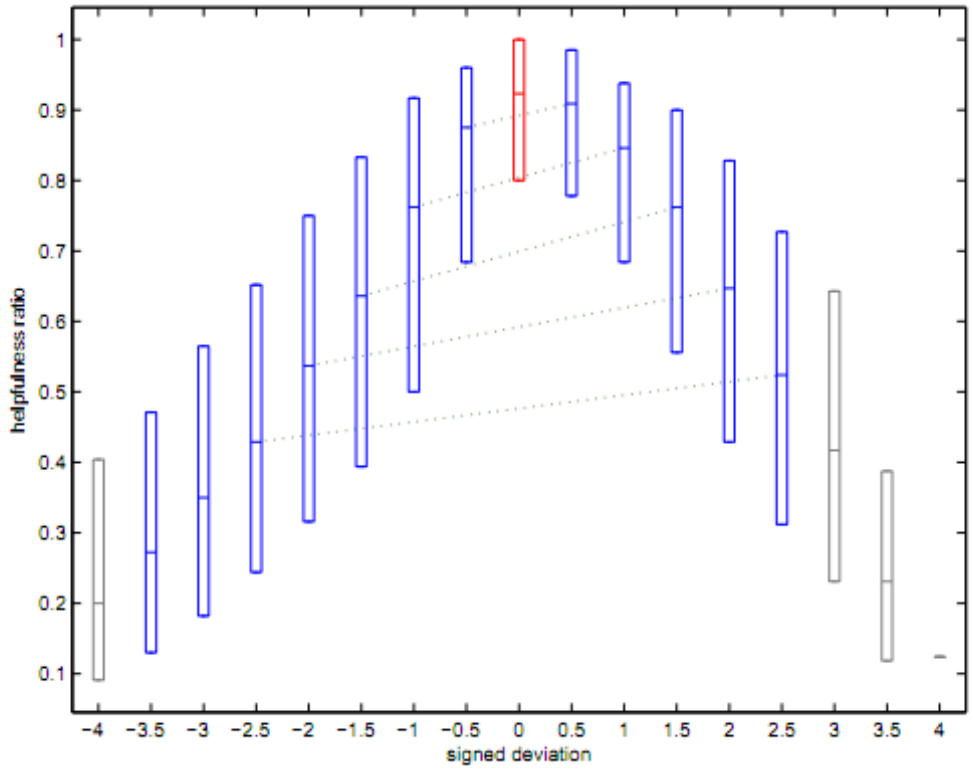
- Conforming reviews are more helpful.



Opinion mining

Rating reviews (cont.)

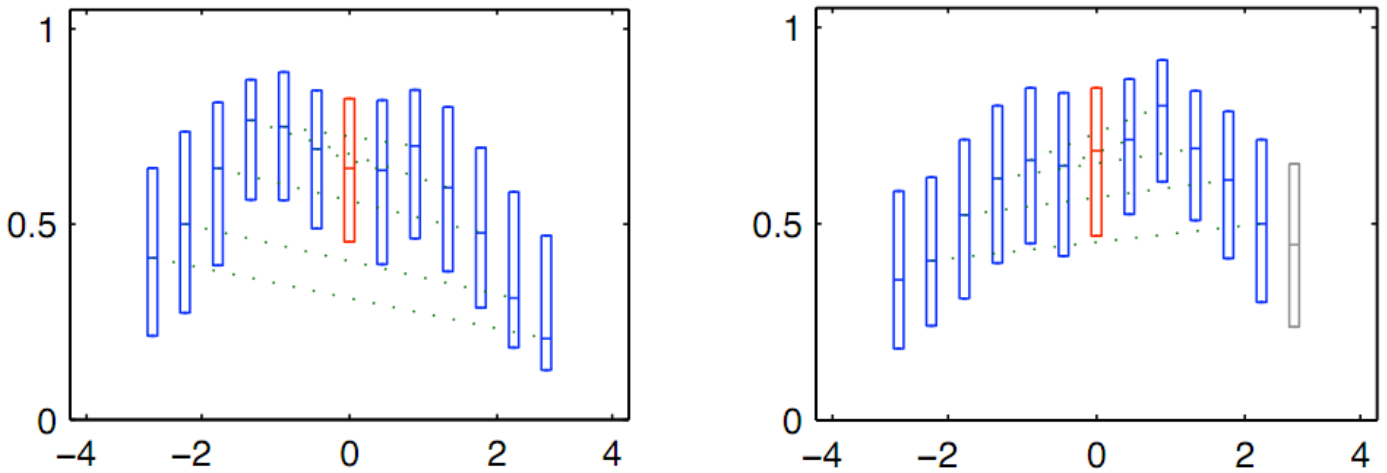
- signed deviation = star rating - average star rating;
- positive reviews are more helpful (*Brilliant-but-cruel* is not working).



Opinion mining

Cultural differences

- Signed deviations vs. helpfulness ratio, in the Japanese (left) and U.S. (right) data. The curve for Japan has a pronounced lean towards the left



Opinion mining

How is it in Russia?

- Yandex.Market

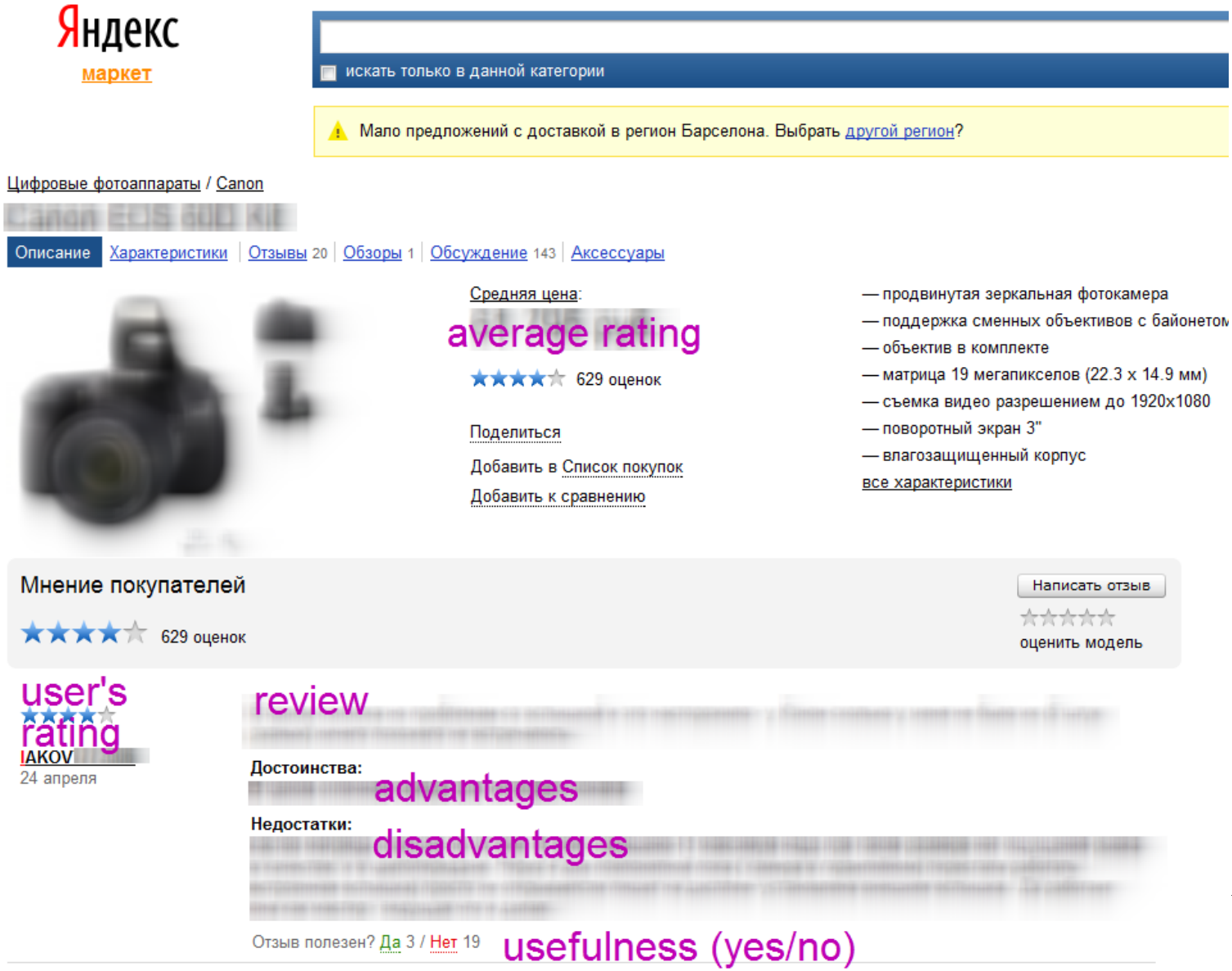
Practical task

Information

- Yandex.Market is the most successful site for reviews in RuNet by the number of reviews and by reviews' quality.
- Link: `bit.ly/russir2011`

Practical task

Yandex Market snapshot



Practical task

Usefulness

- Yandex reviews usefulness:

$$\frac{useful + 1}{numvotes + 2} - \frac{1}{2 * (numvotes + 2)}$$

  - **useful** is the number of votes that rate review as useful;
  - **numvotes** is the number of all votes that rate reviews usefulness;
- The main point: **usefulness = share of useful - error.**
- **Error** is a half of confidence interval;

RuSSIR
Russian Summer School in Information Retrieval

Practical task

Learning set

Yandex.Market data set:

file:reviews.xml Reviews and usefulness for items (digital cameras):

file:modeldata.csv Average rating and usefulness of items;

file:categorydata.csv Average rating and average usefulness of the product items for the selected category (digital cameras);

file:userdata.csv Average usefulness of reviews and the number of the accepted reviews done by an author;

Practical task

Files (1)

file:reviews.xml Reviews and usefulness for items (digital cameras);

- **ID** review id;
- **MODEL ID** item id;
- **AUTHOR D** author id;
- **CR TIME** writing time of the review;
- **RATING** rating of the model by the author of the review (from 1 to 5 (best));
- **TEXT** text of the review;
- **PRO** text about advantages of the model;
- **CONTRA** text about disadvantages of the model;
- **RANK** evaluation by other users of the review usefulness (from 0 to 1 (best)).

RuSSIR
Russian Summer School in Information Retrieval

Practical task

Files (2)

file:modeldata.csv Average rating and usefulness of items;

- **MODEL ID** item id;
- **AVG RANK** average usefulness of items' reviews;
- **RATING** average item rating (from 1 to 5 (best));

Practical task

Files (3)

file:categorydata.csv Average rating and average usefulness of the product items for the selected category (digital cameras);

- **CATEGORY AVG RATING** average rating of the product items for the selected category;
- **CATEGORY AVG RANK** average usefulness of the product items for the selected category.

Practical task

Files (4)

file:userdata.csv Average usefulness of reviews and the number of the accepted reviews done by an author;

- **AUTHOR ID** author id;
- **NUM REVIEWS** the number of accepted reviews done by the author;
- **AVG RANK** average usefulness of the reviews done by the author (from 0 to 1 (best));

Practical task

Tasks (1)

task 1 Given

- text and rating of the item's review,
- average rating of the item,
- average usefulness for the item and for the category,
- average usefulness of the user,
- number of reviews from the user,

to predict

- usefulness of the review by other users.

Practical task

Tasks (2)

task 2 Given

- text of the item's review,
- average rating of the item,
- average usefulness of the user's reviews,
- number of reviews from the user,

to predict

- the user's rating of the item.

Practical task

Links

- **Weka:** www.cs.waikato.ac.nz/ml/weka/
- **LingPipe:** http://alias-i.com/lingpipe/demos/tutorial/logistic-regression/read-me.html
- **Shark:** http://shark-project.sourceforge.net/Tutorials.html
- **Shogun:** http://www.shogun-toolbox.org/

RuSSIR
Russian Summer School in Information Retrieval

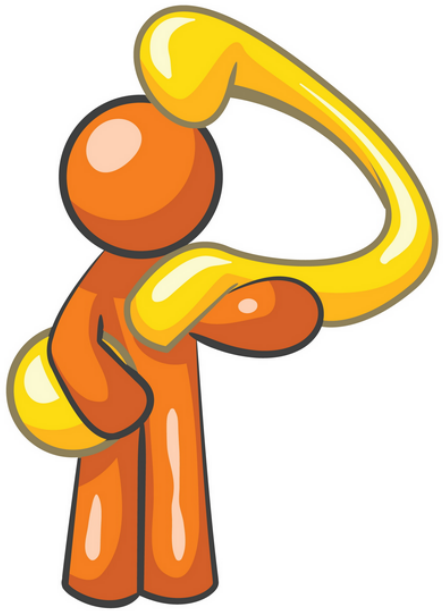
Practical task

Gameplan

- deadline: evening Wed, 17th
- results: Thr, 18th
- winners present their ideas (10 min): Fri, 19th

Practical task

Questions



Bibliography I

L. A. Adamic, D. Lauterbach, C. Y. Teng, and M. S. Ackerman. Rating friends without making enemies,. 2011.

M. Castells. *Communication power*. Oxford University Press, USA, 2009.

C. Danescu, G. Kossinets, J. Kleinberg, and L. Lee. How opinions are received by online communities: a case study on amazon.com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 141–150, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4.

B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

RuSSIR
Russian Summer School in Information Retrieval