Team Project FSS 2018

Mining Product Data from the Web
Hallo

- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- Research Interests:
  - Web Data Integration
  - Data and Web Mining
  - Linked Data Technologies
- Room: B6 - B1.15
- eMail: chris@informatik.uni-mannheim.de
- Consultation: Wednesday, 13:30-14:30
Hallo

- Anna Primpeli
- Graduate Research Associate
- Research Interests:
  - Data Extraction
  - Web Data Integration
  - Active Learning
  - Structured Data on the Web
- Room: B6, 26, C 1.04
- eMail: anna@informatik.uni-mannheim.de
Agenda of Today‘s Kickoff Meeting

1. Introduction
2. Organization and Schedule
3. Specific Subtasks
Motivation of the Team Project

The Web is a rich source of product information

- the same product is described by 100s of websites
  - by merchants, the producer, consumers
- different websites describe different aspects of a product
  - technical spec vs consumer experience
- there are plenty of offers for a product online
  - we can collect price information on global scale
- many websites point us at similar products

Using information about products from the Web, we can

- build comprehensive product catalogues and search engines
- conduct global price comparison engines
- understand consumer behavior and market structure
Project Goals

1. Gather and integrate product, price, and review data from multiple e-shops

2. Mine this data to discover
   • price/feature associations
   • feature/user perception associations
   • understand the market structure
   • understand consumer behaviour
Questions and Subtasks

1. Which e-shops to consider? → **Data Selection and Crawling**
2. Which data to extract? → **Feature Extraction**
3. How to recognize identical products? → **Identity Resolution**
4. How to group similar products? → **Categorization / Cluster Analysis**
5. How to understand user perception? → **Sentiment Analysis**
6. How to combine extracted information? → **Data Fusion**
7. What patterns can be found in the data? → **Data Mining**
How Do Data Scientists Spend Their Days?

- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Building training sets: 3%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Project Organization

Duration: 6 months (02.03.2018 – 02.09.2018)
ECTS: 12
Participants: 8 people
Type of work: Team and subgroup based
Milestones: 4 project phases
Evaluation:
- Individual contribution to the deliverables
- Deliverables: Presentations, reports, code, data
- Every project phase determines 25% of your final grade
The Project Team

1. Heißler, Larissa
2. Bertsch, Matthias Helmuth
3. Demirxhiu, Ersejda
4. Leung, Chung Chi
5. Chowdhury, Abdullah Al Murad
6. Koseoglu, Bengi
7. Aghazada, Adila
8. Gjuzi, Anjeza

- A Short Round of Introductions
  - What are you studying? Which semester?
  - Which DWS courses did you already attend?
  - What are your programming and data wrangling skills?
Main Steps of the Project

1. Source Selection
2. Data Collection
   - Feature Extraction
   - Identity Resolution
   - Product Classification
   - Sentiment Analysis
3. Price/Feature Mining
4. Perception/Feature Mining
## Detailed Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friday, 02.03.2018</td>
<td>Kickoff meeting (today)</td>
</tr>
<tr>
<td></td>
<td>Phase 1 (all members): Decide on a set of products and sources, crawling, basic feature extraction, product catalog construction</td>
</tr>
<tr>
<td>Friday, 09.03.2018</td>
<td>Meet Anna and report plan, division of work</td>
</tr>
<tr>
<td>Friday, 23.03.2018</td>
<td>Drop-out deadline: Dropping out after this date will result in failing the team project</td>
</tr>
<tr>
<td>Friday, 13.04.2018</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; Deliverable: 15 minutes presentation, code &amp; data - Subgroup formation</td>
</tr>
<tr>
<td></td>
<td>Phase 2 (in 4 subgroups): feature extraction, identity resolution, categorization, sentiment analysis</td>
</tr>
<tr>
<td>Friday, 20.04.2018</td>
<td>Meet Anna and report plans</td>
</tr>
<tr>
<td>Friday, 18.05.2018</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; Deliverable: 10 minutes presentation from each subgroup, code &amp; data</td>
</tr>
<tr>
<td></td>
<td>Phase 3 (in 4 subgroups): Refinement of phase 2</td>
</tr>
<tr>
<td>Tuesday, 29.05.2018</td>
<td>Meet Anna and report plan</td>
</tr>
<tr>
<td>Sunday, 01.07.2018</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt; Deliverable: 8-12 pages report from each subgroup, code &amp; data</td>
</tr>
<tr>
<td></td>
<td>Phase 4 (in 2 subgroups): Mining of integrated product data and reviews</td>
</tr>
<tr>
<td>Friday, 20.07.2018</td>
<td>Meet Anna and report plan</td>
</tr>
<tr>
<td>Sunday, 02.09.2018</td>
<td>4&lt;sup&gt;th&lt;/sup&gt; Deliverable: 8-12 pages report from each subgroup, code &amp; data</td>
</tr>
<tr>
<td>Friday, 07.09.2018</td>
<td>Overall presentation 30 min + Feedback</td>
</tr>
</tbody>
</table>
Phase 1: Source Selection, Crawling, Basic Feature Extraction, Catalog Construction

**Participants:** All team members

**Duration:** 02.03.2018 – 03.04.2018

**Deliverables:** 15 minutes presentation, data & code, report who did what

**Tasks (2/5)**

1. Decide on two main product categories
   - Select 2 non-similar main product categories, e.g. laptops and shoes (**NOT** phones, headphones, TVs)
   - Choose 3 – 4 subcategories for each main category, e.g. noise-cancelling, over-ear, on-ear, & sports headphones.

2. Decide on a set of e-shops
   - Analyse data for main players for each product category using sources such as Alexa and WDC
   - Select minimum 20 e-shops for each main product category
   - The selected e-shops should: Be located in 2 countries, be in English, **NOT** be marketplaces.
   - Partly (50%+) support the extraction of structured data by using schema.org and/or HTML tables
Phase 1: Source Selection, Crawling, Basic Feature Extraction, Catalog Construction

Tasks (3/5)

3. Crawl product pages
   - For each main category select 50 seed products. The selected products should not be too distinct from each other, e.g. i-phone 4s, samsung galaxy s8, nokia lumia 635.
   - Crawl products by following links from the seed product pages. Crawl 1000+ products per website.
   - Your crawled results should include closely related products, e.g. i-phone 4 and i-phone 4s.

4. Extract product specifications, prices, category information, product IDs, and reviews
   - Consider the schema.org annotations s:Product, s:Review, and s:Offer
   - Use simple heuristics for locating relevant data: use annotations and identify web tables.

5. Construct a product catalog
   - Consider google shopping to create a product catalog the covering 50 seed products per category
     - The catalog defines a central schema for describing and a single product hierarchy.
     - Perform basic schema matching of product specifications to product attributes in catalog.
## schema.org Terms

### Review
- **Canonical URL:** http://schema.org/Review
- **Thing > CreativeWork > Review**
- A review of an item – for example, of a restaurant, movie, or store.
- **Usage:** Between 250,000 and 500,000 domains

<table>
<thead>
<tr>
<th>Property from Review</th>
<th>Expected Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>itemReviewed</td>
<td>Thing</td>
<td>The item that is being reviewed.</td>
</tr>
<tr>
<td>reviewAspect</td>
<td>Text</td>
<td>This Review or Rating is related to.</td>
</tr>
<tr>
<td>reviewBody</td>
<td>Text</td>
<td>The actual body of the review.</td>
</tr>
<tr>
<td>reviewRating</td>
<td>Rating</td>
<td>The rating given in this review, rating given by the review.</td>
</tr>
</tbody>
</table>

### Product
- **Canonical URL:** http://schema.org/Product
- **Thing > Product**
- Any offered product or service. For example: a pair of shoes, a concert ticket; the rental of a car; a haircut; or an episode of a TV show streamed online.
- **Usage:** Over 1,000,000 domains

<table>
<thead>
<tr>
<th>Property from Product</th>
<th>Expected Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>additionalProperty</td>
<td>PropertyValue</td>
<td>A property-value pair representing an additional characteristic of the entity, e.g. a product feature or another characteristic for which there is no matching property in schema.org. Note: Publishers should be aware that applications designed to use specific schema.org properties (e.g. <a href="http://schema.org/width">http://schema.org/width</a>, <a href="http://schema.org/color">http://schema.org/color</a>, <a href="http://schema.org/gnm113">http://schema.org/gnm113</a>, ...) will typically expect such data to be provided using those properties, rather than using the generic property/value mechanism.</td>
</tr>
<tr>
<td>aggregateRating</td>
<td>AggregateRating</td>
<td>The overall rating, based on a collection of reviews or ratings, of the item. An intended audience, i.e. a group for whom something was created. Supersedes serviceAudience.</td>
</tr>
<tr>
<td>brand</td>
<td>Text</td>
<td>An award won by or for this item. Supersedes awards.</td>
</tr>
<tr>
<td>category</td>
<td>PhysicalActivityCategory</td>
<td>A category for the item. Greater signs or slashes can be used to informally indicate a category hierarchy.</td>
</tr>
<tr>
<td>color</td>
<td>Text</td>
<td>The color of the product.</td>
</tr>
<tr>
<td>depth</td>
<td>Text</td>
<td>The depth of the item.</td>
</tr>
</tbody>
</table>

---

schema.org Annotations: Most Popular Classes

Development of Selected Classes by #PLDs

http://webdatacommons.org/structureddata/
## Adoption by E-Commerce Websites 2014

### Distribution by Top-Level Domain

<table>
<thead>
<tr>
<th>TLD</th>
<th>#PLDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>com</td>
<td>38344</td>
</tr>
<tr>
<td>co.uk</td>
<td>3605</td>
</tr>
<tr>
<td>net</td>
<td>1813</td>
</tr>
<tr>
<td>de</td>
<td>1333</td>
</tr>
<tr>
<td>pl</td>
<td>1273</td>
</tr>
<tr>
<td>com.br</td>
<td>1194</td>
</tr>
<tr>
<td>ru</td>
<td>1165</td>
</tr>
<tr>
<td>com.au</td>
<td>1062</td>
</tr>
<tr>
<td>nl</td>
<td>1002</td>
</tr>
</tbody>
</table>

Adoption by Top-15: 60 %

### Alexa Top-15 Shopping Sites

<table>
<thead>
<tr>
<th>Website</th>
<th>schema:Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon.com</td>
<td>✗</td>
</tr>
<tr>
<td>Ebay.com</td>
<td>✓</td>
</tr>
<tr>
<td>NetFlix.com</td>
<td>✗</td>
</tr>
<tr>
<td>Amazon.co.uk</td>
<td>✗</td>
</tr>
<tr>
<td>Walmart.com</td>
<td>✓</td>
</tr>
<tr>
<td>etsy.com</td>
<td>✗</td>
</tr>
<tr>
<td>Ikea.com</td>
<td>✓</td>
</tr>
<tr>
<td>Bestbuy.com</td>
<td>✓</td>
</tr>
<tr>
<td>Home depot.com</td>
<td>✓</td>
</tr>
<tr>
<td>Target.com</td>
<td>✓</td>
</tr>
<tr>
<td>Groupon.com</td>
<td>✗</td>
</tr>
<tr>
<td>Newegg.com</td>
<td>✓</td>
</tr>
<tr>
<td>Lowes.com</td>
<td>✗</td>
</tr>
<tr>
<td>Macys.com</td>
<td>✓</td>
</tr>
<tr>
<td>Nordstrom.com</td>
<td>✓</td>
</tr>
</tbody>
</table>
# Properties used to Describe Products 2014

<table>
<thead>
<tr>
<th>Top 15 Properties</th>
<th>PLDs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
</tr>
<tr>
<td>schema:Product/name</td>
<td>78,292</td>
</tr>
<tr>
<td>schema:Product/image</td>
<td>59,445</td>
</tr>
<tr>
<td>schema:Product/description</td>
<td>58,228</td>
</tr>
<tr>
<td>schema:Product/offers</td>
<td>57,633</td>
</tr>
<tr>
<td>schema:Offer/price</td>
<td>54,290</td>
</tr>
<tr>
<td>schema:Offer/availability</td>
<td>36,789</td>
</tr>
<tr>
<td>schema:Offer/priceCurrency</td>
<td>30,610</td>
</tr>
<tr>
<td>schema:Product/url</td>
<td>23,723</td>
</tr>
<tr>
<td>schema:Product/aggregateRating</td>
<td>21,166</td>
</tr>
<tr>
<td>schema:AggregateRating/ratingValue</td>
<td>20,513</td>
</tr>
<tr>
<td>schema:AggregateRating/reviewCount</td>
<td>14,930</td>
</tr>
<tr>
<td>schema:Product/manufacturer</td>
<td>10,150</td>
</tr>
<tr>
<td>schema:Product/brand</td>
<td>9,739</td>
</tr>
<tr>
<td>schema:Product/productID</td>
<td>9,221</td>
</tr>
<tr>
<td>schema:Product/sku</td>
<td>7955</td>
</tr>
</tbody>
</table>
Stuff that will help you later …

1. Product IDs
   - GTINs, MPNs, SKUs, ISBNs
   - solve the identity resolution problem

2. HTML Tables
   - contain key/value pairs
   - main source of structured product specifications

3. Category Information
   - (Schema.org) bread crumbs
   - Schema.org category
   - URLs fragments

- Make sure that 50%+ of the selected shops provide these things
Main Steps of the Project

- Source Selection
  - Data Collection
    - Feature Extraction
    - Identity Resolution
    - Product Classification
    - Sentiment Analysis
    - Price/Feature Mining
    - Perception/Feature Mining
Phases 2 & 3: Feature Extraction, Identity Resolution, Categorization, Sentiment Analysis

Duration: 04.04.2018 – 15.05.2018 (Phase 2), 16.05.2018 – 01.07.2018 (Phase 3)

Participants: 4 subgroups of 2 persons each

Deliverables:
- 10 minutes presentation from each subgroup after phase 2
- 8 -12 pages report, data & code from each subgroup after phase 3

Tasks (2/4)

Subgroup 1: Feature Extraction
- Perform more sophisticated feature extraction techniques such as application of regex expressions, further consideration of table structure
- perform more sophisticated schema matching with product catalog

Subgroup 2: Identity Resolution
- Apply basic IR techniques and machine learning, exploit features produced by subgroup 1 in phase 2
- Apply transfer learning: use product identifiers as labeled data to learn a classification model and evaluate the model on unseen data
Tasks (4/4)

Subgroup 3: Product Categorization

- Apply multi-level classification techniques
- Consider product features for categorization
- Exploit existing categorization information / integrate category trees
- Exploit identity resolution results from team 2 in second phase.
- Cluster products into additional more fine grained categories (premium products within subcategory?)

Subgroup 4: Sentiment Analysis

- Gather and preprocess the reviews of the crawled websites
- Extract sentiment information for the specific product features and overall for the product
- Extract background information about raters (crawl additional review portals if necessary)

Phase 3 will be a refinement of Phase 2!
Phase 4: Data Mining

Participants: 2 subgroups of 4 persons each

Duration: 02.07.2018 – 01.09.2018

Deliverables: report, code & data from each subgroup, final overall presentation

Tasks

Bring your results together and extract interesting facts

- Combine the information extracted from the subtasks of phase 2 and 3
- Mine your data by performing correlation analysis
- Discover and report interesting facts concerning

  Subgroup 1: **Price**
  
  e.g. Which features determine the price? How are prices distributed? By location? Product type?

  Subgroup 2: **Reviews**
  
  e.g. Which specific customer groups prefer which sub-categories/products?, which features matter most for specific customer groups?, how do certain product features affect customer satisfaction?
Formal Requirements
Formal Requirements & Consultation

- Deliverables
  - Reports should be 8-12 pages single column
    - including appendixes
    - not including the bibliography
    - every additional page reduces your grade by 0.3
    - Created with Latex template of the Data and Web Science group
      (http://dws.informatik.uni-mannheim.de/en/thesis/masterthesis/)
    - Every deliverable (presentation/report) should be accompanied with an excel sheet stating which team member conducted which subtask.

- Final grade
  - 25% for every phase, individual grade / not per team
  - Late submission: reduction of grade by 0.3 per day

- Consultation
  - Send one e-mail per team stating your questions to Anna, she answers questions or meets with you
  - Chris does second level support and gives feedback at presentations
How to structure your deliverables?

1. Problem definition
2. Profiling of selected data
3. Methodology
4. Evaluation (for phases 2 & 3)
5. Error Analysis (for phases 2 & 3)
6. Conclusion

Accompany your deliverables with the code and data you used
  ! The phase deadlines apply for the submission of your code and data as well
Submission of Deliverables

Presentations
The presentations will take place according to the schedule. For the exact time you will be informed via e-mail. The presentation slides should be provided by the end of the meeting.

Team and Subgroup Reports
Send one e-mail per team or subteam until the deadline date according to the schedule

Data and Code
Add your data and code in a zipped folder and send (URL) via e-mail

Member to subtask report
Send one excel sheet per team explaining who did what together with the deliverables.

All deliverables should be sent to Chris & Anna!
Potentially Useful Software

- Crawling
  - Scrapy: [https://scrapy.org/](https://scrapy.org/)
  - Any23

- Data Integration
  - Winte.r Framework: [https://github.com/olehmberg/winter](https://github.com/olehmberg/winter)
  - Silk Framework: [https://github.com/silk-framework/silk](https://github.com/silk-framework/silk)

- Data Mining, Machine Learning
  - RapidMiner: [https://rapidminer.com/](https://rapidminer.com/)

- Natural Language Processing
  - Stanford NLP: [https://nlp.stanford.edu/software/](https://nlp.stanford.edu/software/)
Related Work (1/3)


Related Work (3/3)


Learning Targets

Improve your technical skills

- Work as a **Data Scientist**: gather, clean, profile, integrate, classify data in order to extract knowledge
- Understand the nature of **Web Data**
- Improve your technical expertise / programming skills

Improve your soft skills

- Work as part of a bigger team on a more complex project
- Organize yourself and assign tasks based on your skills
- Communicate and coordinate your work
Questions?
Project Infrastructure?

- **Shared Document Space**
  - for todo lists, brainstorming documents
  - Google Docs? Wiki?

- **ILIAS Group**
  - mailing to all participants
  - for sharing files

- **Code Repository**
  - GitHub?

- **Data Repository**
  - Google Drive? Dropbox?

- **Anything else?**