Team Project FSS 2018

Mining Product Data from the Web
Phase 2
Progress and Focus of these Phases

✓ 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**

Phase 3 will be a refinement of phase 2
Main Steps of the Project

Source Selection

Data Collection

Feature Extraction
Identity Resolution
Product Classification
Sentiment Analysis

Price/Feature Mining
Perception/Feature Mining
### Detailed Schedule for Phase 2

<table>
<thead>
<tr>
<th>Date</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friday, 13.04.2018, 9:15am</td>
<td>Introduction to Phase 2, subgroup formation</td>
</tr>
<tr>
<td>Friday, 20.04.2018, 8:30am</td>
<td>Meet Anna and discuss plans</td>
</tr>
<tr>
<td>Friday, 27.04.2018, 9:15am</td>
<td>Meet Chris and Anna, report profiling results and specific goals for phase 2</td>
</tr>
<tr>
<td>Friday, 18.05.2018, 9:15am</td>
<td><strong>2nd Deliverable:</strong> 10 minutes presentation from each subgroup, code &amp; data</td>
</tr>
</tbody>
</table>
Results from Phase 1

What should you have from Phase 1?

✓ Crawled corpus of Bag and Camera products
  - Min. 20 e-shops located in 2 countries
  - Estimated high product overlap
  - Different subcategories
  - Pages rich in annotations and specification tables
  - Pages describing different products with similar attributes ex: Polaroid vs Polaroid Kit

✓ Basic Feature Extraction and Profiling
  - Identify and profile product specifications, prices, category information, product IDs, and reviews

✓ Product Catalogs
  - Min 50 bag products
  - Min 50 camera products
Phase 2 – Subgroup 1: Feature Extraction

Goal: Extract clean feature - value pairs from the product pages and perform schema matching

How?

1. Identify where features are located: tables, lists, free text following certain patterns, schema.org annotations

2. Extract product features and map them to the catalog
   - Generic Approach: Consider table and list structure, schema.org annotations and DOM structure
     1. Create a gold standard for schema matching
     2. Perform schema matching
        - Label based
        - Instance based
   - Catalog-oriented Approach: Apply regex expressions exploiting the knowledge in the catalog

3. Compare the two approaches
Phase 2 – Subgroup 2: Identity Resolution

Goal: Match entities between your product corpus and the product catalog

How?
1. Create your gold standard
   - Annotate manually min. 100 product pairs (product page – catalog entry)
   - Make sure you include good negative examples!
2. Consider Bag of Words models from different parts of the product page, e.g. Tables, annotations, free text
3. Preprocess the data
4. Apply basic IR techniques
5. Apply machine learning methods
   - Learn one model for all product pairs
   - Learn multiple models – one for each product (you need a big gold standard)
   - Feature vectors: tokens (binary or tfidf weights), similarity scores computed with different measures
6. Evaluate and if necessary refine your gold standard
Phase 2 – Subgroup 3: Categorization

Goal: Learn a model to assign the correct category to every product

How?

1. Define an initial hierarchy of product categories
2. Select a hierarchical classification method [1]
3. Create a gold standard considering the requirements of your method
   • Min. 200 annotated products in the form of <product_a : action camera>
4. Use product features and apply hierarchical classification
   • Your features should be simply induced, e.g. Bag of Words model.
5. Evaluate and if necessary refine your gold standard

Phase 2 – Subgroup 4: Sentiment Analysis

Goal: Perform aspect based sentiment analysis on product reviews

How?

1. Profile the review information and if necessary crawl more reviews from the product pages
   - How many reviews?
   - Are there predefined aspects/ information about reviewers?

2. Identify reviewed features and subfeatures
   - e.g. „The display is of great quality“ VS „Although the display screen is big, I am not satisfied with its resolution“

3. Extract the sentiment for each feature and overall for the product
   - Usage of linguistic patterns, e.g. Adjective + Noun : This bag is of great material
   - Consideration of negation and degree words
   - Usage of polarity dictionaries

4. Evaluate your scoring against the extracted score
Phase 2 Results & Deliverable

**Duration:** 13.04.2018 – 15.05.2018

**Deliverables:**

1. **A 10 min presentation from each subgroup**
   The presentation slides should be provided by the end of the meeting.

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

3. **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/)
  - RiTa library: [http://rednoise.org/rita/download.php](http://rednoise.org/rita/download.php)
Related Work for Feature Extraction


Related Work for Identity Resolution


Related Work for Categorization


Related Work for Sentiment Analysis


Questions?
Subtask Assignment

- Subgroup 1: Feature Extraction
  - Members: Chung, Adela
  - Timeslot for 27.04: 9:15 – 9:35

- Subgroup 2: Identity Resolution
  - Members: Larissa, Ersejda
  - Timeslot for 27.04: 9:35 – 9:55

- Subgroup 3: Categorization
  - Members: Bengi, Anjeza
  - Timeslot for 27.04: 10:15 – 10:35

- Subgroup 4: Sentiment Analysis
  - Members: Matthias, Murad
  - Timeslot for 27.04: 10:35 – 10:55