Team Project FSS 2018

Mining Product Data from the Web
Phase 3
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 is 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 3

<table>
<thead>
<tr>
<th>Date</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friday, 18.05.2018, 9:15am</td>
<td>Introduction to Phase 3</td>
</tr>
<tr>
<td>Tuesday, 29.05.2018, 8:30am</td>
<td>Meet Anna and discuss plans</td>
</tr>
<tr>
<td>Sunday, 01.07.2018</td>
<td>3\textsuperscript{rd} Deliverable: 8-12 pages report from each subgroup, code &amp; data</td>
</tr>
</tbody>
</table>


Results from Phase 2

What should you have from Phase 2?

Feature Extraction team
- Dictionary-based feature extraction method
- A baseline generic approach
- Gold standard
- Evaluation of your methods
- Set of extracted features per page and profiling statistics about it

Identity Resolution team
- Baseline BoW identity resolution approach
- Machine learning based BoW identity resolution approach
- Gold standard
- Experiments with different sets of features (e.g. schema.org annotations, all html)
- Evaluation of your methods
- Set of correspondences and profiling statistics about it
Results from Phase 2

What should you have from Phase 2?

Product Classification team
- Taxonomy of product categories for bags and cameras
- Baseline hierarchical classifier
- Experiments with different sets of features
- Gold standard
- Evaluation of your method
- Profiling statistics about the categories of the pages of your corpus

Sentiment Analysis team
- Gather bigger review corpus and profiling statistics about it
- Define your aspects for sentiment analysis (think about future mining e.g. include price)
- Baseline method for extracting polarity of aspects in reviews
- Evaluate baseline method
Phase 3 – Subgroup 1: Feature Extraction

Goal: Extract clean feature-value pairs from the product pages and perform schema matching (refinement and wider set of attributes)

How?

1. Perform error analysis on your current results
   • What went wrong in the feature extraction?
   • What went wrong in the schema matching?

2. Enhance your feature extraction techniques
   • Extract tables and classify them into specification/non-specification ones. Consider [1,2,3]

3. Enhance your schema matching techniques
   • Use the results of the identity resolution team (or their gold standard) for duplicate-based schema matching

Phase 3 – Subgroup 2: Identity Resolution

Goal: Match entities between your product corpus and the product catalog
(using a wider set of attributes and additional webpages to learn product recognizers)

How?
1. Use the results of the feature extraction subteam for phase 2 and test more expressive learning algorithms
   1. linear regression on attribute similarity scores
   2. tree-based models on attribute similarity scores
2. Perform error analysis on best BoW- and Attribute-based models
3. Use product IDs (GTIN) to learn better product recognizers
   • Learn product recognizers (one binary classifier per product) based on the data in the catalog
   • Use products identifiers to find additional web pages describing the product (google the IDs).
   • Learn improved product recognizers using the catalog and the new pages as training data
   • Compare performance of base recognizers and improved recognizers: Find out how much the results can be improved using additional webpages that were found using product IDs (like GTIN).
Phase 3 – Subgroup 3: Categorization

Goal: Profile/Cluster categorization taxonomies from different websites in order to identify alternative categorization approaches.

How?

1. Improve hierarchical classification
   1. Use the results of the feature extraction subteam for phase 2 and update your features
   2. Apply more sophisticated hierarchical classification methods

2. Profile Categorization Taxonomies of different websites
   1. Extract alternative categorization taxonomies/paths from the shops using schema.org breadcrumb, schema.org product or offer category, or regular expressions.
   2. Profile the categorization taxonomies/paths
   3. Manually group the taxonomies/paths by categorization „idea“ / level of detail.

3. Automatically Cluster Categorization Taxonomies
   1. Apply taxonomy clustering method to automatically find grouping (using results from 3. as ground truth).
Goal: Perform aspect based sentiment analysis on product reviews (refinement)

How?

1. Perform error analysis
   • Identify the most difficult factors for extracting sentiments, i.e. degree words, sentence structure, negation

2. Use feature extraction methods or feature names/values from the Feature Extraction subteam to locate more easily the aspects in your textual reviews

3. Use supervised models to learn important polarity words for every aspect

4. Create a gold standard of the form <Review_id, Aspect_id, Polarity>
Phase 3 Results & Deliverable

**Duration:** 18.05.2018 – 01.07.2018

**Deliverables:**

1. **8 – 12 from each subgroup**
   - 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

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/
  - Any23

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

- Data Mining, Machine Learning
  - RapidMiner: https://rapidminer.com/

- Natural Language Processing
  - Stanford NLP: https://nlp.stanford.edu/software/
  - RiTa library: 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?