

TrustYou Cookbook Semantic Analysis

Version	0.9.1		
Date	09.08.2010		
Authors	Markus Reil	Email:	markusr@trustyou.net
		Phone:	
		Email:	
		Phone:	

Table of Contents

1.Introduction.....	3
2.Export Format.....	3
2.1.Entities.....	3
2.1.1.Example.....	4
2.2.Reviews.....	5
2.2.1.Example.....	6
3.Analysis and Matches.....	8
3.1.Example (entity.xml).....	8
4.Generation.....	10
4.1.Linking reviews to entities.....	10
4.2.Linking reviews to matches.....	10
4.3.Extended match context.....	11
4.4.Categories.....	13
4.5.Tops & Flops.....	14
5.Changes.....	16

1. Introduction

The purpose of this document is to help partners understand and analyse the results of the Semantic Analysis process.

During this process TrustYou analyses reviews and aggregates all the results and generates summaries for every entity.

This document describes how this aggregated information can be linked to the original review texts and how different data views and reports can be generated.

2. Export Format

The output of the Semantic Analysis consists of 2 separate output files:

- entities.xml contains the aggregated information for every entity
- reviews.xml contains the semantic information for all reviews.

2.1. Entities

Every entity is stored in an <entity> element and contains the following information (s. also entities.xsd):

Element	Description
entity	Container element for one entity
type	The entity type (s. Entity types)
id	The entity id as defined by the partner
trustyou_id	TrustYou's internal id, if available
name	Entity name
address	Entity address
module_url	TrustYou widget URL

synth_phrases	One sentence generated from the Semantic Analysis result, that describes, what people think about the hotel.
category_list	List of categories that this hotel was assigned.
matches	Result of the Semantic Analysis process
matchpos	All positive matches. freq: number of matches
matchneu	All neutral matches
matchneg	All negative matches
match	The details for every match object freq: number of occurrences in all reviews id: object id of the match
submatch	The details for every match freq: number of occurrences in all reviews aid: attribute id of the match id: object subid of the match
norms	Container for the normalised matches.
norm	A normalized match. freq: number of occurrences in all reviews lang: match language stage: normalisation stage (1=not normalised, 9=fully normalised)
acc	The accusative form of the normalised match (if available)
nom	The normalised nominative form of the match (stage="9") or the original match (stage="1")
neg	The negated form of the normalised match (if available)
translations	Container for match translations.
trans	A match translation. lang: the target language quality: quality of the translation

2.1.1. Example

```
<?xml version="1.0" encoding="UTF-8"?>
<root>
  <entity>
    <type>1</type>
    <id>80383</id>
    <trustyou_id>0077b0bb-9799-491b-b664-0223c546610e</trustyou_id>
    <name>Hotel Restaurant Ibis</name>
    <address>
      <city>Zurich</city>
      <country>Switzerland</country>
      <countrycode>CH</countrycode>
    </address>
  </entity>
</root>
```

```

    <street>Schiffbaustrasse 11</street>
    <zip>8005</zip>
</address>
<module_url>http://widgets.trustyou.com/widgetizer?id=0077b0bb-9799-491b-b664-0223c546610e&
    amp;module=1&lang=de&partnerid=abcde</module_url>
<matches>
  <matchpos freq="1">
    <match freq="1" id="15">
      <submatch freq="1" aid="p05" id="">
        <norms>
          <norm stage="9" lang="en" freq="1">
            <acc/>
            <nom>outstanding service</nom>
            <neg/>
          </norm>
          <norm freq="1" lang="de" stage="9">
            <acc>sehr freundliche und hilfsbereite Personal</acc>
            <nom>sehr freundliches und hilfsbereites Personal</nom>
            <neg/>
            <translations>
              <trans lang="en" quality="85">very friendly and helpful staff</trans>
            </translations>
          </norm>
        </norms>
      </submatch>
    </match>
  </matchpos>
</matches>
</entity>
</root>

```

2.2. Reviews

Every review is stored in a <review> element and contains the following information:

Element	Description
review	Container element for one review
entity_id	The entity the review belongs to (corresponds to an entity's id)
cluster_id	TrustYou's internal id, if available
author	The review's author

mark	The review's rating
text	The review text
analyse	The result of the Semantic Analysis lang: the review language
text	The review text len: number of words
tokens	All the tokens of a review. The index is used in the matches to indicate where a match occurred (see below).
matches	The matches found in the text total: the number of matches found
concordance	the grammatical context of the match
aunion/a	the matched attribute id : the attribute id
ounion/obj	id : the object id subid : the object subid
analyse_mask	Container for all analyse ids
analyse_id	Id that identifies the presence of a specific match.

2.2.1. Example

```
<?xml version="1.0" encoding="UTF-8"?>
<review>
  <uid>ccc544096f742eb03fd9dc8e7e30ef4e_85320_17</uid>
  <entity_id>80383</entity_id>
  <cluster_id>0077b0bb-9799-491b-b664-0223c546610e</cluster_id>
  <source>10</source>
  <author>Corinna</author>
  <mark>
    <total>55</total>
  </mark>
  <date>1254348000</date>
  <address>
    <url>http://www.holidaycheck.de/hotel-Reiseinformationen_Hotel+Ibis+Zuerich+City+West-
hid_80383.html</url>
    <city>Zurich</city>
    <country>Switzerland</country>
  </address>
  <text>This hotel is supposed to be new, I seems renovations weren't thought through, since the
hotel seems like someone just re-painted it and left an older hotel underneath. Staff was Friendly
and the hotel was clean enough, what really is disappointing is the hotel's location, near nothing.
Location is dreadful. Front desk was very helpful.</text>
  <analyse lang="en">
    <text len="340"/>
  </analyse>
</review>
```

```
<tokens total="124">
  <tok index="0">This</tok>
  <tok index="1"/>
  <tok index="2">hotel</tok>
  <tok index="3"/>
  <tok index="4">is</tok>
  <tok index="5"/>
  ...
  <tok index="119">very</tok>
  <tok index="120"/>
  <tok index="121">helpful</tok>
  <tok index="122">.</tok>
  <tok index="123"/>
</tokens>
<matches total="4">
  <concordance end="67" index="0" start="63">
    <context>... . <span class="match">Staff was Friendly</span>    and ...</context>
    <match v="R0710"><ounion><obj id="15" subid="">Staff</obj></ounion>    was<aunion><a
id="p15" type="p"> Friendly</a></aunion></match>
    </concordance>
  </matches>
</analyse>
<analyse_mask>
  <analyse_id>15|p15|p</analyse_id>
  <analyse_id>15|p</analyse_id>
</analyse_mask>
</review>
```

3. Analysis and Matches

The Semantic Analysis process recognises text fragments in the reviews that each represent one of the author's opinions about an entity. Each of the recognised fragments, so-called matches, consists of an object and an attribute.

For example, if a reviewer mentions "friendly staff" the process identifies the object "staff" (object id 15) and the attribute "friendly" (attribute id p15).

The same object id / attribute id combination would be found if the reviewer mentioned "the nice staff" or if a German author was happy about "das freundliche Personal".

If more specific object descriptions are found in the match, the object id will get a suffix, a so-called sub-id. E.g. an opinion about the waiter or the management would result in a match with object id 15 (same as "service"), but with the additional sub-ids "c" and "g" respectively.

Additionally, every match has a type, which is one of "p", "n" and "o", and which identifies the match as positive, negative or neutral.

3.1. Example (entity.xml)

Let's have a look at an entity and see what information about a match we can retrieve from the XML:

```
<matches>
  <matchpos freq="15">
    <match freq="4" id="20">
      <submatch freq="2" aid="p07" id="a">
        <noms>
          <nom freq="2" lang="en" stage="1">
            <acc/>
            <nom>The pool was nice</nom>
            <neg/>
          </nom>
        </noms>
      </submatch>
    </match>
  </matchpos>
</matches>
```

```
</norms>
```

```
...
```

We can see, that at least one reviewer thought, that "the pool was nice".

The object "pool" is assigned the object id "20". This **is** tells us that the match is about the "Grounds". Since the pool is part of the grounds, the id is used here, too. Being a more specific description than "grounds", the object "pool" gets an additional sub-id, "a". The attribute id "p07" is assigned for the word "nice".

4. Generation

4.1. Linking reviews to entities

We can find all reviews belonging to one entity via their `<entity_id>` and `<id>` elements respectively.

In the example the entity contains the element `<id>80383</id>` and the review that belongs to it contains `<entity_id>80383</entity_id>`.

4.2. Linking reviews to matches

In order to create the `<entity>` element an aggregation of the semantical data of multiple reviews are being aggregated. Therefore it is not possible to directly recognise which reviews are responsible for a specific match in the entity.

But it still possible to identify those reviews via the so-called "analyse id".

This id has the format **has the format:**

ObjectId | ObjectSubId | AttributeId | Type

Let's take a look at an example match and build its "analyse id":

```
<matchpos freq="1">
  <match freq="1" id="15">
    <submatch freq="1" aid="p15" id="">
      <noms>
        <norm stage="9" lang="en" freq="1">
          <acc/>
          <nom>friendly staff</nom>
          <neg/>
        </norm>
        <norm freq="1" lang="de" stage="9">
          <acc>sehr freundliche und hilfsbereite Personal</acc>
          <nom>sehr freundliches und hilfsbereites Personal</nom>
          <neg/>
        </norm>
      </noms>
    </submatch>
  </match>
</matchpos>
```

```
<translations>
  <trans lang="en" quality="85">very friendly and helpful staff</trans>
</translations>
</norm>
...
```

This match consists of the object id "15" (`<match freq="1" id="15">`). Its object sub-id is empty (`<submatch freq="1" aid="p15" id="">`), and its attribute id is "p15" (`<submatch freq="1" aid="p15" id="">`). Its type is "p", because it is a positive match (it occurs in the `<matchpos>` element and its attribute id begins with "p"). So the analyse id for this match is

15||p15|p

We can now search all the entity's reviews for an `<analyse_id>` element containing this id.

Let's take a look at the `<analyse_mask>` of the example review:

```
<analyse_mask>
  <analyse_id>15|p</analyse_id>
  <analyse_id>15||p15|p</analyse_id>
</analyse_mask>
```

This mask contains the analyse id 15||p15|p and does indeed contain the match we are looking for ("friendly staff").

4.3. Extended match context

In a review, only the part of the text that was matched during the Semantic Analysis process plus an additional word to the left and to the right of the match, is included in the concordance `<context>` element:

```
<concordance end="76" index="2" start="68">
  <context>... The <span class="match">Hotel is near the airport</span> and ...</context>
  <match>
```

```
<ounion><obj id="14" subid="h">Hotel</obj></ounion>liegt  
<aunion><a id="p1431c" type="p">near the airport</a></aunion>  
</match>  
</concordance>
```

It is, however, possible to create a larger context using the **start** and **end** attributes and the **<tokens>** element:

```
<tokens>...  
<tok index="64">.</tok>  
<tok index="65"/>  
<tok index="66">The</tok>  
<tok index="67"/>  
<tok index="68">Hotel</tok>  
<tok index="69"/>  
<tok index="70">is</tok>  
<tok index="71"/>  
<tok index="72">near</tok>  
<tok index="73"/>  
<tok index="74">the</tok>  
<tok index="75"/>  
<tok index="76">Airport</tok>  
<tok index="77"/>  
<tok index="78">and</tok>  
<tok index="79"/>  
<tok index="80">the</tok>  
<tok index="81"/>  
<tok index="82">city</tok>  
<tok index="83"/>  
<tok index="84">center</tok>  
<tok index="85"/>  
<tok index="86">.</tok>  
<tok index="87"/>  
...
```

We can expand the original token range (index 68 to 76) in both directions until we reach a punctuation character. We would reach token 64 (".") at the beginning and token 86 (".") at the end of the new range and can then display the whole match context: **"Das Hotel is near the airport and the city center"**.



Illustration 1: Highlighted extended contexts on trustyou.com

4.4. Categories

Entities can easily be assigned to different categories with the help of their matches' object and attribute ids and the categories reference list (s. Categories).

The list contains information about which object id / attribute id combinations justify the hotel's assignment to a specific category. E.g. a hotel that has reviews which contains matches with object id 11 ("room") and attribute id p10 ("clean") can be assigned to category "cat1" ("Saubere Zimmer", "Clean Rooms").

Categories

The hotel is perceived by reviewers as follows:

What they say about you

Wellness/Sport Area (2)	Clean Rooms (7)
Central Location (3)	Good Breakfast (10)
Quiet Location (8)	Good Deal (4)
Good Food (17)	Big Rooms (2)
Good Service (26)	Free Internet (1)

Hotel type

[City Trip Hotels](#) (3)

Type of travellers

[Sportsmen](#) (2)

Illustration 2: Category overview on Trustyou Analytics

4.5. Tops & Flops

In addition to assigning hotels to categories, we can also find the realms that reviewers are especially **excited disappointed** about. Let's examine the <matches> element in an entity in order to find out what people like most about a hotel.

Since we want to find the positive matches, we have to check the <matchpos> element.

We then need to examine every <match> element's freq attribute:

```
<matchpos freq="81">
  <match freq="12" id="20">
    ... (<submatch>es and <norm>s left out)
  <match freq="7" id="18">
  <match freq="62" id="15">
</matchpos>
```

We can see, that 81 positive matches were found in total. Most of the matches (freq="62") refer to object id 15 ("Grounds").

We can now dig into the details of that match group and continue finding the most frequent submatch or the most frequent exact match:

```

...
<submatch freq="37" aid="p05" id="a">
  <norms>
    <norm freq="5" lang="en" stage="1">
      <acc/>
      <nom>great pool</nom>
      <neg/>
    </norm>
  </norms>
...

```

A list of a maximum of the five most frequent positive and negative matches regarding "Food+Drink" could look like this:

Tops and Flops:

Food+Drink	<ul style="list-style-type: none"> <u>+ amazing restaurant</u> (11) <u>+ lots of good restaurants</u> (4) <u>+ The restaurant and bar were very nice</u> (2) <u>+ The breakfast is a very good</u> (2) <u>+ grill offers the best</u> (2) <u>- the food wasn't good</u> (1) <u>- the chicken was a bit dry</u> (1) <u>- the food was awful</u> (1) <u>- Wasn't too pleased with the breakfast menu</u> (1)
------------	---

Illustration 3: "Tops and Flops" on trustyou.com

Once we have found the matches that we are interested in, we can use their analyse id (s. Linking reviews to matches) in order to find the corresponding reviews.

5. Changes

Version	Author	Changes	Date
0.9	Markus Reil	Initial version	23.02.2010
0.9.1	Markus Reil	Updated images, examples	08.09.2010

Appendix I: XML Schema definition

Appendix II: Entity types

Appendix III: Match objects

See file "Referenzliste_Objects.xls" for a complete list of objects and their ids.

Appendix IV: Categories

See file "Referenzliste_Metacats.xls" for a complete list of categories and their object id / attribute id combinations.