Imprint

gpGeocoder Documentation

Documentation Overview release notes gpGeocoder-1.8.25.0-SNAPSHOT

Result Classification

Requirements

- Return a quality description of the results compared to the input.
- Describe classes for the result qualities.
- Describe characteristical features of the result.
- Allow a sorting of the results.
- Give Criteria for batch geocoding.

Explicitly **not** in the focus of the classification are the following issues:

- Give a continuous number to rate the result quality.
- Provide information about internals of the geocoder algorithm.

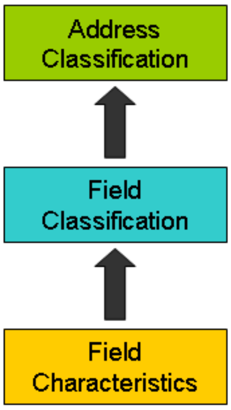
From these requirements, the classification was specified as follows.

Specification

Introduction

The classification of a result address is a three-stage process:

- 1. **Field Characteristics** For each field all applicable characteristics are determined. The characteristics describe properties of a field, in the majority of cases applying comparison criterions on input and result.
- 2. **Field Classification** Based on only these characteristics, the classification of each single field is computed.
- 3. **Address Classification**The classification of the complete result address is derived from the classification of all fields.



A top-down description of each stage follows below:

Address Classification

The classification of the complete address is done according to the rules that are given in the following table.

Classification	Condition
Exact	(P == Exact || NoInput) && (C == Exact || NoInput) && (S == Exact || NoInput) && (H == Exact || NoInput)
Partially exact	(P >= Partially exact || NoInput || NoResult) && (C >= Partially exact || NoInput) && (S >= Partially exact || NoInput) && (H >= Partially exact || NoInput || NoResult)
High	(P >= High || NoInput || NoResult) && (C >= High || NoInput) && (S >= High || NoInput) && (H >= Partially exact || NoInput || NoResult)
Medium	(((P >= Partially exact || C >= Partially exact) && (S >= Partially exact || NoInput)) || (P >= Partially exact || NoInput || NoResult) && (C >= Partially exact || NoInput) && (S >= Partially exact || NoInput || NoResult) || (P >= Medium || NoInput || NoResult) && (C >= Medium || NoInput) && (S >= Medium || NoInput)) && (H >= Partially exact || NoInput || NoResult)

Low	Rest

Legend	
P	Postalcode
C	City/City2
S	Street
H	House Number

Field Classification

Classification	Availability				Definition
	Postalcode	City/City2	Street	HNr	
Exact	✓	✓	✓	✓	For details for each single field see below.
Partially exact	✓	✓	✓	✓	For details for each single field see below.
High	✓	✓	✓	⛔	Not contained in superior classes and Match quality high is set.
Medium	✓	✓	✓	⛔	Not contained in superior classes and Match quality medium is set.
Low	✓	✓	✓	✓	Not contained in any other class
NoInput	✓	✓	✓	✓	The input field is empty.
NoResult	✓	✓	✓	✓	The result field is empty and the input field is not empty.

Postalcode	
Exact	Match exact
Partially exact	Input is prefix of result

City/City2 + Street	
Exact	Match exact && Input is prefix of result
Partially exact	(Input is prefix of Multiword result || Match except abbreviations || Match except separators) && Match quality high && Result words include input

House Number	
Exact	Match exact
Partially exact	Match exact except additions

Field Characteristics

Definitions

- The **definition** of a term is written bold letters. The *usage* of a defined term is written in italic letters.
- The steps listed in the column Preprocessing are applied in the given order on both input and result. Then, the characteristic is true if the preprocessed data fulfills the definition. Following preprocessing steps are available:
  - **SN** Special character normalisation: Special characters and umlauts are replaced (e.g ü->ue, é -> e, ...).
  - **CN** Case normalisation: Upper and lower case characters are treated in the same way.
  - **AN** Abbreviation normalisation: The replacements defined in the abbreviation dictionary are applied.
  - **XN** Affix normalisation: All affixes defined in the affix dictionary are removed from the string.
  - **PN** Phonetic normalisation: The phonetic replacements are applied.
  - **SepN** Separator normalisation: All separators (Space, -, /, ., etc.) are treated in the same way. Multiple consecutive separators are projected to a single one.
- Prefix: x is Prefix of another word y if x is not longer than y and characters on position i (i = 1 ... length(x)) in x and y are identical.
- Multiword Prefix: A list of words w is a Multiword Prefix of a list of words z, if an injective mapping from the words w_1 ... w_n to the words z_1 ... z_m exists such that each word w_1 ... w_n is Prefix of its corresponding word in z_1 ... z_m. (Note that m >= n)

Characterization bitfield values

The characterization is returned as a number that represents a bitfield. The values are given in the enumeration **eResultCharacteristics**.

Characteristic	Availability				Preprocessing	Definition	Examples	
	Postalcode	City/City2	Street	HNr			Positive	Negative
Input is prefix of result **rcInputIsPrefixOfResult**	✓	✓	✓	⛔	CN, SN, SepN	Input is *Prefix* of result	I: karl O: Karlsruhe	I: karle O: Karlsruhe

Input is phonetic prefix of result **rcInputIsPhoneticPrefixOfResult**	⛔	✓	✓	⛔	CN, SN, SepN, PN	Input is *Prefix* of result	I: karl O: Karlsruhe I: carl O: Karlsruhe	I: karle O: Karlsruhe
Input is prefix of Multiword result **rcInputIsPrefixOfMultiwordResult**	⛔	✓	✓	⛔	CN, SN, SepN	Input is *Multiword Prefix* of result	I:Wei Beu O:Beutelsbach Weihersdorf I: Beu O:Beutelsbach Weihersdorf	I:BeuWei O: Beutelsbach Weihersdorf I: Beu Wei O:Beutelsbach
Input is prefix of phonetic Multiword result **rcInputIsPrefixOfPhoneticMultiwordResult**	⛔	✓	✓	⛔	CN, SN, SepN, PN	Input is *Multiword Prefix* of result	I:Karls Turla O:Karlruhe Durlach I:Carls O:Karlruhe Durlach	I:BeuWei O: Beutelsbach Weihersdorf I: Beu Wei O:Beutelsbach
Match except abbreviations **rcMatchExceptAbbreviations**	⛔	✓	✓	⛔	CN, SN, SepN, AN	Input is *Multiword Prefix* of result	I:Stumpfstrasse O:Meine Stumpfstr	I:Stumpfstrasse O:Stumpfweg I:Stumpfstrasse O:Stampfstr
Match except affixes **rcMatchExceptAffixes**	⛔	✓	✓	⛔	CN, SN, SepN, XN	Input is *Multiword Prefix* of result	I:Stumpfweg O:Stumpfstr	I: Stumpfstra O: Stumpfstr
Result words include input **rcResultWordsIncludeInput**	⛔	✓	✓	⛔	CN, SN, SepN, AN	Input is *Multiword Prefix* of result	I: Beu O:Beutelsbach Weihersdorf	I: Beu Wei O:Beutelsbach
Input words include result **rcInputWordsIncludeResult**	⛔	✓	✓	⛔	CN, SN, SepN, AN	Result is *Multiword Prefix* of input	I: Beutelsbach Wei O:Beutelsbach	I: Beu O:Beutelsbach Weihersdorf
Match representative postal code **rcMatchRepresentativePostalCode**	✓	⛔	⛔	⛔	CN, SepN	The representative postal code matches to the input	I:76000 O:76*** I:76000 O:76229, RepPostcode: 76***	I:76000 O:75000
Match length **rcMatchLength**	✓	✓	✓	⛔	CN, SN, SepN	The length of the input and the result is the same	I:Beutelsbach - Weihersdorf O: Beutelsbach Weihersdorf I: Karlsruhe O: Stuttgart	I:Beu Wei O: Beutelsbach Weihersdorf
Match exact **rcMatchExact**	⛔	✓	✓	⛔	CN, SN, SepN, AN	Result is *Multiword Prefix* of input and Input is *Multiword Prefix* of result	I: Weihersdorf Beutelsbach O: Beutelsbach Weihersdorf	I: Weihersdorf Beutelsdorf O: Beutelsbach Weihersdorf
Match exact **rcMatchExact**	✓	⛔	⛔	⛔	CN, SepN	The strings are identical	I: ab3-b5 O: AB3 B5	I: AB3B5 O: AB3 B5
Match exact **rcMatchExact**	⛔	⛔	⛔	✓	(none)	The strings are identical	I: 42 O: 42	I: 42a O: 42
Match except separators **rcMatchExceptSeparators**	✓	✓	✓	⛔	(delegated to Match exact)	After removing all separators Match exact applies	I:Media Park O:Mediapark	I:Media Par O:Mediapark
Match quality high **rcMatchQualityHigh**	✓	✓	✓	⛔	(various)	The rating algorithm of the gpGeocoder indicates a high quality	I: Stampfstr O: Stumpfstr	I: Bbuch O: Beutelsbach Weihersdorf
Match quality medium **rcMatchQualityMedium**	✓	✓	✓	⛔	(various)	The rating algorithm of the gpGeocoder indicates a medium quality	I: Stamfstr O: Stumpfstr	I: Karlsruhe O: Berlin
Match quality low **rcMatchQualityLow**	✓	✓	✓	✓	(various)	The rating algorithm of the gpGeocoder indicates a low quality	I: Karlsruhe O: Berlin	I: Karlsruhe I: Stamfstr O: Stumpfstr
CityFieldPermutation **rcCityFieldPermutation**	⛔	✓	⛔	⛔	CN, SN, SepN, AN	Not(Considering input City field and result City field Input is Multiword Prefix	I: C: Durlach O:C: Karlsruhe C2: Durlach	I: C: C2: Durlach O: C: Karlsruhe C2: Durlach

						of result is fulfilled. The same condition must hold for city2.)		
Match exact except additions **rcMatchExactExceptAdditions**	⛔	⛔	⛔	✓	(delegated to Match exact)	If additions are removed, input and result Match exact	I: 42a O: 42	I: 4 2a O: 42

Notes

- **>** The examples in above tables are only for explanatory purposes. The authoritative information is given in the definition.
- **>** Note that always exactly one of Match quality high, Match quality medium or Match quality low is true.

PTV Geoplatform, created on Tue Sep 7 15:37:22 2010