Task 2.12.1 Keyword Text Analysis Mar 20, 2015

Overview

The listing data contains a number of text fields, such as property description, agent comments and photo captions from which we want to extra information. We want to scan these text fields to locate predefined word phrases and increment a counter when the phrase is found.

The entire process is to be table driven so that new phrases can be added as needed. The phrases are assigned a QDNA key and 4 digit token. Basically the approach involves searching for keyword phrases in the text and optionally examining the words before and after the phrase to refine the QDNA category.

For example, the key word phrase “on the 9th hole” can take on a number of forms such as;

on the nth hole

over looks the nth hole

on nth hole

backs onto nth hole

The approach must be able to recognize the different variations and mapped them to the same QDNA and token. The Keyword\_Phrase table defines the keywords and the relative location of secondary words (adjectives) associated with the keyword.

Process

* Find the keyword in the text

If Relative=0 and contains a single word

* + Extract the 4 words proceeding and following the keyword from the text. Any word proceeding or following a period, comma, asterisk, “and”, “/”, “&”, “!” or other special character or short connecting word should be deleted. Short shorts at the beginning or end should also be deleted, “the”, “a”, “to”, “it”, “for”, “in”, etc (see examples below)
  + Add QDNA & Token to Real\_QWords table including the resulting string

Example 1: Keyword = Updated

Text: There are three bedrooms on the main level and one **updated** bathroom. There is a huge addition off the 3rd bedroom.

Extracted words: main level and one **updated** bathroom. There is a

Expected Results: one **updated** bathroom

Example 2: Keyword = Updated

Text: \* Cozy Sunroom \* Spacious Back Patio \* Updated Plumbing, Electrical Box & Water Heater

Extracted words: Spacious Back Patio \* Updated Plumbing, Electrical Box & Water

Expected Results: Updated Plumbing, Electrical Box

Example 3: Keyword = Updated

Text: Dream master bedroom with updated 5 piece bath and walk-in closet.

Extracted words: Dream master bedroom with updated 5 piece bath and

Expected Results: Dream master bedroom with updated 5 piece bath

Example 4: Keyword = Updated

Text: Three levels with WO basement. 3 Bedroom septic. Updated in 2008 New roof.

Extracted words: basement. 3 Bedroom septic. Updated in 2008 New roof.

Expected Results: Updated in 2008 New roof

If Relative=0 and contains a more than one word

* + Add QDNA & Token to Real\_QWords table

If Relative<0

* + Locate the adjective words before the keyword. Only search the number of words before by the number given (ie. If Relative=-2, search a max of 2 words before the keyword)

If Relative>0

* + Locate the adjective words after the keyword. Only search the number of words after by the number given (ie. If Relative=2, search a max of 2 words after the keyword)

If Token=0

* + SKIP. No keyword match.

Note:

Seq controls the order in which the keyword search is performed. If Sequence not specified, the order doesn’t matter.