Discover and use real-world terminology with IBM Watson Content Analytics

Build sample domain dictionaries for data analysis

Jean-Marc Langé
Client Technical Professional, IBM Software Group
IBM

04 December 2014

Use linguistic analysis in IBM® Watson™ Content Analytics (WCA) to explore domain-specific terminology, and build domain dictionaries that reflect users' "real-life" vocabulary preferences. Use these dictionaries in WCA Studio to build concept annotators.

Introduction

The case for structuring unstructured data

There is much interest about the wealth of information that society produces in ever-growing quantities (be it within the enterprise, on the web, or in social networks). You can use that data in several ways to derive insights that might improve health, democracy, or the way you do business. These data-based insights are the traditional playground of Analytics or Business Intelligence (BI), which typically rely on structured data, such as dates, financial amounts, quantities, or company names. However, most data is in unstructured form — texts, images, movies — in proportions that vary from 70% for enterprise data to almost 100% in social media.

Any analytics application that uses only structured data therefore does away with about four fifths of the available information. Extracting structured information from unstructured sources appears a must in the big data era. This tutorial focuses on textual data and shows how to extract terminological information that is relevant for a business domain.

IBM Watson Content Analytics

IBM Content Analytics with Enterprise Search is a search and analytics platform. It uses rich-text analysis to surface new, actionable insights from many sources and types of textual content, including enterprise content, web content (including social media), email, or databases.

In practice, IBM Watson Content Analytics (WCA) can be used in two general ways:
• Immediately use WCA analytics views to derive quick insights from sizeable collections of contents. These views often operate on facets. *Facets* are significant aspects of the documents that are derived from either metadata that is already structured (for example, date, author, tags) or from concepts that are extracted from textual content.
• Extracting entities or concepts, for use by WCA analytics view or other downstream solutions. Typical examples include mining physician or lab analysis reports to populate patient records, extracting named entities and relationships to feed investigation software, or defining a typology of sentiments that are expressed on social networks to improve statistical analysis of consumer behavior.

WCA uses Natural Language Processing technology (NLP) for extracting information from unstructured data (or texts). That information can be found in the following forms:

• Atomic concepts or entities, such as persons, places, companies, aircraft parts, manufacturing actions;
• Combinations of the preceding information, generally involving some level of relationship between concepts. Examples might be a person and her job, a company and its industry domain, a maintenance operation of a specific aircraft part, a patient medical antecedent that involves a family link and a health issue.

WCA processes raw text from the content sources through a pipeline of operations that is conformant with the UIMA standard. UIMA (Unstructured Information Management Architecture) is a software architecture that is aimed at the development and deployment of resources for the analysis of unstructured information. WCA pipelines include stages such as detection of source language, lexical analysis, entity extraction, or application of custom concept extraction. Custom concept extraction is performed by annotators, which identify pieces of information that are expressed as segments of text. Annotators can be created with IBM Content Analytics Studio (WCA Studio), a graphical, Eclipse-based environment that facilitates the design and testing of annotators based on dictionaries and rules.

The focus in this paper is how to streamline the creation of domain dictionaries. The acquisition of dictionaries can seem an easy task when domain terminologies are available. However, the reality of contents is that the authors do not necessarily follow the canonical terminology as is seen further on. Hence the need for a corpus of texts that is representative of the domain under study. In this paper I use a corpus of complaints from automobile users. I describe the exploration of the corpus in search of domain terminology, with the help of WCA native linguistic and analytic functionalities. I show how these operations can be streamlined to semi-automatically produce dictionaries that can be used in WCA Studio to perform further annotation tasks. In the final section, I describe one possible use of the dictionaries, that is, tagging new complaint records with information on the components that are possibly involved in the problem.

The paper assumes basic knowledge of Watson Content Analytics. For more information on WCA, see Resources.
Building the source corpus

The sample source

The United States Department of Transportation established the National Highway Traffic Safety Administration (NHTSA), so that information on all vehicle safety-related issues is available to consumers. NHTSA accepts complaints on auto safety issues, through their website or other channels such as email or phone.

On the website (https://www-odi.nhtsa.dot.gov/VehicleComplaint/index.xhtml), users can enter information about the vehicle (Make - Model – Year) and the incident. The latter part contains both constrained choice fields: date, whether there was a fire / crash / injuries, mileage, speed, affected parts (out of a list of 17 high-level car parts), and free text input under "Tell us what happened." As a result, the data available for an incident contains the typical mixture of structured and unstructured data:

Figure 1. NHTSA form for filing a safety complaint

The resulting data is made publicly available on NHTSA site. These contents are anonymous in that they do not contain personal identities or license plates. On the complaint data, I found that the field for "component" is filled with a richer set of values than the few allowed in the online form, where users can choose only three of 17 high-level car parts. This richness suggests that the NHTSA data is enriched with information after user input, although it is not documented on their site.

Loading NHTSA data into WCA

The NHTSA data can be obtained at the FLAT FILE COPIES OF NHTSA/ODI DATABASES page. The WCA product knowledge center explains how to import a .csv file into a collection. This process is not detailed further here.
I downloaded over 230,000 records corresponding to user complaints that span the years 2005 - 2011.

I imported the records in a Watson Content Analytics collection of documents. In the configuration for the collection, I specified that index fields and facets be created for all structured fields such as incident date, vehicle make and model, occurrence of a fire, crash, injuries, or vehicle components that are mentioned in the complaint.

**Figure 2** shows the Document view in Content Analytics Miner, the analytics application that is provided with WCA. A list of facets is presented on the leftmost column and summaries of different records on the right side. The request terms, either typed in or selected through facet navigation, are highlighted in these summaries:

**Figure 2. Document view in Content Analytics Miner**

![Document view in Content Analytics Miner](image)

Using linguistic facets to discover domain terminology

Now look at how linguistic facets provided with IBM Watson Content Analytics can help discover domain-specific vocabulary.

**WCA linguistic facets how-to**

WCA provides immediately usable facets for part-of-speech information for single words, such as noun, verb, or adjective, and for phrases, such as noun phrases. The Facets view in the Content Analytics Miner shows the values that a specific facet can take in the set of documents that are selected by the current query. These values can be sorted by **frequency**, that is, the number of documents that contain the specific facet value, or **correlation**. **Correlation** is a measure of how strongly the facet value is related to the set of documents that are selected by the current query, compared to the other documents in the collection.

To better understand the difference between frequency and correlation, look at the list of nouns that are found in the whole collection of documents, which are sorted by frequency. Select the **Facets** view, and facet **Part of speech > Noun > General Noun** in the Facet explorer. (**General Noun** represents those words that are identified as nouns in the general language dictionary, whereas **Others** represent unknown words.)
Figure 3. Topmost values for linguistic facet "noun"

Figure 3 shows the most frequent nouns in the whole collection of documents. Not surprisingly, users speak of their problem with vehicle, dealer, or car.

Note: It is in the list because, since the NHTSA data is in uppercase, it was mistakenly identified as a noun in certain contexts.

Now enter the search query "storm" (which yields 644 documents in the collection), and observe how the list of selected nouns dynamically changes:

Figure 4. Topmost nouns with query "storm"

Apart from storm itself and rain, the other topmost words: problem, car, vehicle, time, while not related to the concept of storm, still appear here. They are frequent in the whole corpus and therefore also appear in those documents that are related to storm.

Sorting the results by correlation tells a different story:
Figure 5. Topmost nouns with query "storm", sorted by correlation

Here, you intuitively understand the close proximity between storm and snow, rain, or water, and you know that it bears some relationship with wiper, windshield, or window. You see families of words that are semantically linked. **Correlation** is clearly a better indication than **frequency** for the semantic proximity between the noun facet values and the query terms.

Now look at the noun phrases, more specifically noun sequences, which are sorted by correlation. Select those linguistic facets under the **Phrase Constituent** super facet:

Figure 6. Topmost noun sequences with query "storm"

The list in Figure 6 makes sense in the context of storm. The top noun sequences are composed mostly of words that appear in the list of top nouns, which seems logical.

Another interesting phrase facet is the **modified noun** (that is, typically adjective + noun):

Figure 7. Topmost modified nouns with query "storm"

Again, you find phrases that contain the word "storm" or that are related to adverse weather conditions such as heavy rain, late winter, or inclement weather.
Finding words or phrases that are associated with a specific concept

If you already have a series of facets that identify general concepts with some business interest, you can use them instead of plain search queries to navigate the collection. Consider, for example, specific vehicle components and what vocabulary is associated with them. For that purpose, select the Component facet in the facet explorer, turn to the Facets view, and select one or more of the values.

Before that, maximize the number of facet values displayed (100 by default), since many more values are possible than the 100 set by default in WCA. For that purpose, click the Preferences icon, pick the Facets tab, and set Count to analyze to 500:

**Figure 8. Setting number of facets in display**

Now look at the possible values for facet Component in the collection. Filter those values that contain "wheel", then select value "WHEELS:LUGS/NUTS/BOLTS" and add it to the current query as in **Figure 9**:

**Figure 9. Selecting documents that are related to Component "WHEELS:LUGS/NUTS/BOLTS"**

This operation results in 276 complaints that are tagged by NHTSA as related to WHEELS:LUGS/NUTS/BOLTS. How do users refer to these components in their own words? Return to the linguistic facets and check which nouns are highly correlated with the set of documents that are returned by your facet navigation:
Figure 10. Top nouns that are related to component "WHEELS:LUGS/NUTS/BOLTS"

Here you find words from the facet name (lug, nut, bolt, wheel), and other words that are related to the subject, such as "stud", which describes the male part on which the lug nuts are fastened, or the wheel hub, to which bolts can be screwed to secure the wheel. These words reflect the fact that users can employ variable terminology when they freely express their concerns. Users do not necessarily use the standard terminology, but instead might prefer related terms, synonyms, or paraphrases as exemplified in this complaint excerpt:

"[...] HAD AFTERMARKET WHEELS AND TIRES REPLACE FOR A BETTER LOOK. ON 09/06, A YEAR LATER, HAD TO HAVE WHEEL STUD REPLACED. DEALERSHIP BILLED ME STATING NOT UNDER WARRANTY BECAUSE HAD WHEELS & TIRES REPLACED. JUST YESTERDAY 11/12/06, SAME INCIDENT OCCURRED AGAIN. ONLY TWO STUDS HAVE TO BE REPLACED THIS TIME"

Suppose you wanted to automatically categorize complaints by matching their words with the keywords contained in the category titles. Here you see no lug, nut, or bolt that allows you to choose the specific category WHEELS:LUGS/NUTS/BOLTS. Instead, you need to add keywords such as "stud" or "hub" to the list of keywords that characterize this category.

As you navigate the Component concept with WCA linguistic facets, it is easy to find similar examples:

- Words that are associated with "COMMUNICATIONS:HORN ASSEMBLY" include the verb honk, which intuitively sounds relevant;
- Component "STRUCTURE:BODY:DOOR" yields nouns slide, door, handle, opening, and verbs slide, close, open, latch, or snap. Again, these words intuitively belong to the lexical field of doors, either with door parts (handle) or things you can do with doors (close).

You see how WCA can help you discover the vocabulary that is typically associated, in the voice of users, with a specific subject area. You use this property later in this article, when you tag NHTSA data categories. Before you do so, let's explore other aspects of terminology identification with WCA.
Known words that come as a surprise, unknown words that deserve to be known

The word list in Figure 10 also contains words that are not directly related to the component category you selected. One indirectly related word is common in the North American automobile domain, *turnpike*, but another word is surprising: *Dane* – what’s a Dane doing here? Watson Content Analytics version 3.5, among many new features, offers the possibility to easily change layouts of the analytics application. In particular, the Advanced analytics layout allows viewing the relevant document summaries alongside another view, AND it changes the selection and keyword highlighting of documents when something is selected in this other view. For example, if you click “Dane” in the facet view, the document view shows relevant documents with the word highlighted, which immediately tells you that this Dane comes from the commercial name of a trailer:

**Figure 11. Finding instances of "Dane" in complaints**

This facility becomes precious when you need to compare your terminology intuitions to the reality of text.

Besides finding domain terms (*rain storm*), related terms (*windshield wipers, hub*) or noise (*turnpike, Dane*), WCA linguistic facets also can help find other items of interest. So far you used the General noun subfacet of Noun. Look at the subfacet "Noun / others", which includes tokens that are not found in WCA dictionaries. You are still in category "WHEELS:LUGS/NUTS/BOLTS":

**Figure 12. Exploring out-of-vocabulary words**

The results include a number of misspellings (*chekcing, comprhensive*), acronyms (*NSA*), rare or newly coined terms (*overtightened*), proper nouns, or specific technical words that WCA does not have in its general vocabulary (*Ohio, embrittlement*). In other cases, there are admissible
variants of a form, such as McPherson for MacPherson. All these findings can be of interest from a terminology perspective — for example, frequently occurring misspellings, or acronyms that are used to shorten a domain word, might have a place as alternative forms in your dictionaries.

Implementing more linguistic facets
During your exploration of technical data, it turned out that some of the relevant terms for a component class were not identified by WCA linguistic facets. Such is the case for terms such as right front wheel, main drive pulley, or independent repair shop, which follow a grammatical pattern adjective-noun-noun. This pattern is not detected by WCA "modified noun" linguistic facet, which is limited to adjective-noun patterns.

To enhance WCA built-in facets, I used WCA Studio to create an annotator that locates such patterns. Once deployed on WCA pipeline, the annotator feeds a facet that I dubbed Additional Linguistic Patterns/AdjNP.

Figure 13 shows the new facet Adj NP (for Adjective-Noun phrase) in your WCA implementation, along with a sample of the phrases that it identified still for WHEELS:LUGS/NUTS/BOLTS):

![Figure 13. Implementation of the new linguistic facet Adj-Noun-Phrase](image)

Summing up this first section, you used WCA analytics capabilities to identify words or phrases that are related to certain classes of components. Some of these words are terms that belong to the semantic domain of the component (such as synonyms), others are variations (acronyms, misspellings, alternative forms). In the next section, I will use this capability to create domain dictionaries that can be further used for tasks such as automatic tagging of new complaints.

Domain dictionary creation
You know enough now to create your own domain dictionaries, with relevant terminology that can be used for several purposes. I will show how to create such domain dictionaries manually by using the Content Analytics Miner functionality, and then how to automate dictionary creation with the help of WCA application programming interface (API).

Dictionary creation, the manual way
Exporting facet values
Facet values that are displayed in the "Facets" view can be saved into a comma-delimited (.csv) file. Click the Report button and choose the relevant option. Each facet value comes with its frequency and correlation. For example, after you select the "WHEELS:..." component, the Noun
sequence linguistic facet gives the following output once you load it in a spreadsheet software and sort it by frequency (in Figure 14) or correlation (in Figure 15):

**Figure 14. Export of "Noun Sequence" facet values (sorted by frequency)**

<table>
<thead>
<tr>
<th>facet value</th>
<th>frequency</th>
<th>correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oakl traffic crash report</td>
<td>56</td>
<td>58.49</td>
</tr>
<tr>
<td>failure rulings</td>
<td>40</td>
<td>4.90</td>
</tr>
<tr>
<td>lug nut</td>
<td>42</td>
<td>174.13</td>
</tr>
<tr>
<td>tire wheel</td>
<td>21</td>
<td>16.13</td>
</tr>
<tr>
<td>wheel stud</td>
<td>20</td>
<td>102.25</td>
</tr>
<tr>
<td>front tire</td>
<td>18</td>
<td>4.73</td>
</tr>
<tr>
<td>police report</td>
<td>14</td>
<td>3.98</td>
</tr>
<tr>
<td>rear wheel</td>
<td>11</td>
<td>18.13</td>
</tr>
<tr>
<td>wheel housing</td>
<td>10</td>
<td>2.89</td>
</tr>
<tr>
<td>fender flaps</td>
<td>8</td>
<td>1.62</td>
</tr>
<tr>
<td>repair shop</td>
<td>6</td>
<td>0.56</td>
</tr>
<tr>
<td>passenger side tire</td>
<td>4</td>
<td>0.65</td>
</tr>
<tr>
<td>wheel lug</td>
<td>7</td>
<td>158.08</td>
</tr>
<tr>
<td>oil fill</td>
<td>7</td>
<td>26.36</td>
</tr>
<tr>
<td>NHTSA campaign id number</td>
<td>7</td>
<td>8.41</td>
</tr>
<tr>
<td>lug bolt</td>
<td>7</td>
<td>119.69</td>
</tr>
<tr>
<td>wheel assembly</td>
<td>6</td>
<td>18.87</td>
</tr>
<tr>
<td>tire rotation</td>
<td>6</td>
<td>4.49</td>
</tr>
<tr>
<td>passenger wheel</td>
<td>5</td>
<td>12.42</td>
</tr>
<tr>
<td>lug stud</td>
<td>5</td>
<td>138.51</td>
</tr>
<tr>
<td>left front wheel</td>
<td>5</td>
<td>61.90</td>
</tr>
<tr>
<td>starter unit</td>
<td>5</td>
<td>105.65</td>
</tr>
<tr>
<td>ford explorer</td>
<td>5</td>
<td>8.43</td>
</tr>
<tr>
<td>front header</td>
<td>4</td>
<td>8.18</td>
</tr>
<tr>
<td>passenger side</td>
<td>4</td>
<td>6.27</td>
</tr>
</tbody>
</table>

**Figure 15. Export of "Noun Sequence" facet values (sorted by correlation)**

In the spreadsheet, you can sort and filter the list according to different criteria. For example, you might decide to keep only those terms that have a correlation value above a certain threshold, ensuring good relevance of the results.

You need to repeat what you did for noun sequences for other term-productive parts of speech, such as simple nouns, modified nouns, or the enhanced "adjective noun phrase" pattern. Figure 16 shows a list of the topmost "terms" obtained with these different patterns and sorted by correlation, for component category "FUEL SYSTEM, GASOLINE" — the whole process took only minutes:
In Figure 16, you see a mix of single and multi-word units that in most cases appear to be relevant terms for the "engine" domain. Some noise appears, notably with car brand or model names. Some terms are incomplete or too extended. For example, a quick examination of occurrences of "sludge build" in the documents shows that the complete term is a combination of "sludge" and "build up". Similarly, "engine failure due" should be shortened to "engine failure" as the "due" is always part of a "due to..." phrase. The enhanced analytics layout in WCA Content Analytics miner is great help in that process, since a click a facet value filters the documents view for all occurrences of the corresponding term:

![Figure 16. Export of all linguistic facet values for component category "FUEL SYSTEM, GASOLINE", sorted by correlation](image)

Figure 17. Looking for facet value "sludge build" in the complaints

This filtering allows for a quick review of the list by domain specialists who accept or reject the candidates in the terms list, based on their own domain knowledge and evidence from the documents.

You might use the resulting term list for further analysis, this list is imported into a WCA Studio dictionary. In that perspective, you need to add part-of-speech information about each term. Since you deal with noun sequences, the resulting part of speech should be "noun". This information can be easily added for all or part of the entries with simple spreadsheet commands.
Using the dictionary in WCA Studio

Importing the dictionary into WCA Studio is straightforward with the dictionary import wizard. The creation and import of a dictionary in WCA Studio are described in detail in "Chemical Dictionaries in ICA Studio," a developerWorks article.

The resulting dictionary can be used to feed a subfacet of "ENGINE...", where the values would be engine parts such as rod bearing, cylinder head, oil pump, spark plug... The dictionary can also be part of a higher-level rule that combines engine parts with another type of concept to generate a new annotation. The annotations that are built from dictionaries and rules can be deployed to the WCA run time where they can feed new facets in WCA miner.

Streamlining dictionary creation

In the previous section, you saw how to export terminology candidates from WCA into WCA Studio for a single linguistic facet (noun). The process must be repeated for every linguistic facet that is interesting for terminology purposes: noun, verb, noun phrase (also known as "noun sequence" in the WCA miner interface).

To streamline this operation, use the WCA REST API to bulk-extract all those facets. For the sake of concision, I outline only the basics here. WCA IBM Knowledge Center explains how to access the REST API documentation.

First, try to get a list of facets associated with the whole collection, by entering the relevant REST call in a browser address field:


which returns something like (picture truncated):

![Figure 18. List of facets that are obtained through WCA REST API](image)

The XML member facet has a label attribute that gives the display name of the facet (for example, "Noun") and an id attribute for the "internal" name (for example, ".word.noun") which is the one that is used in the API calls. Now you can search for possible values of the "Noun" facet with the REST call:

http://localhost:8393/api/v10/search/facet?collection=test_NHTSA&facet={"namespace":"keyword","id":".word.noun"}&query=*

---

Discover and use real-world terminology with IBM Watson Content Analytics
The search/facet REST call returns the possible values for the Noun facet, but also the associated statistics. Here all facet values have a correlation of 1 because the query parameter returns all documents in the collection. The query parameter uses the same syntax as plain WCA queries. Modify your REST call to search for documents that contain the word "engine", and return 500 possible facet values:

```
http://localhost:8393/api/v10/search/facet?collection=test_NHTSA&facet=["namespace":"keyword","id":"$._word.noun","count":"500"]&query="engine"
```

If you extract facet values and correlation from the resulting XML document, and sort the list by decreasing correlation, you obtain the topmost nouns that are listed in Table 1:

**Table 1. Topmost nouns for query "engine"**

<table>
<thead>
<tr>
<th>Noun</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>engine</td>
<td>5.945369512024206</td>
</tr>
<tr>
<td>sludge</td>
<td>5.083791128653582</td>
</tr>
<tr>
<td>check</td>
<td>4.5708567925714405</td>
</tr>
<tr>
<td>misfire</td>
<td>4.434578339511516</td>
</tr>
<tr>
<td>liter</td>
<td>4.100485492528164</td>
</tr>
<tr>
<td>compartment</td>
<td>3.8341929699503408</td>
</tr>
<tr>
<td>cool</td>
<td>3.8088538657369537</td>
</tr>
<tr>
<td>timing</td>
<td>3.6459742040763126</td>
</tr>
<tr>
<td>crankshaft</td>
<td>3.529813794496392</td>
</tr>
<tr>
<td>oxygen</td>
<td>3.4589095714483973</td>
</tr>
<tr>
<td>piston</td>
<td>3.2950541462145404</td>
</tr>
</tbody>
</table>
Next, you want to perform the same operation with a query on the NHTSA "component" facet "WHEELS:LUGS/NUTS/BOLTS"; but how do you express it as a WCA query? Switch back to the Content Analytics Miner application, in the Facets view, select this facet value and click the "Add to query..." button: the query field in WCA now shows the exact syntax that you need to add to your REST call:

&query=keyword::/"component"/"WHEELS:LUGS/NUTS/BOLTS"

Table 2 shows the resulting list of topmost nouns (sorted by correlation) contains familiar entries:

**Table 2. Topmost nouns that are associated with component "WHEELS:LUGS/NUTS/BOLTS"**

<table>
<thead>
<tr>
<th>Noun</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>lug</td>
<td>229,0527407353</td>
</tr>
<tr>
<td>stud</td>
<td>178,3998666573</td>
</tr>
<tr>
<td>nut</td>
<td>81,0483078168</td>
</tr>
<tr>
<td>OHIO</td>
<td>30,8503584036</td>
</tr>
<tr>
<td>BERM</td>
<td>25,8579292585</td>
</tr>
<tr>
<td>Dane</td>
<td>23,2626241945</td>
</tr>
<tr>
<td>bolt</td>
<td>18,5593450909</td>
</tr>
<tr>
<td>wrench</td>
<td>15,8542751694</td>
</tr>
<tr>
<td>dual</td>
<td>13,9362582749</td>
</tr>
<tr>
<td>turnpike</td>
<td>12,7908656193</td>
</tr>
</tbody>
</table>

If you want to automatically create a dictionary that is associated with the different components, Listing 1 shows the typical pseudocode to use:

**Listing 1. Pseudocode to extract linguistic facet values for different components**

For each value \( V_c \) of the "components" facet     # for example WHEELS:LUGS/NUTS/BOLTS...
  For each value \( V_1 \) of linguistic facets      # for example, Noun,Verb,Noun Phrase...
    Get values \( V_{1c} \) of facet \( V_1 \) restricted with a query on \( V_c \)
      For each \( V_{1c} \)
        Output \( V_{1c}, V_c, \) part-of-speech, frequency, correlation
      End for
  End For
End For

This basic code can be refined to include filters — for example to reject candidate terms under certain correlation thresholds — or to include several possible components for a word.
As an example, I implemented a script in Microsoft Powershell that performs such extraction. For each term candidate, it returns the two associated components with the top correlation scores. Figure 20 shows the topmost results. (Output is sorted by correlation. Some component names are truncated.):

Figure 20. Automatically extracted terms and associated component codes

As can be seen, some term candidates have more than one associated component code: for example, **shackle** belongs to **SUSPENSION:REAR....:SHACKLE**, but also has a significant association with **STRUCTURE:FRAME AND MEMBERS**.

Other cases occur when the component codes are related: **tether** or **tether strap** occur significantly in NHTSA complaints that are related to **CHILD SEAT**, but even more in those tagged **CHILD SEAT:TETHER (STRAP)**. The latter are detailed component codes with a greater depth in the hierarchy of components.

In the lower values of correlation, it is no surprise to find more noise:

Figure 21. Automatically extracted terms with lower values of the correlation score
A relationship between brake pressure, or even squeal, and SERVICE BRAKES:... is understandable. It is not clear that uneven road can lead to a problem with SERVICE BRAKES, or how local dealership is linked to SUSPENSION issues.

Similarly, while specific car models had some steering problems in the past, the situation might change as manufacturers take corrective actions. You do not want to pollute your dictionaries with car makes and models, or even peripheral terminology such as local dealership, unless it is relevant for your purpose.

Dictionary building can be streamlined with the help of Watson Content Analytics / WCA Studio, but the resulting dictionaries still need careful revision.

**A possible application: Tagging NHTSA safety complaints**

The dictionary that you created in the previous section can be imported in WCA Studio to build annotators that extract more information from the source texts, and thus possibly provide better insights by using WCA analytics facets.

Your purpose is to help automate tagging of NHTSA complaints with relevant component information. To do so, attempt to find snippets of text that help to identify which vehicle components are at the core of the complaint. For that purpose, use the dictionary that you automatically extracted in the previous section, with its associated contents (component code and correlation score).

The import into a WCA Studio dictionary isn't detailed here but encompasses the following steps:

- Save your terms list to a delimited format that includes columns such as component code and correlation score.
- Create a WCA Studio Dictionary Database.
- In the Dictionary Database creating wizard, add columns for associated values: component code and correlation score.
- Import the delimited terms list into that Dictionary Database. The import wizard proposes a match between the column names in the delimited file and those names in the database, with a possibility to redefine the match.
- Compile the Dictionary Database.

This process is straightforward and can be completed in a few minutes. Figure 22 shows an excerpt of this database, which was obtained with a raw dictionary import (no attempt was made to remove car brands or models nor any kind of noise). The component codes appear in a shortened form to facilitate the process:
Once the dictionary database is completed, it can be used in a lexical analysis step of an UIMA annotation profile. Sample texts analyzed with that profile reveal which of their contents are matched by dictionary entries:

In **Figure 23**, the outline view on the right shows instances of a particular annotation, in this case the component dictionary. In the text file, these instances are highlighted, and hovering the mouse over one instance shows the details of the dictionary entry: *tensioner pulley* is correlated (83.73) with component code ENGINE AND ENGINE COOLING:ENGINE:GASOLINE:BELTS AND ASSOCIATED PULLEYS. If you look at the other snippets in this entry, *serpentine belt* is also correlated to the same component code (corr. 108.3), while steering failure is
associated with STEERING (corr. 12.67) and STEERING:HYDRAULIC POWER ASSIST SYSTEM (corr. 11.68). In all logic, this complaint should be tagged ENGINE AND ENGINE COOLING:ENGINE:GASOLINE:BELTS AND ASSOCIATED PULLEYS: this tagging is indeed the case in the original NHTSA data.

From this quick example, you can easily derive a process to assign a component code to a new complaint. Locate all snippets in a complaint by annotating with a terminology dictionary such as described earlier, and compile an overall score for each component code value that is found in the different snippets, based on the individual correlation scores.

To give an example, I will use the simplest aggregated score, summing up individual scores. In the earlier sample complaint:

- ENGINE AND ENGINE COOLING:ENGINE:GASOLINE:BELTS AND ASSOCIATED PULLEYS would come on top with a global score of 192.03 (83.73 + 108.3),
- then, STEERING with 12.67,
- then, STEERING:HYDRAULIC POWER ASSIST SYSTEM with 11.68.

Such a score, using this simple sum or a more complex formula, can be computed by using Java code as a custom stage in the UIMA pipeline. The custom code would use the correlation figures extracted with the dictionary annotations in the previous annotation stage.

**Conclusion**

In this tutorial, you learned how to explore domain-specific terminology by using the capabilities of Watson Content Analytics, in particular its linguistic facets. These terminologies can be extracted automatically thanks to WCA REST API, and imported into WCA Studio to build simple dictionary annotations or higher-level annotators. Bear in mind that such automatic extractions still contain entries that are incomplete or irrelevant terms, and need careful revision based on the use that you make of the resulting dictionaries.
Resources

Learn

• Learn more about Watson Content Analytics
• Learn how to explore unstructured data with Content Analytics Miner in the WCA IBM Knowledge Center.
• Get in-depth information about the features and capabilities of WCA with the IBM Redbook IBM Watson Content Analytics: Discovering Actionable Insight from Your Content.
• In November 2014, IBM Watson Content Analytics converges with IBM Watson Explorer. Learn how Watson Explorer can help you reduce your research time and work smarter and faster.

Get products and technologies

• Evaluate IBM products in the way that suits you best: Download a product trial or try a product online.

Discuss

• Get involved in the developerWorks community. Connect with other developerWorks users while you explore the developer-driven blogs, forums, groups, and wikis.
About the author

Jean-Marc Langé

Jean-Marc Langé has been with IBM 30 years, working in many different fields that include Natural Language Processing, Speech and Telephony solutions, or Business Process Management middleware. He advises customers on IBM Watson Explorer and IBM Watson Content Analytics products. Jean-Marc lives in Paris.