[MS-FSST]:
Spelltuning File Format Specification

Intellectual Property Rights Notice for Open Specifications Documentation

- **Technical Documentation.** Microsoft publishes Open Specifications documentation for protocols, file formats, languages, standards as well as overviews of the interaction among each of these technologies.
- **Copyrights.** This documentation is covered by Microsoft copyrights. Regardless of any other terms that are contained in the terms of use for the Microsoft website that hosts this documentation, you may make copies of it in order to develop implementations of the technologies described in the Open Specifications and may distribute portions of it in your implementations using these technologies or your documentation as necessary to properly document the implementation. You may also distribute in your implementation, with or without modification, any schema, IDL's, or code samples that are included in the documentation. This permission also applies to any documents that are referenced in the Open Specifications.
- **No Trade Secrets.** Microsoft does not claim any trade secret rights in this documentation.
- **Patents.** Microsoft has patents that may cover your implementations of the technologies described in the Open Specifications. Neither this notice nor Microsoft's delivery of the documentation grants any licenses under those or any other Microsoft patents. However, a given Open Specification may be covered by Microsoft's Open Specification Promise (available here: http://www.microsoft.com/interop/osp) or the Community Promise (available here: http://www.microsoft.com/interop/cp/default.mspx). If you would prefer a written license, or if the technologies described in the Open Specifications are not covered by the Open Specifications Promise or Community Promise, as applicable, patent licenses are available by contacting iplg@microsoft.com.
- **Trademarks.** The names of companies and products contained in this documentation may be covered by trademarks or similar intellectual property rights. This notice does not grant any licenses under those rights.
- **Fictitious Names.** The example companies, organizations, products, domain names, e-mail addresses, logos, people, places, and events depicted in this documentation are fictitious. No association with any real company, organization, product, domain name, email address, logo, person, place, or event is intended or should be inferred.

Reservation of Rights. All other rights are reserved, and this notice does not grant any rights other than specifically described above, whether by implication, estoppel, or otherwise.

Tools. The Open Specifications do not require the use of Microsoft programming tools or programming environments in order for you to develop an implementation. If you have access to Microsoft programming tools and environments you are free to take advantage of them. Certain Open Specifications are intended for use in conjunction with publicly available standard specifications and network programming art, and assumes that the reader either is familiar with the aforementioned material or has immediate access to it.

Revision Summary

Date	Revision History	Revision Class	Comments
11/06/2009	0.1	Major	Initial Availability
02/19/2010	1.0	Major	Updated and revised the technical content
03/31/2010	1.01	Editorial	Revised and edited the technical content
04/30/2010	1.02	Editorial	Revised and edited the technical content
06/07/2010	1.03	Editorial	Revised and edited the technical content
06/29/2010	1.04	Editorial	Changed language and formatting in the technical content.
07/23/2010	1.04	No change	No changes to the meaning, language, or formatting of the technical content.
09/27/2010	1.04	No change	No changes to the meaning, language, or formatting of the technical content.
11/15/2010	1.04	No change	No changes to the meaning, language, or formatting of the technical content.
12/17/2010	1.04	No change	No changes to the meaning, language, or formatting of the technical content.

Table of Contents

1 Introduction

This document specifies the file formats that are used to make query spelling suggestions more relevant..

1.1 Glossary

The following terms are defined in [MS-GLOS]:

Augmented Backus-Naur Form (ABNF)

The following terms are defined in [MS-OFCGLOS]:

dictionary
spell tuning
term frequency

The following terms are specific to this document:

MAY, SHOULD, MUST, SHOULD NOT, MUST NOT: These terms (in all caps) are used as described in [RFC2119]. All statements of optional behavior use either MAY, SHOULD, or SHOULD NOT.

1.2 References

1.2.1 Normative References

We conduct frequent surveys of the normative references to assure their continued availability. If you have any issue with finding a normative reference, please contact dochelp@microsoft.com. We will assist you in finding the relevant information. Please check the archive site, http://msdn2.microsoft.com/en-us/library/E4BD6494-06AD-4aed-9823-445E921C9624, as an additional source.

[ISO-639-1] International Organization for Standardization, "Codes for the representation of names of languages -- Part 1: Alpha-2 code", 2002, http://www.iso.org/iso/catalogue_detail?csnumber=22109

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997, http://www.ietf.org/rfc/rfc2119.txt

[RFC3066] Alvestrand, H., "Tags for the Identification of Language", RFC 3066, January 2001, http://www.ietf.org/rfc/rfc3066.txt

[RFC5234] Crocker, D., Ed., and Overell, P., "Augmented BNF for Syntax Specifications: ABNF", STD 68, RFC 5234, January 2008, http://www.ietf.org/rfc/rfc5234.txt

1.2.2 Informative References

[MS-FSLRDS] Microsoft Corporation, "Linguistic Resource Data Structure", November 2009.

[MS-GLOS] Microsoft Corporation, "Windows Protocols Master Glossary", March 2007.

[MS-OFCGLOS] Microsoft Corporation, "Microsoft Office Master Glossary", June 2008.

[RFC1952] Deutsch, P., "GZIP file format specification version 4.3", May 1996, http://www.ietf.org/rfc/rfc1952.txt

1.3 Structure Overview (Synopsis)

This document describes the file formats that the **spell tuning** component uses to update the spell checking dictionaries:

- The format of a configuration file that enables or disables the item processing stage for term extraction.
- The format of the term frequency files that the item processing stage for term extraction uses to save information about **term frequencies** in the items in the current item batch. The files contain information about the language of the terms, the extracted terms, and the frequencies of the extracted terms. Every extracted term is a single token; that is, it does not contain any spaces. After the item batch is processed, these files are uploaded to the resource store, from which the spell tuning component later retrieves them to update the term frequencies in the spell checking **dictionaries**

1.4 Relationship to Protocols and Other Structures

[MS-FSLRDS] defines the format of the dictionaries that the spell tuning component uploads to the resource store.

1.5 Applicability Statement

The file formats that this document describes are relevant for tuning the spell checking dictionaries in the search engine.

1.6 Versioning and Localization

None.

1.7 Vendor-Extensible Fields

None.

2 Structures

2.1 Term Extraction Configuration File

The term extraction configuration file MUST have the format specified by the following **Augmented Backus-Naur Form (ABNF)** (as specified in [RFC5234]).

```
lines = 1*line
line = content lineend
content = command / comment
comment = "#" *(WSP / VCHAR);
lineend  = LF / (CR LF)
command = "active yes" / "active no"   ; enable or disable the Term extraction
                                       ; "active yes" means enable it
                                       ; "active no" means disable it
```

2.2 Term Extraction Output Files

2.2.1 File Name

Each file MUST conform to the format specified by the following Augmented Backus-Naur Form (ABNF) (as specified in [RFC5234]).

```
filename = item-processor-port underscore hostname underscore timestamp dot extension
item-processor-port = 1*DIGIT          ; the number of the port on which
                                       ; the item processor is running
hostname = 1*(DIGIT / ALPHA)           ; the name of the host,
                                       ; without domain component
timestamp = 1*(DIGIT) [dot 1*3(DIGIT)]; the time of the file creation,
                                       ; the number
                                       ; of seconds that elapsed 1970-01-01
                                       ; UTC (Coordinated Universal Time)
                                       ; as a decimal number. The fractional
                                       ; part including the decimal
                                       ; separator '.' is optional and can
                                       ; contain up to 3 digits.
extension = "out.gz"                   ; fixed filename extension
underscore = %x5f                      ; the underscore character
dot = %x2e                             ; the dot character
```

2.2.2 File Format

The files MUST be compressed through the gzip.exe tool. For more details, see [RFC1952].

In decompressed form, the lines are plain text files where each line MUST have the format specified by the following Augmented Backus-Naur Form (ABNF) (as specified in [RFC5234]).

```
lines = 1*line
line  = language whitespace term whitespace frequency lineend
language = "" / "zh-cn" / "zh-tw" / isolang
term = 1*(DIGIT / ALPHA)  ; A string encoded in UTF-8, representing a single
                          ; token extracted from a item batch
isolang = 2 * ALPHA       ; a two-letter language code as described below
frequency = 1* DIGIT      ; integer representing the frequency of the term
```

```
whitespace = %x20          ; whitespace character
lineend  = LF
```

The language code is the empty string if the language has not been identified by the item processing. The values "zh-cn" and "zh-tw" conform to [RFC3066], and the two-letter language codes MUST conform to [ISO-639-1].

3 Structure Examples

3.1 Term Extraction Configuration File

To enable the term extraction, the configuration file will look as follows:

enable the Term extraction

active yes

To disable the term extraction, the configuration file will look as follows:

disable the Term extraction

active no

3.2 Term Extraction Output File Name

In the following file name example, "12200" is the port that the item processor runs on, "myhost" is the host name, and "1228825794.17" is the time stamp:

12200_myhost_1228825794.17.out.gz

3.3 Term Extraction Output File Content

These are three example lines from the file named 12200_myhost_1228825794.17.out.gz after decompression of the file:

en aaa 12

en baz 7

de bar 4

Here, the word "aaa" occurred 12 times and the word "baz" occurred 7 times in all items from the current item batch identified as English. The word "bar" occurred 4 times in all items of the current item batch identified as German.

4 Security Considerations

None.

5 Appendix A: Product Behavior

The information in this specification is applicable to the following Microsoft products or supplemental software. References to product versions include released service packs:

- Microsoft® FAST™ Search Server 2010

Exceptions, if any, are noted below. If a service pack or Quick Fix Engineering (QFE) number appears with the product version, behavior changed in that service pack or QFE. The new behavior also applies to subsequent service packs of the product unless otherwise specified. If a product edition appears with the product version, behavior is different in that product edition.

Unless otherwise specified, any statement of optional behavior in this specification that is prescribed using the terms SHOULD or SHOULD NOT implies product behavior in accordance with the SHOULD or SHOULD NOT prescription. Unless otherwise specified, the term MAY implies that the product does not follow the prescription.

6 Change Tracking

No table of changes is available. The document is either new or has had no changes since its last release.

7 Index