

## Documentation for:

- **Stage Two Parsing for the LoughranMcDonald\_MasterDictionary**
- **10-X Document Dictionaries – Doc\_Dict\_2011.csv**
- **10-X Document Data – Doc\_Data\_10X\_2011.csv**

**Author:** Professor Bill McDonald  
Mendoza College of Business  
University of Notre Dame  
Notre Dame, IN 46556  
[mcdonald.1@nd.edu](mailto:mcdonald.1@nd.edu)

## Local Notes:

### Updates

1. Delete word count columns from current dictionary.
2. Save results as new csv dictionary file.
3. Run  
N:\Research\Natural\_Language\_Processing\Programs\Build\_Dictionary\LookForNewWords\ (change input for load master dictionary)
4. Edit PotentialNewWords\_... files for potential additions. (Have considered those with frequency $\geq$ 100)
5. Run  
\Research\Natural\_Language\_Processing\Programs\Build\_Dictionary\Build\_10X\_MasterAndDoc\_Dictionary on compressed files (\Research\Edgar\Programs\Compress\_10X)
6. Removed "CASUALTY" 10/3/2012
7. Removed "CASUALTIES" 4/23/2013 (not in web version – will update with 2012 dictionary)

## Overview

In a second stage of parsing we generate the LoughranMcDonald\_MasterDictionary based on all Form 10-X variants filed with on the SEC's EDGAR website. In addition, document dictionaries and data summaries are created for each file parsed as we build the Master Dictionary. All files input into this stage of parsing are assumed to have been preprocessed based on the [Stage One Parsing process](#).

## Stage Two Parsing

The word counts and other statistics associated with each word in the LoughranMcDonald\_MasterDictionary are based on documents that are taken from the Stage

One parse and then parsed into word counts and other document attributes. The Stage Two parsing procedure is itemized below:

1. Delete all header information including header data added in the Stage One parse in addition to the SEC header.
2. Delete any residual HTML.
3. Eliminate all parens.
4. Delete end-of-line hyphens.
5. Replace hyphens preceding titlecase words with space.
6. Delete names preceded by titles, camel case words, obvious geographic proper nouns, and unambiguous proper nouns.
7. Delete ticker symbol throughout document.
8. Delete "'s" and "s'" in possessives.
9. Delete phrase "Table of Contents". (Many times this appears as a link at the top of each page.)

### **Individual Document and Data Dictionaries**

As an artifact of parsing each 10-X document to create the Master Dictionary, we generate a document dictionary containing word counts for each filing and a separate document data file with sentiment counts and other document characteristics for each filing. This file is posted on the website.

#### Document Dictionaries

A single file contains a document dictionary record for each 10-X filing in each year. The file does not contain a header record. Each record is a single line with multiple delimiters:

CIK, filing\_date, form\_type, file\_name | word\_sequence\_number1:count1,  
word\_sequence\_number2:count2, ...

This file is approximately 10 gig in size. If you wish to obtain a copy of the file, please contact [mcdonald.1@nd.edu](mailto:mcdonald.1@nd.edu) and request the document dictionary file.

#### Document Data

Separately in this parsing process a file is created containing summary data for each filing. This file contains a header record with labels and is comma delimited. Each record reports:

1. CIK – the SEC Central Index Key
2. Filing\_date – the filing date (YYYYMMDD) for the form
3. Fiscal\_Year\_End – fiscal year end
4. Form\_type – the specific form type (e.g., 10-K, 10-K/A, 10-Q405, etc.)
5. File\_name—the local file name for the filing

6. SIC number—the four digit SIC reported in the header of the filing. If this number does not appear in the header, then the primary web page for all filings from that firm at EDGAR is parsed in an attempt to identify the SIC number. If all of these methods fail, an SIC of -99 is assigned.
7. Fama/French Industry—the Fama-French 48 industry classification based on the SIC number. All missing SIC's are assigned to the miscellaneous category.
8. Number of words—the count of all words, where a word is any token appearing in the Master Dictionary.
9. Number of unique words--the number of words occurring at least once in the document.
10. A sequence of sentiment counts—negative, positive, uncertainty, litigious, weak modal, strong modal,constraining.
11. Negation—a count of cases where negation occurs within four or fewer words from a word identified as positive. Negation words are (“no, not, none, neither, never, nobody”, see Gunnel Totie, 1991, *Negation in Speech and Writing*). Thus net positive words is the positive word count minus the count for Negation.

*Statistics derived from the Stage One Parse*

12. GrossFileSize—the total number of characters in the original filing.
13. NetFileSize—the total number of characters in the filing after the Stage One Parse.
14. ASCIIEncoded Characters – the total number of ASCII Encoded characters (e.g., &#amp;#38;);
15. HTMLChars—the total number of characters attributable to HTML encoding.
16. XBRLChars—the total number of characters attributable to XBRL encoding.
17. TableChars—the total number of characters attributable to tables that have been removed in the Stage One Parse.