Microsoft Fuzzy Lookup Add-In for Excel

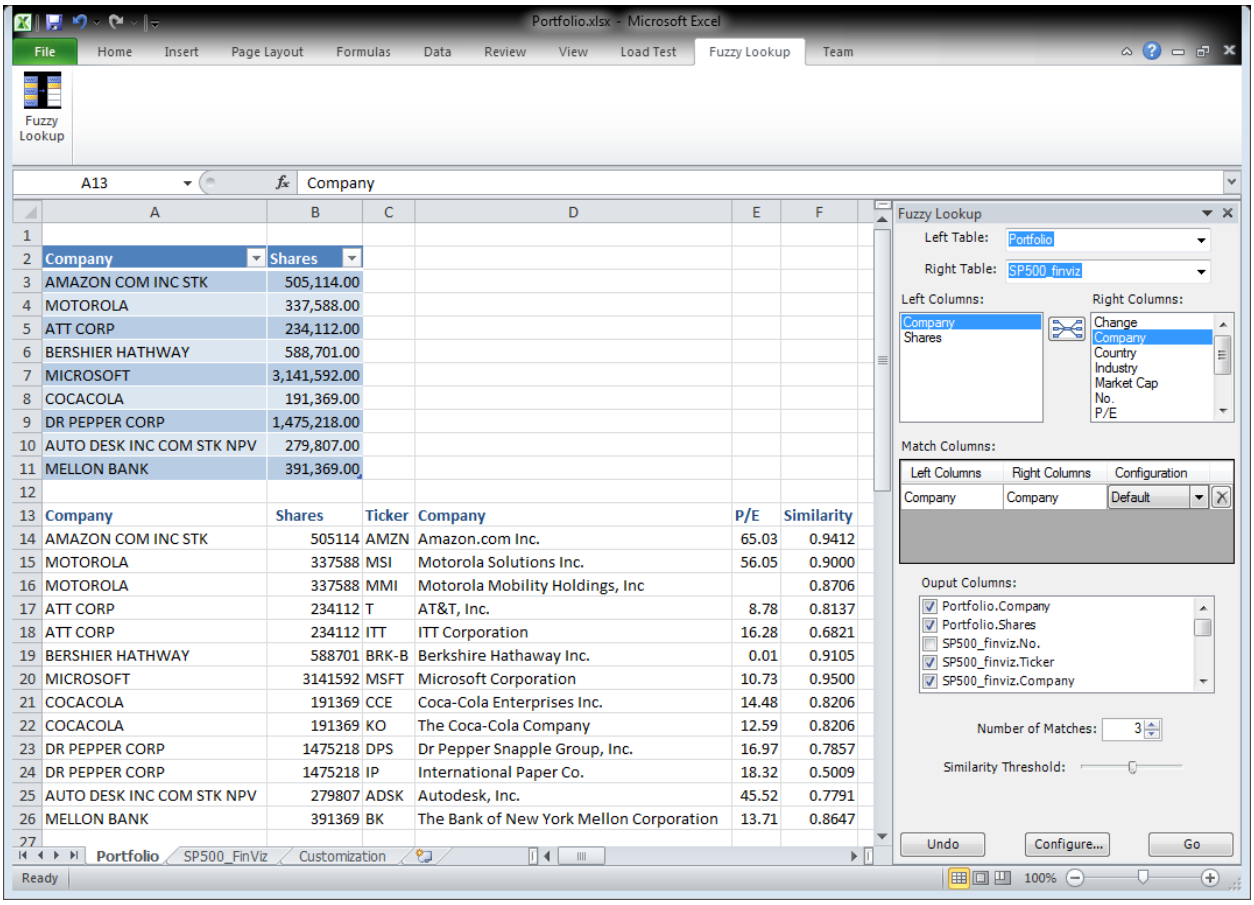
Introduction

A challenging problem in data management is that the same entity may be represented in multiple ways throughout the dataset. For instance, customer “Andy Hill” might also be present as “Mr. Andrew Hill” or “Hill, Andrew R.”. Variations can result from merging independent data sources, spelling mistakes, inconsistent naming conventions and abbreviations, or records with additional/missing information.

Fuzzy Lookup technology, developed by Microsoft Research, allows you to quickly identify data records which are textually similar. You can identify fuzzy duplicates within a single table or perform a fuzzy join between two different tables. The default configuration works well for a wide variety of data, but the matching may also be customized for specific domains.

Installation

See the ReadMe.txt file for installation instructions. Once installed, you should see a FuzzyLookup tab on the Excel Ribbon bar.



Company	Shares
AMAZON COM INC STK	505,114.00
MOTOROLA	337,588.00
ATT CORP	234,112.00
BERSHIER HATHWAY	588,701.00
MICROSOFT	3,141,592.00
COCACOLA	191,369.00
DR PEPPER CORP	1,475,218.00
AUTO DESK INC COM STK NPV	279,807.00
MELLON BANK	391,369.00

Company	Shares	Ticker	Company	P/E	Similarity
AMAZON COM INC STK	505114	AMZN	Amazon.com Inc.	65.03	0.9412
MOTOROLA	337588	MSI	Motorola Solutions Inc.	56.05	0.9000
MOTOROLA	337588	MMI	Motorola Mobility Holdings, Inc		0.8706
ATT CORP	234112	T	AT&T, Inc.	8.78	0.8137
ATT CORP	234112	ITT	ITT Corporation	16.28	0.6821
BERSHIER HATHWAY	588701	BRK-B	Berkshire Hathaway Inc.	0.01	0.9105
MICROSOFT	3141592	MSFT	Microsoft Corporation	10.73	0.9500
COCACOLA	191369	CCE	Coca-Cola Enterprises Inc.	14.48	0.8206
COCACOLA	191369	KO	The Coca-Cola Company	12.59	0.8206
DR PEPPER CORP	1475218	DPS	Dr Pepper Snapple Group, Inc.	16.97	0.7857
DR PEPPER CORP	1475218	IP	International Paper Co.	18.32	0.5009
AUTO DESK INC COM STK NPV	279807	ADSK	Autodesk, Inc.	45.52	0.7791
MELLON BANK	391369	BK	The Bank of New York Mellon Corporation	13.71	0.8647

Portfolio Sample

This section describes how to use the Fuzzy Lookup Add-In for Excel with the spreadsheet **Portfolio.xlsx**.

Imagine you have a stock portfolio described by two columns **Company** and **Shares** and that you are interested in computing the average price/earnings (P/E) ratio of the companies in the portfolio. To do this, you need to join your portfolio table with another table containing P/E ratios. The spreadsheet contains a second tab called SP500 which contains company data imported from the stock screener at the http://finviz.com website. Looking at the data, one can see that an exact join on the Company columns of the two tables would fail as the string representations of the companies differ (e.g., "AMAZON COM INC STK" and "Amazon.com Inc.").

A fuzzy join of the two tables can be performed as follows:

1. Turn the each data range into an Excel table by selecting a region and pressing **CTRL-L**. You can assign a name to the table clicking on it and selecting the Design tab in the Excel ribbon.
2. Open the Fuzzy Lookup pane by clicking on the Fuzzy Lookup button in the Fuzzy Lookup tab of the Excel ribbon.
3. Pick the left and right tables from the drop down menus. Matching rows from the right table will be returned for each row in the left table.
4. Select the columns to match on. If the two tables share one or more column names in common, a default join will already have been added. If you wish to match on different columns, first delete the existing join by pressing the "X" button on the join row in the Match Columns table. To create a new column binding, select one or more columns from each table (multiple columns may be selected by holding down SHIFT or CTRL and click on the column names). Next, press the button in between the two lists of columns to add a row to the Match Columns table.
5. Select one or more output columns to be output for each match.
6. Select the maximum number of matches to be returned for each left row.
7. Set the similarity threshold. All matches returned must have a similarity greater than or equal to this value.
8. Move the current cell selected in the Excel spreadsheet to an empty cell which has empty space to the right and below it. The Fuzzy Lookup matches will be output starting at this cell.
9. Press the "Go" button to perform the match.

One should see the results as indicated in the screenshot above. Notice that each returned match includes a similarity score indicating how close the two records are. 1.0 means an exact match while lower scores indicate less similarity.

Note that Fuzzy Lookup can also be used to identify matches in a single table by setting the left and right tables to be the same.

Advanced Concepts

Fuzzy Lookup technology is based upon a very simple, yet flexible measure of similarity between two records.

Jaccard similarity

Fuzzy Lookup uses Jaccard similarity, which is defined as the size of the set intersection divided by the size of the set union for two sets of objects. For example, the sets {a, b, c} and {a, c, d} have a Jaccard similarity of 2/4 = 0.5 because the intersection is {a, c} and the union is {a, b, c, d}. The more that the two sets have in common, the closer the Jaccard similarity will be to 1.0.

Weighted Jaccard similarity and tokenization of records

With Fuzzy Lookup, you can assign weights to each item in a set and define the weighted Jaccard similarity as the total weight of the intersection divided by the total weight of the union. For the weighted sets {(a, 2), (b, 5), (c, 3)}, {(a, 2), (c, 3), (d, 7)}, the weighted Jaccard similariyt is (2 + 3)/(2 + 3 + 5 +7) = 5/17 = .294.

Because Jaccard similarity is defined over sets, Fuzzy Lookup must first convert data records to sets before it calculates the Jaccard similarity. Fuzzy Lookup converts the data to sets using a Tokenizer. For example, the record {“Jesper Aaberg”, “4567 Main Street”} might be tokenized into the set, {“ Jesper”, “Aaberg”, “4567”, “Main”, “Street”}. The default tokenizer is for English text, but one may change the LocaleId property in Configure=>Global Settings to specify tokenizers for other languages.

Token weighting

Because not all tokens are of equal importance, Fuzzy Lookup assigns weights to tokens. Tokens are assigned high weights if they occur infrequently in a sample of records and low weights if they occur frequently. For example, frequent words such as “Corporation” might be given lower weight, while less frequent words such as “Abracadabra” might be given a higher weight. One may override the default token weights by supplying their own table of token weights.

Transformations

Transformations greatly increase the power of Jaccard similarity by allowing tokens to be converted from one string to another. For instance, one might know that the name “Bob” can be converted to “Robert”; that “USA” is the same as “United States”; or that “Missispi” is a misspelling of “Mississippi”. There are many classes of such transformations that Fuzzy Lookup handles automatically such as spelling mistakes (using Edit Transformations described below), string prefixes, and string merge/split operations. You can also specify a table containing your own custom transformations.

Jaccard similarity under transformations

The Jaccard similarity under transformations is the maximum Jaccard similarity between any two transformations of each set. Given a set of transformation rules, all possible transformations of the set are considered. For example, for the sets {a, b, c} and {a, c, d} and the transformation rules {b=>d, d=>e},

the Jaccard similarity is computed as follows: Variations of {a, b, c}: {a, b, c}, {a, d, c} Variations of {a, c, d}: {a, c, d}, {a, c, e} Maximum Jaccard similarity between all pairs: J({a, b, c}, {a, c, d}) = 2/4 = 0.5 J({a, b, c}, {a, c, e}) = 2/4 = 0.5 J({a, d, c}, {a, c, d}) = 3/3 = 1.0 J({a, d, c}, {a, c, e}) = 2/4 = 0.5 The maximum is 1.0. Note: Weghted Jaccard similiary under transformations is simply the maximum weighted Jaccard similarity across all pairs of transformed sets.

Edit distance

Edit distance is the total number of character insertions, deletions, or substitutions that it takes to convert one string to another. For example, the edit distance between “misissipi” and “mississippi” is 2 because two character insertions are required. One of the transformation providers that’s included with Fuzzy Lookup is the EditTransformationProvider, which generates specific transformations for each input record and creates a transformation from the token to all words in its dictionary that are within a given edit distance. The normalized edit distance is the edit distance divided by the length of the input string. In the previous example, the normalized edit distance is 2/9 = .222.

Technical resources

For more technical details on Fuzzy Lookup, see the following resources:

Microsoft Research Data Cleaning Project

Transformation-based Framework for Record Matching

Efficient Exact Set-Similarity Joins