What is Fuzzy Matching?

Why CPAs Should Know About This Powerful Tool

By Shivam Arora, CPA

In accounting, we regularly encounter situations where work is being performed manually and there is substantial scope of automation. For example, a CPA was downloading sales tax permits for their client's vendors to perform scoping for potential tax refunds. After downloading a permit, they would move it manually to the respective vendor's folder.

This process, they complained, took hours of their time, especially since multiple state permits had to be downloaded per vendor. However, the even more frustrating fact was that all file names "almost" contained their vendor's name, with omissions of letters and additions of special characters that did not seem to follow any one pattern.

This is a common problem. There are many business cases where practitioners spend time over text matching tasks that are intuitively obvious but do not follow a pattern and therefore, must be performed manually. Fortunately, solutions exist, such as fuzzy matching.

Fuzzy Matching

Fuzzy matching encompasses an umbrella of statistical techniques that compare and match approximately equal strings. These techniques employ statistical rules to arrive at a relative degree of truth on the similarity between two strings, in contrast to a Boolean approach, which uses a separate, hard-coded format for each task to provide a Yes/No answer.

The concept of fuzzy matching is analogous to the substance over form principle in accounting. It's the same reason why passthrough entities do not pay income tax even though they are technically separate legal entities to their owners or why the IRS sometimes classifies unusually large salary payments to owners as dividends even though they are technically salary payments.

Fuzzy matching, like these cases, gives preference to substantial equivalence between strings over their technical form.

The Underlying Logic

There are several approaches available to fuzzy match data, but I'm going to go briefly over the most common one. The Levenshtein Distance (LD) is commonly used to establish similarity between two strings. It is the minimum number of single character edits that are required for changing either of the two strings into the other. An edit can refer to a character's insertion, deletion or replacement. Consider the following strings:

charlie_vendor
$charl_vndr$

Assume that the naming convention of vendors in a system is "{name}_vendor." It is intuitively obvious that the file downloaded is for the vendor Charlie. However, unless all downloaded files follow the same naming convention as above, a Boolean approach to matching will declare both strings unequal.

When I run LD-based fuzzy matching (LDFM) in Python, I get an LD of 6. This means that the shortest number of single-character changes to exactly match the file name and the vendor's name is 6. Converting it into the LD ratio (using a formula I will not delve into), I obtain .77.

What I now have is a quantified degree to which both strings are similar: my computer understands that both strings are about 77% similar. It still knows that they are

not equal; it has just established equivalence.

Applications of Fuzzy Matching in Accounting

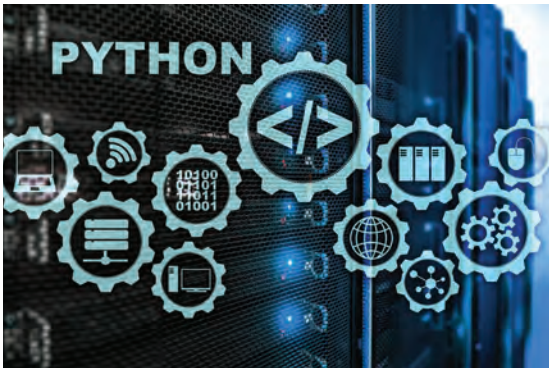
There can be several applications of fuzzy matching in accounting. A few of them follow.

File Renaming. As with the above case, fuzzy matching can be used to rename downloaded files and match them to their respective group. Names of files downloaded from the internet often contain either truncated text or unwanted characters.

Support Accounting Processes. Fuzzy matching can support accounting processes such as bank reconciliations, inventory tracking and evidence gathering for various types of audits.

Internal Controls. Fuzzy matching can detect duplicate AP payments with minor variations, compare purchase orders to deliver invoice/bill of lading and enforce data entry checks. In case of fraud, it can also aid in identifying matches across different databases or comparing fraudulent acts across different time periods.

Preprocessing for ML. With the arrival of artificial intelligence, organizations are increasingly utilizing machine learning (ML) techniques. A substantial amount of ML in the financial space occurs on data generated by



accounting systems. By facilitating preprocessing of data using fuzzy matching techniques, organizations can develop robust and accurate ML models.

A Coding Exercise

One can perform fuzzy matching in Excel (refer to the article "Excel: Fuzzy Matching" by Bill Jelen in *Strategic Finance* magazine). Unsurprisingly though, the functionality is extremely limited and there is little clarity on what technique is used. A better alternative may be the programming language Python.

Python is a high-level programming language that is general-purpose; it can be used to code for a wide variety of situations. The beauty of Python is that it is intuitive and relatively easy to learn, which is why it is used extensively in business. It hosts numerous libraries that are specifically designed for business-related tasks.

Case Examples

Consider the below as examples of how helpful fuzzy matching can be for accountants and auditors.

1. A CPA is performing a quarter-end bank reconciliation. There are 300+ entries on both sides. The CPA notices that transaction descriptions on bank statements are similar to those in the books, albeit with expected differences such as truncations, word order and unwanted characters. Using LDFM, the CPA can match 270 transaction descriptions between the bank and the books. The CPA also verifies that

the corresponding amounts across these transactions are equal.

The CPA now begins to reconcile the remaining few transactions on both sides. Using fuzzy matching has greatly reduced the manual workload.

2. A tax consultant is working on a reverse audit for one of their clients. The consultant must download sales/use tax permits for the client's 500+ vendors to ascertain the type of permits held in the relevant states.

For simplicity, assume that a single file contains all permits for one vendor. The consultant has an Excel file with a list of all vendors. Instead of manually linking each vendor to their corresponding permit, the consultant employs LDFM. This results in a >90% confidence match for 460 vendors.

After a cursory review of the matches to ensure accuracy, the consultant needs to only focus on the unmatched vendors for linking permits manually. If it takes 30 seconds for the consultant to browse through all permits to find the correct permit for each vendor and they have a code already available to perform LDFM, they have just reduced the task time by close to four hours.

3. One warehouse of a manufacturing company uses LDFM to compare raw materials ordered on a purchase order to those received and listed on the invoice. This helps the warehouse detect not only discrepancies between the quantity of items ordered, but also between their type.

Over the years, the warehouse has been able to reduce purchase return-related costs by up to 40% by refusing delivery of suboptimal orders. You can read the outstanding use case of fuzzy matching in fraud examination in an article written by Ehsanelahi in *Data Ladder* titled "Fuzzy Matching 101: Cleaning and Linking Messy Data."

A Powerful Tool

Given the nature of accounting work, fuzzy matching techniques can be a powerful tool in a CPA's arsenal.

Fuzzy matching can be performed in Excel but is much more powerful when performed in a programming language such as Python, which is an intuitive programming language that, in addition to core software development, has extensive use cases in a business setting.

As evident, it is not hard to follow most (if not all) aspects of the Python exercise above even without basic knowledge of the language. CPAs should consider learning a programming language to automate much of the manual tasks they perform.

About the Author: Shivam Arora, CPA, is a data scientist. Arora holds dual master's degrees: an MS in Accounting and an MS in Business Analytics. As an applied Artificial Intelligence (AI) consultant at one of the largest consulting firms in the world, Arora specializes in applied AI for accounting and finance. Research interests include financial modeling and statistical relationships in the financial markets, application of AI to accounting and Robotic Process Automation (RPA). Email shivam.arora@mavs.uta.edu.

Editor's Note: For more information on AI, check out TXCPA's CPE program Artificial Intelligence for Accounting and Financial Professionals in the Education section of the TXCPA website.