BIG DATA: SMALL STEPS WITH EXCEL

Ting Jie (TJ) Wang, Ph.D.
Associate Professor of Accounting
College of Business
Governors State University
Phone: 708-534-4965 / Fax: 708-534-8457
1 University Parkway
University Park, Illinois 60484
twang@govst.edu

INTRODUCTION

Big data have attracted a lot of attention from both accounting profession and academics in the past several years. For instance, KPMG is partnering with Ohio State University and Villanova University in developing a master accounting degree program in data analytics (Lynch, 2017). Deloitte is in alliance with Intermountain Healthcare around big data and data analytics (McCann, 2013). Accounting literature shows that big data gain an important role in accounting since it has impacts on how data for financial and managerial accounting are measured and represented (Moffitt and Varsarhelyi, 2013; Warren, Jr., Moffitt, and Byrnes, 2015).

There are many big data tools available such as Jaspersoft Business Intelligence Suite, Karmasphere Studio and Analyst, Pentaho Business Analytics, RapidMiner Studio, Skytree Server, Splunk, Tableau Desktop and Server, and Talend Open Studio, etc. However, many schools don't have resources or access to big data tools unlike schools with data analytics programs such as West Virginia University, Ohio State University, Villanova University, etc. Furthermore, many instructors may not have knowledge and/or skills to teach or use any of these big data tools, and/or may not have sufficient amount of time to cover this topic/application into their already full AIS courses.

Basically, big data consists of four different types of data: (1) structured data (e.g., data with a defined schema), (2) semi-structured data (e.g., data tagged with XML, and (3) unstructured data (e.g., text and video), and (4) multi-structured data (e.g., integrated data with different types and structural levels). Accounting related data can be found from many textual data sources (i.e., unstructured data), such as EDGAR (10-K & 10-Q), websites, social media, newspapers, professional magazines, etc.

According to the Excel team from Microsoft, Excel is able to handle some aspects of data analytics (Sharma and Ellis, 2016). The "Data Model" feature in Microsoft Excel provides support for large volumes of data with the 1M row limit per worksheet. Excel 2016 is built with Power Query capable of querying big data from large relational sources, such as Software as a Service (e.g., Microsoft Dynamics CRM, DalesForce), Hadoop Distributed File System, etc. Excel Power Query is also capable of transforming data into a clean set of data to work with. In addition, Power Query can also handle semi-structured data with a built-in support for extracting structure out of a certain formatted data. Furthermore, Excel Data Model can embrace the Tables, Columns, Relationships as first-class objects.

There are three patterns of applying Excel with big data: (1) Import data into Excel, (2) Live query of an external source, and (3) Export from an application to Excel (Sharma and Ellis, 2016). When importing data into Excel, there are a few challenges: (1) querying big data, (2) transforming data, (3) handling large data sources, (4) handling semi-structured data, and (5) handling large volumes of data in Excel.

The purpose of this paper is to present a big data project/assignment which will allow students to gain some fundamental skills and experience in data analytics as well as to help students advance their Excel skills. This project is easily adaptable for AIS instructors. This paper presents a certain part of the project created and used by the authors, which includes (1) Gathering and

Arranging Data, (2) Summarizing Data Using PivotTable, and (3) Presenting Data Using PivotChart. The basic tool that students will be needed for this project is Microsoft Excel. In this paper, we will focus on the tasks involving Gathering and Arranging Data (i.e., transforming unstructured data) since data analytics professionals spend most of their time on these tasks. Survey shows that data analytics professionals spend about 50 percent to 90 percent of their time cleaning the data (Taft, 2015).

PROJECT LEARNING OBJECTIVES

The goals of this project are to allow students to acquire skills in using advanced Excel functions to manipulate and transform unstructured data (i.e., tasks in data analytics), and to gain experience with Excel features such as PivotTable and PivotChart. Students will first learn to form logistics conceptually in gathering and transforming unstructured data, and then perform Excel functions in importing, extracting, and transforming data. The learning objectives of this project are that students after completing the project are able to:

- Conceptually plan and design a logistic for data extraction and transformation,
- Use Excel features such as Import, PivotTable, and PivotChart,
- Apply Excel functions such as LEFT, RIGHT, MID, SEARCH/FIND, LEN, IF, CONCATENATE, VLOOKUP, etc.,
- Employ nested functions in Excel to manipulate and transform unstructured data, and
- Create a formal business memo reporting the plan.

PROJECT REQUIREMENTS

This big data project requires students to deal with unstructured data, i.e., text, and handle tasks such as importing text data into Excel, extracting and merging data, and transforming data into certain layouts and formatting in order for further data analyses. There are three parts in this project: (1) Gathering and Arranging Data, (2) Summarizing Data Using PivotTable, and (3) Presenting Data Using PivotChart. Students will be given monthly sales records in four different text files with different data formats and contents. For the first part of the project, students are required to compose two data worksheets/sets from these four text files: one that contains a complete sales records and the other a master customer list. These two data worksheets/sets will then be used in the second and third parts of the project. Instructors can easily add or alter the requirements and/or specify the usage of specific Excel tools. Students may be impulsive to implement whatever comes to their minds right away, but it is important for students to come up with a conceptual plan for gathering, merging and transforming the data before they actually work on the data. Remind students that the primary purpose here is to come up with a standard operating procedure (SOP) so that the company can compose the data and reports every month. This paper will focus on the first part of the project, i.e., Gathering and Arranging Data. Information on PivotTable and PivotChart, i.e., the second and third part of the project, can be easily found in the literatures (Collins, 2017; Filzen and Simkin, 2015). Some teaching notes are presented in Appendix A. upon request, the authors will provide the actual project, sample data files, and suggested solution for this project.

GATHERING AND ARRANGING DATA

The first part of the project requires students to gather and arrange data for further data analyses in the second and third parts of the project. Students are required to compose a complete/merged sales records and a master customer list with complete/no missing data in the cells. Addresses should be separated into Street Address, City, State and Zip Code so that further analyses on geographic segmentation can be performed later. Four text files are provided to students as shown in Figures 1-4 below:

Figure 1. Adam.txt (Name, Address, Gender, and Sales)

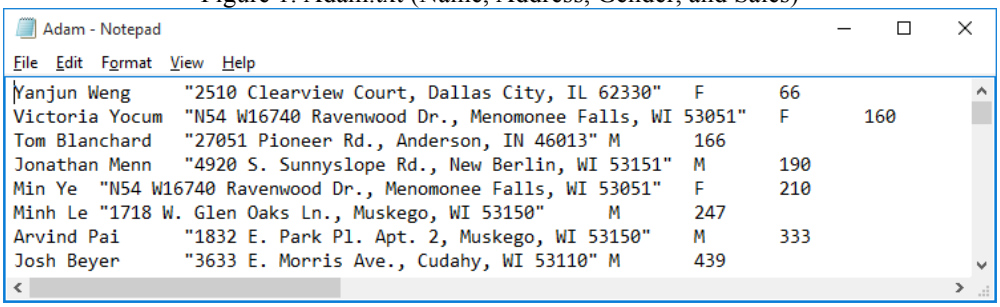


Figure 2. Becky.txt (Name, Address, Gender, and Sales)

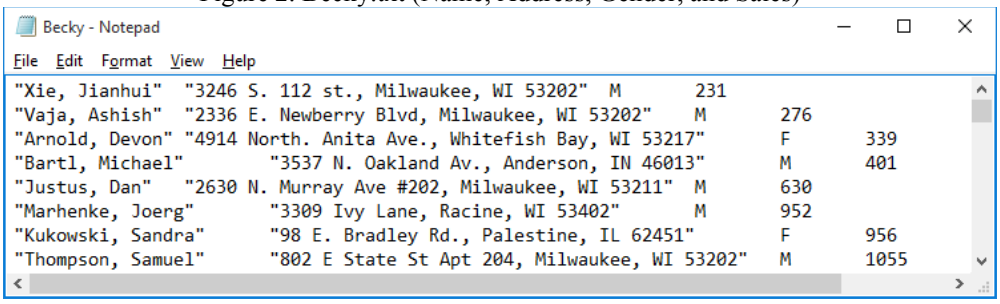


Figure 3. Chris.txt (Name, Telephone, and Sales)

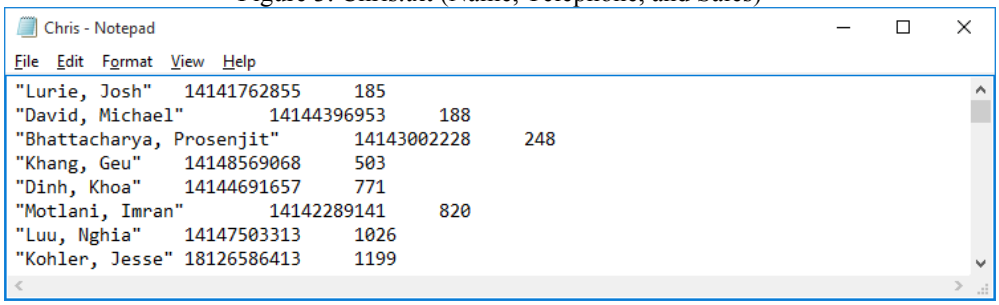
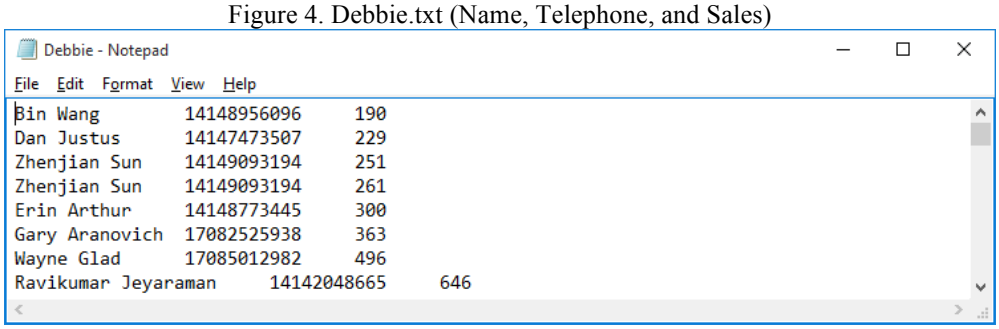


Figure 4. Debbie.txt (Name, Telephone, and Sales)



The first two files, i.e., Adam.txt and Becky.txt, provide information about customers' name, address, gender, and sales amount for the month. The formats of the data fields in these two files are the same except for the "Name" field. In the Adam.txt data file, the "Name" follows First Name, a space, and Last Name. In the Becky.txt data file, it follows Last Name, a comma, a space, and First Name.

The last two files, i.e., Chris.txt and Debbie.txt, provide customers' name, telephone number, and sales amount. The formats of the data fields are the same except for the "Name" field. In the Chris.txt data file, the "Name" follows Last Name, a comma, a space, and First Name. In the Debbie.txt data file, it follows First Name, a space, and Last Name.

These text files contain monthly sales records from four sales representatives, i.e., Adam, Becky, Chris, and Debbie. Both Adam and Becky recorded customers' name, address, gender, and sale amount information except telephone information. However, both Chris and Debbie recorded customers' name, telephone, and sale amount information except address and gender. Customers in these four files are not completely exclusive. Some customers may have dealt with different sales representatives during the month, i.e., appear in more than one file. We will assume that customers' names are unique. That is, customers with the same names are the same customers in this project.

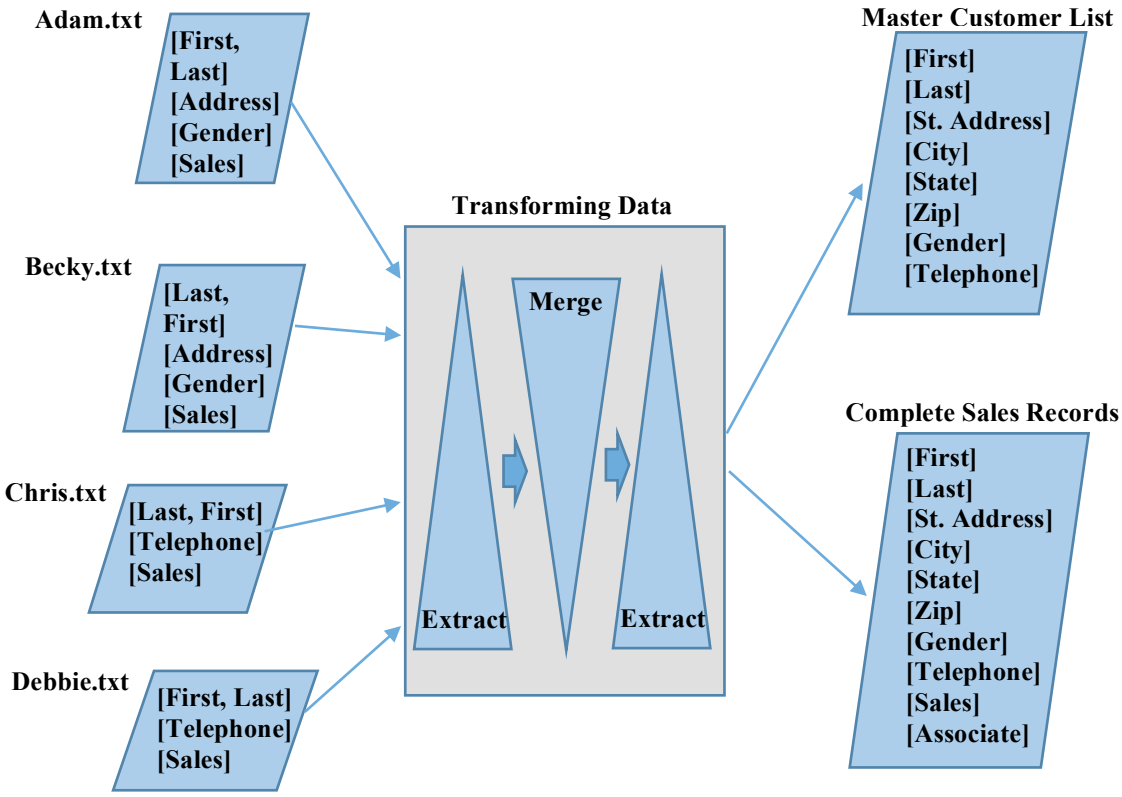
The most crucial part of the project is to help students come up with a plan (i.e., steps) to extract, merge, and transform data into complete sales records and a master customer list before they actually implement the plan. Figure 5 depicts the process for the gathering and arranging data in this part of the project. Instructors can use this diagram to help students walk through the process conceptually. For example, the following plan/steps can be recommended:

1. Importing four text data files individually into Excel worksheets,
2. Using customer's name as a "key" for identification purpose (after extracting and arranging the name) so that a master customer list can be composed (duplicates should be removed),
3. Creating a sales associate's column before merging these worksheets (it will not be possible after merging the files),
4. Merging four worksheets (i.e., SalesRecords—a draft list of Sales Records with some cells missing address, gender, or telephone),
5. Creating two additional worksheets with exactly the same as the merged worksheet from the previous step (copy and paste); one as a draft master list for the customer (CustomerMaster) and the other as a transition file for customers with telephone

information (CustomerTel). Remove rows with empty address cells in CustomerMaster file and rows with empty telephone cells in CustomerTel. Finally, remove duplicates in them based on the identification column,

6. Looking up telephone information for CustomerMaster from CustomerTel (using VLOOKUP function)—this is the complete master customer list,
7. Looking up address and telephone information for SalesRecords from CustomerMaster,
8. Extracting and Separating address information into Street Address, City, State, and Zip Code in SalesRecords and CustomerMaster.

Figure 5. Diagram of the Process for the Gathering and Arranging Data



The order of certain steps from the above process may be mandatory but others may not. For example, creating a sales associate's column must be done before the merge of the four worksheets. Otherwise, the identity of the sales associates will be lost once these worksheets are merged. A "key" column using customer's full name can be created before or after the merge of the four worksheets. It does not affect the process.

In order to allow students the opportunities to learn some text related functions in Excel, we do not allow students to use the "Text to Columns" feature in Excel. This way, students get an opportunity to look into Excel functions for ways to handle unstructured data, such as LEFT, RIGHT, MID, SEARCH, LEN, etc.
Excel skills that students will be exposed to along with the process include but not limit to:

1. Importing data into Excel worksheet (Data > Get External Data > From Text),
2. Extracting, arranging, and transforming unstructured data with Excel functions such as LEFT/RIGHT/MID/SEARCH/LEN/CONCATENENT/IF,
3. Removing duplicates (Data > Remove Duplicates), and
4. Looking up information from other sources with VLOOKUP function

CONCLUSION

Big data gain an important role in accounting while information technology advances. There is no doubt that future accountants are expected to analyze companies' finances and financial activities with large databases and/or various data sources. Since many schools especially small schools may not have access to data analytics software, and/or many AIS instructors may not have knowledge of these data analytics software and/or time to incorporate them into their AIS courses, students are likely to miss out the learning opportunity for handling big data. However, the project presented in this paper serves as one of the first steps in helping accounting students advance their knowledge and experience in handling tasks involving unstructured data (i.e., text data). It exposes students to Excel functions in manipulating and transforming unstructured data and requires students to semanticize (i.e., give a meaning to) the functions that they apply especially with those nested functions, IF logical operator, and lookup functions. It also enhances students' skills and experience in using PivotTable and PivotChart features.

REFERENCE

Collins, J. Carlton. 2017. Data Mining Your General Ledger With Excel. *Journal of Accountancy*. January 1, 2017.

Filzen, Joshua J. and Mark G. Simkin. 2015. Audit Accounting Data Using Excel Pivot Tables: An Aging of Accounts Receivable Example. *Information Systems Audit and Control Association (ISACA)*. Volume 1, 2015.

Lynch, Elizabeth. 2017. Inaugural Class Selected for KPMG Masters Program at Ohio State and Villanova Universities' Business Schools. *KPMG Press Releases*. February 7, 2017. (https://home.kpmg.com/us/en/home/media/press-releases/2017/02/inaugural-class-selected-for-kpmg-masters-program-at-ohio-state-and-villanova-universities-business-schools.html)

McCann, Erin. 2013. Deloitte, Intermountain Tap Big Data. *Healthcare IT News*. June 24, 2013. (http://www.healthcareitnews.com/news/deloitte-intermountain-tap-big-data)

Moffitt, Kevin C. and Miklos A.Vasarhelyi. 2013. AIS in an Age of Big Data. *Journal of Information Systems* 27(2): 1-19.

Sharma, Ashvini, and Charlie Ellis. 2016. Excel and Big Data. *Office Blog*. June 23, 2016. (https://blogs.office.com/2016/06/23/excel-and-big-data)

Taft, Darryl K. 2015. One-Third of BI Pros Spend Up to 90% of Time Cleaning Data. *eWEEK*. July 5, 2015. (http://www.eweek.com/database/one-third-of-bi-pros-spend-up-to-90-of-time-cleaning-data)

Warren, Jr. J. Donald, Moffitt, Kevin C., and Paul Byrnes. 2015. How Big Data Will Change Accounting. *Accounting Horizons* 29(2): 397-407.

APPENDIX A: TEACHING NOTES

Instructors should make sure that students are familiar with the following Excel functions before they start to do the project: *LEFT*, *RIGHT*, *MID*, *FIND*/*SEARCH*, *LEN*, *IF*, *CONCATENATE*, and *VLOOKUP*. Instructors can find examples for students from the Excel Help (i.e., F1 key), Google, and/or Youtube. It would be the best if instructors can help student semanticize (i.e., give a meaning to) the functions and processes, not just blindly apply but also make meaning out of these functions and processes.

I. Understand the Functions

Definitions and common usages of the Excel functions below can be easily found in Excel (click on the drop box of AutoSum in the Editing from the Ribbon on the top of the Excel software interface). Instructors can come up with some simple exercises to help students familiarize with these functions. For example:

- *FIND*(*find_text,within_text,start_num*): *FIND* returns the starting position of one text string within another text string. *FIND* is case-sensitive. For example, *FIND*("t","Microsoft Excel",1) should return a numerical value of 9. Be aware that quotation marks are used around a text value in Excel.
- *IF*(*logical_text,value_if_true,value_if_false*): *IF* checks whether a condition is met, and returns one value if it is TRUE and another value if it is FALSE. For example, *IF*(5>7, "Greater","Lesser") will return a text value of "Lesser". Be aware that the returned value can be either a text or numeric value.
- *LEFT*(*text,num_chars*): *LEFT* returns the specified number of characters from the start of a text string. For example, *LEFT*("Microsoft Excel",9) will return a text value of "Microsoft".
- *LEN*(*text*): *LEN* returns the number of characters in a text string. For example, *LEN*("Armstrong") will return a numerical value of 9.
- *MID*(*text,start_num,num_chars*): *MID* returns the characters from the middle of a text string, given a starting position and length. For example, *MID*("Microsoft Excel Functions",11,5) will return a text value of "Excel".
- *RIGHT*(*text,num_chars*): *RIGHT* returns the specified number of characters from the end of a text string. For example, *RIGHT*("Microsoft Excel Functions",9) will return a text value of "Functions".
- *SEARCH*(*find_text,within_text,start_num*): *SEARCH* returns the starting position of a specific character or text string, reading left to right (not case-sensitive). For example, *SEARCH*("o","Microsoft Excel Functions",9) will return a numerical value of 23.
- *CONCATENATE*(*text1,text2,…*): *CONCATENATE* joins several text strings into one text string. For example, *CONCATENATE*("Microsoft","Excel","Functions") will return a text string of "Microsoft Excel Functions".
- *VLOOKUP*(*lookup_value,table_array,col_index_num,range_lookup*): *VLOOKUP* looks for a value in the leftmost column of a table, and then returns a value in the same row from a column you specify. By default, the table must be sorted in an ascending order. For example, *VLOOKUP(77,J1:K5,2)* will return a text value of C, given that J1 contains 0, J2 60, J3, 70, J4 80, and J5 90, and K1 contains F, K2 D, K3 C, K4 B, and K5 A.

II. Semanticize the Functions

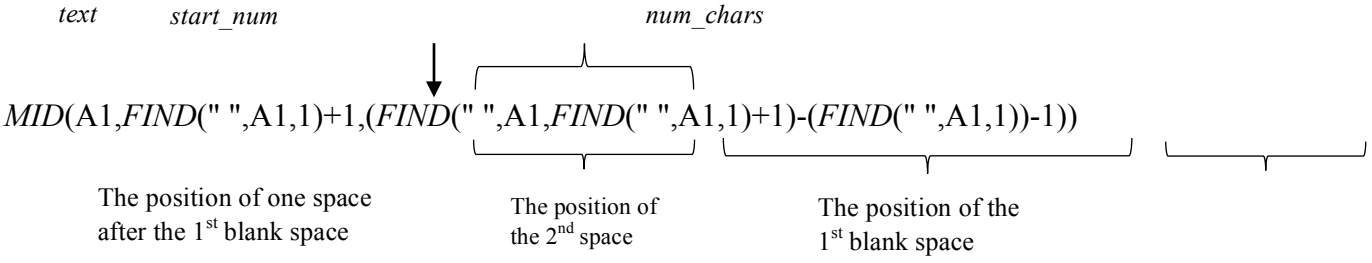
Instructors can also help students enhance their understanding of Excel functions by helping them semanticize (i.e., make a meaning out of) these functions. For example, what will be the function to extract the first/second "word" from a text string? Students by now should know that *FIND*, *SEARCH*, and *LEN* functions return a numerical value while *LEFT*, *MID*, and *RIGHT* return a text string. These functions can be used within each other as long as the returned value of a function is comparable with the parameter of the function (i.e., nested functions). For example, A1 cell contains a text string of "Microsoft Excel Functions" (or any text string with at least two blank spaces between words or three words in it). *LEFT*(A1,9) returns a text value of "Microsoft". *FIND*(" ",A1,1) returns a numerical value of 10, which is the starting position of the first blank space from the left. So, using these two functions within each other, we can obtain the first word from a text string: *LEFT*(A1, *FIND*(" ",A1,1)-1)

> since *FIND*(" ",A1,1)-1, the *num_chars* parameter/argument of the *LEFT* function, will return a numerical value of 9, i.e., the character position before the first blank space from the left side or the length of the first word from the text string.

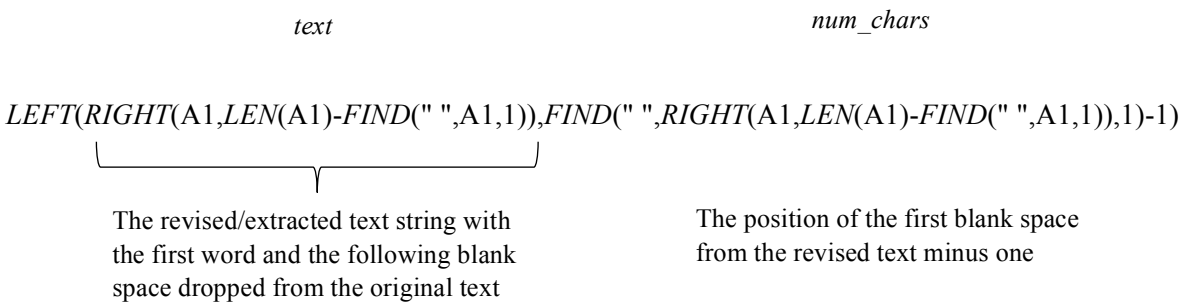
In the English language, blank spaces are used to separate words. As a result, we are able to use a blank space as an indication for the beginning/ending of a word.

How do we obtain the second word? To obtain the second word, we will need to find the second blank space in the text string. For example, *FIND*(" ",A1,*FIND*(" ",A1,1)+1) will give us the position of the second blank space in A1. The *FIND* function used as the *start_num* parameter within another *FIND* function will return a position that goes by passing the first blank space since it adds one to the first blank space as the starting position for finding the second blank space. So, the function to extract the second word is: *MID*(A1,*FIND*(" ",A1,1)+1,(*FIND*(" ",A1,*FIND*(" ",A1,1)+1)-(*FIND*(" ",A1,1)+1))). Of course, there may be other ways to extract the second word. For example, *LEFT*(*RIGHT*(A1,*LEN*(A1)-*FIND*(" ",A1,1)),*FIND*(" ",A1,*FIND*(" ",A1,1)+1)-(*FIND*(" ",A1,1)+1)) can also extract the second word from the text string.

Let's decompose these functions in order to semanticize them. The *MID* function has three parameters: first is the *text*, second the starting position or *start_num*, and third the number of characters or *num_chars*. In this case, A1 contains the text (i.e., "Microsoft Excel Functions"), the starting position, *start_num*, is the position of the first blank space plus one, and the number of characters or the length of the second word should be the distance between the position of the space after the first blank space and the position of the space before the second blank space (or the difference between the positions of the first and second blank spaces minus one; the concept is the same).



Another way to extract the second word from a text string is to first drop the first word and the following blank space from the original text and then take the first word from the revised/extracted text string as the function shown below:



Instructors may require students to provide written explanations for the functions that they apply in order to be sure that they really semanticize the functions. Remind students about possible blank spaces undetected when they apply nested functions since blank spaces at the beginning or the end of the extracted text strings may not be easily visualized in the outcome from the screen. A good way to double check for blank spaces at the beginning or ending of a text string is to use a *LEN* function to verify the number of characters in the outcome.

III. Conceptualize the Process

Once students are exposed to the functions and how to semanticize them as mentioned above, they probably would jump into the project right away without giving any thoughts to the whole process of obtaining the results, i.e., complete sales records and master customer list. Although students may be able to complete the project using a trial and error approach, they may have to go back and forth a few times to redo certain steps due to their lack of planning. Sometimes they may have to start from the beginning depending on what they have done or not done/skipped. The authors require students to submit their plan in writing as one of the requirements for the project. Students may work on coming up with a plan and implementation simultaneously. However, having a plan for gathering, merging, and transforming data will help them prevent from time wasted in going back and forth in the process. For example, students should first Import Data into Excel worksheets, then create a KEY data field for each row so that the KEY can be used for the merge of the worksheets later on (in this case, this KEY is the identification for the customers). Below is an example of a plan and design for the process. Be aware there are many ways to handle this task. Even with the same steps/procedures, the sequences of some steps may be fixed and others not.

Recommended Procedures:

1. Import Data into Excel Worksheets (Data > Get External Data > From Text)
   - Clean Up & Arrange:
     i. Importing 4 text data files into individual worksheets
     ii. Making sure each data field is placed into a column during the import
     iii. Creating a “Sales Associate” column and entering associate’s name
     iv. Separating “First Name” and “Last Name” using Excel functions

v. Separating “Street Address”, “City”, “State”, and “Zip Code” using Excel functions (Note: this can also be performed after sales records are all merged)

2. Create a KEY attribute/column using customer’s full name (e.g., First+Last)
   - Clean Up & Arrange:
     i. Removing blank spaces from the beginning and ending of the “First Name” and “Last Name” using Excel functions
     ii. Creating an identification field, i.e., “ID”, by merging the revised “First Name” and “Last Name”
3. Create a draft Sales Worksheet (**SalesRecords**) by Combining 4 arranged worksheets (Note: this step can be performed before Step #2)
4. Create a Master Customer List
   - Clean Up & Arrange:
     i. Creating two worksheets by copying the draft Sales worksheet (**SalesRecords**); one for addresses, i.e. **CustomerMaster**, and the other telephone **CustomerTel**
     ii. Cleaning up the “**CustomerMaster**” worksheet:
         1. Deleting records with blank street address
         2. Removing duplicates in the “ID”

         (Note: sequences do not matter in this step)
     iii. Cleaning up the “**CustomerTel**” worksheet:
         1. Deleting records with blank telephone numbers
         2. Deleting all other columns except “ID” and “Telephone”
         3. Placing “ID” column on the left of “Telephone” column
         4. Removing duplicates in the “ID”

         (Note: sequences do not matter in this step)
     iv. Updating telephone information on **CustomerMaster**
         1. Using a VLOOKUP function to extract “Telephone” from **CustomerTel** worksheet

         (Note: use the *range_lookup* parameter to find an exact match; assign a “False” value to it for an exact match)
5. Update Sales Records
   - Clean Up & Arrange:
     i. Deleting all other columns except “ID”, “First Name”, “Last Name”, “Sales Associate”, and “Sales Amount”
     ii. Updating “Street Address”, “City”, “State”, “Zip Code”, “Gender” and “Telephone” from **CustomerMaster** using VLOOKUP function

Students may want to use “copy and paste” for Steps 4.iv and 5.ii to fill data in the empty cells. Let students be aware that they may have to perform this process on a regular monthly basis (and, will they “copy and paste” when there are ten times of the amount of data next month?). So, it will be better if they can use functions instead.

IV. Account for the Records (Check Figures)

The following are some results from the process. Instructors can provide these to students as check figures if you want. Adam.txt contains 247 records; Becky.txt contains 152 records; Chris.txt contains 129 records; Debbie.txt contains 277 records; Sales records should be 805 and Master Customer List should be 201 records. All 201 customers as well as sales records have complete address and telephone information.