Spreadsheet functions

Text data and analysis

Overview

Introduction

Text data and analysis functions in spreadsheets refer to a set of **built-in functions** that allows users to **manipulate and analyse text data within cells**. Advantages of using these functions include:

01. **Time-saving**

They allow quick and easy manipulation of large datasets, which saves time compared to manually editing each cell.

02. **Consistency**

They ensure consistency in formatting and cleaning up text data, which reduces errors and increases accuracy and reliability.

03. **Scalability**

They can be used on large datasets, making it easier to analyse and visualise the data.

04. **Flexibility**

They allow users to extract specific information or manipulate text data in various ways, depending on their needs. This flexibility allows users to create customised solutions for different types of data.

05. **Increased productivity**

The ability to manipulate and analyse text data within a spreadsheet increases productivity, as users can quickly extract relevant information without having to switch between multiple programs.

Data overview

To investigate how spreadsheet functions can be used to analyse text data, we will use a **Tweets on climate change** dataset that has 100 rows and the following columns:

1. **ID**

A numeric string that is associated with and uniquely identifies a single Tweet within the dataset. It makes it possible to access and interact with a specific Tweet.

2. **Text**

An aggregated Tweet pertaining to climate change.

Dataset

	A	B
1	ID	Text
2	1028954403129184256	Gotta love the facts. https://t.co/bZ2G8AZuo9
3	1028954356572250112	RT @ToolangiForest: A great day of action for our message of "Dear Dan"! Toolangi community & friends joined together to respectfully ask @…
4	1028954497341480960	@jonkudelka Harvey Norman reckons climate change is bunkum because his mates who own coal companies need people to buy polluting stuff
5	1028954494133043200	RT @jayrosen_nyu: Why does skepticism about immigration walk hand in hand with skepticism about the science of climate change? I know we're…
6	1028954811511844864	RT @FranceinIreland: On 5th November we call all creative citizens w/ practical solutions to fight against #climatechange to join us for a…
7	1028954782457909250	RT @jayrosen_nyu: Why does skepticism about immigration walk hand in hand with skepticism about the science of climate change? I know we're…

Tweets on climate change

The SUBSTITUTE function

The **SUBSTITUTE** function is used to **replace a specific character or string of characters** in a cell with a different character or string.

SUBSTITUTE and **REGEXREPLACE** are similar but **SUBSTITUTE** is preferred over **REGEXREPLACE** when the text being replaced is in multiple columns.

`=SUBSTITUTE(text_to_search, search_for, replace_with, [occurrence_number])`

- `text_to_search` – The text within which to search and replace.

- `search_for` – The string to search for within `text_to_search`.
  `search_for` will match parts of words as well as whole words; therefore, a search for "vent" will also replace text within "eventual".

- `replace_with` – The string that will replace search_for.

- `occurrence_number` – [OPTIONAL] The instance of `search_for` within `text_to_search` to replace with `replace_with`. If `occurrence_number` is specified, only the indicated instance of `search_for` is replaced.

The SUBSTITUTE function

Example use:

Remove all **URLs** and **mentions** from all the Tweets.

- We will use **OR logic** since a Tweet can have both, either of the two, or neither of the two.

- The **REGEXEXTRACT** function will be used to identify and extract the **URLs** and **mentions** then the **SUBSTITUTE** function will be used to replace them with a **blank string**.

On Twitter, a **mention** is a way to **tag or reference another user** in a Tweet by including their **username** in the tweet. Mentions are commonly used to start a conversation with someone, to acknowledge someone in a Tweet, or to give credit to someone for their work.

Mentions are prepended with the "@" symbol.

Recall that the **pipe symbol (|)** represents the **OR** operator in **regular expressions.**

6

The SUBSTITUTE function

Example use:

- The regex expression to match URLs is **https/?:\/\/[^\s/$.?#].[^\s]*** while that of mentions is **\B@\w{1,15}**.

- To remove the URLs and mentions extracted by the **REGEXREPLACE** function, we will make **replace_with** on **SUBSTITUTE** an empty string (**""**).

01. Enter **=SUBSTITUTE(B2, (REGEXEXTRACT(B2,"\B@\w{1,15}|https?:\/\/[^\s/$.?#].[^\s]*")), "")** on cell **C2**.

02. Replicate the formula to the other rows by dragging the fill handle down.

ID	Text	URLs removed
1028954403129184256	Gotta love the facts. https://t.co/bZ2G8AZuo9	Gotta love the facts.
1028954810781814784	You send me crap It's 5 minutes to midnight for a mute https://t.co/FFhYHCitKb	You send me crap It's 5 minutes to midnight for a mute

The TRIM and CLEAN functions

The **TRIM** function is used to remove leading, trailing, and repeated spaces in text while the **CLEAN** function returns the text with the non-printable ASCII characters removed.

```
=TRIM(text)
=CLEAN(text)
```

- **text** – The string or reference to a cell containing a string to be trimmed or the text whose non-printable characters are to be removed.

For example:

- **=TRIM(" Hello, World! ")** → "Hello, World!"
  - Removing leading and trailing spaces.
- **=CLEAN("Hello   World!")** → "HelloWorld"
  - Removing tab character between "Hello" and "World".

ⓘ Spreadsheets **do not show non-printable characters** in the user interface, so using the **CLEAN** function will typically not result in any visible changes.

The TRIM and CLEAN functions

Example use:

- **Remove** leading, trailing, and repeated **spaces** as well as **non-printable characters** from all Tweets.

01. Enter **=CLEAN(TRIM(B2))** on cell **C2**.

02. Replicate the formula to the other rows by dragging the fill handle down.

ID	Text	Cleaned and trimmed text
1028954810781814784	You send me crap It's 5 minutes to midnight for a mute https://t.co/FFhYHCitKb	You send me crapIt's 5 minutes to midnight for a mute https://t.co/FFhYHCitKb
1028954474805710849	RT @Peters_Glen: It is always a good reminder to see how the global average temperature has changed over the last 150 years... https://t.c…	RT @Peters_Glen: It is always a good reminder to see how the global average temperature has changed over the last 150 years...https://t.c…

The SEARCH and FIND functions

The **SEARCH** and **FIND** functions both return the position at which a string is first found within text. **SEARCH**, however, **ignores case** while **FIND** is **case-sensitive**.

```
=SEARCH(search_for, text_to_search, [starting_at])
=FIND(search_for, text_to_search, [starting_at])
```

- **search_for** – The string to look for within **text_to_search**.

- **text_to_search** – The text to search for the first occurrence of **search_for**.

- **starting_at** – [OPTIONAL] The character within **text_to_search** at which to start the search. 1 by default.

For example:

- **=SEARCH("World", "Hello, World!")** → 8
  - 8 is the position of the letter **W** in the word **World**.
- **=FIND("World", "Hello, world!")** → #VALUE!
  - **FIND** is case-sensitive and will therefore not find a match.

The SEARCH and FIND functions

Example use:

Identify all tweets that mention the hashtag **#climatechange**.

- We will use **SEARCH** so that we can identify all relevant hashtags irrespective of sentence case.
- Since **SEARCH** will return an **error** if a Tweet does not contain the hashtag, we will use an **IFERROR** statement to replace the error value with **0**.

01. Enter `=IFERROR(SEARCH("#climatechange",B2),0)` on cell C2.

02. Replicate the formula to the other rows by dragging the fill handle down.

ID	Text	#climatechange
1028954652832882688	RT @6esm: Halfway to boiling: the city at 50C https://t.co/jccTA8tDCS - #climatechange	73
1028954995469811713	#climatechange #spaceweather https://t.co/hB3tQgPeys	1

The SPLIT function

The **SPLIT** function divides text around a specified character or string and puts each fragment into a separate cell in the row.

`=SPLIT(text, delimiter, [split_by_each], [remove_empty_text])`

- **text** – The text to divide.
- **delimiter** – The character or characters to use to split text.

  By default, each character in delimiter is considered individually, e.g. if delimiter is "the", then text is divided around the characters "t", "h", and "e". Set **split_by_each** to **FALSE** to turn off this behaviour.

- **split_by_each** – [OPTIONAL] Whether or not to divide text around each character contained in delimiter. **TRUE** by default.
- **remove_empty_text** – [OPTIONAL] Whether or not to remove empty text messages from the split results. Default behaviour is to treat consecutive delimiters as one (if **TRUE**). If **FALSE**, empty cells' values are added between consecutive delimiters.

The SPLIT function

Example use:

Split all the Twitter texts into individual words.

- We will use the space character as the `delimiter` since words are separated by spaces.
- It is advisable to use the **SPLIT** function after the **last column** since its results populate the cells **horizontally**.

01. Enter **=SPLIT(B2, " ")** on cell C2.

02. Replicate the formula to the other rows by dragging the fill handle down.

ID	Text	Split text				
1028954403129184256	Gotta love the facts. https://t.co/bZ2G8AZuo9	Gotta	love	the	facts.	https://t.co/bZ2G8AZuo9

The **CONCATENATE** function

The **CONCATENATE** function appends strings to one another.

`=CONCATENATE(string1, [string2, ...])`

- `string1` – The initial string.
- `string2` – [OPTIONAL] Additional strings to append in sequence.

For example:

- `=CONCATENATE("Hello", " ", "World!")` → "Hello World!"
  - A white space must be included in the function where needed.

14

The CONCATENATE function

Example use:

Combine each Twitter text with its corresponding ID.

- We will use a colon followed by a whitespace as a **delimiter** between the ID and text.

01. Enter `=CONCATENATE(A2,": ",B2)` on cell C2.

02. Replicate the formula to the other rows by dragging the fill handle down.

ID	Text	Combine text with ID
1028954403129184256	Gotta love the facts. https://t.co/bZ2G8AZuo9	1028954403129184256: Gotta love the facts. https://t.co/bZ2G8AZuo9
1028954995469811713	#climatechange #spaceweather https://t.co/hB3tQgPeys	1028954995469811713: #climatechange #spaceweather https://t.co/hB3tQgPeys

The UPPER, LOWER, and PROPER functions

The **UPPER** function converts a specified string to uppercase, **LOWER** converts a specified string to lowercase, and **PROPER** capitalises each word in a specified string.

```
=UPPER(text)
=LOWER(text)
=PROPER(text_to_capitalise)
```

- **text** – The string to convert to uppercase or lowercase.

- **text_to_capitalise** – The text which will be returned with the first letter of each word in uppercase and all other letters in lowercase.

Some applications include:

- Converting inconsistent capitalisation for uniformity and consistency.

- Ensuring that all data entered into a cell is in a consistent format.

- Manipulating text, e.g. converting text to lowercase and then using other functions to extract specific characters from the string.

- Creating titles (UPPER and PROPER).

The UPPER and LOWER functions

Example use:

Find all Tweets containing the word climate, regardless of case.

- Since the **FIND** function is case-sensitive, we can start by converting the text to uppercase or lowercase before applying the **FIND** function.
- It is common practice to convert text to lowercase during analysis so we will use the **LOWER** function.

01. Enter `=IFERROR(FIND("climate", LOWER(B2)), 0)` on cell C2.

02. Replicate the formula to the other rows by dragging the fill handle down.

ID	Text	Find climate
1028954635443171328	Climate change and wildfires — how do we know if there is a link? https://t.co/2SqyvW7asF via @ConversationUS	1
1028954995469811713	#climatechange #spaceweather https://t.co/hB3tQgPeys	2