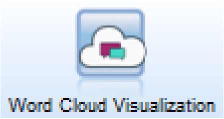
Word Cloud Visualization

Create a word cloud to visualize text data



Product: IBM® SPSS® Modeler

Extension type: Visualization

Table of Contents

Description…………………………………………………………………………………………………3

Requirements…………………………………………………………………………………………………3

Installation…………………………………………………………………………………………………3

R Packages used…………………………………………………………………………………………………3

User Interface…………………………………………………………………………………………………4-5

Examples…………………………………………………………………………………………………6-13

Modeler Data Column Comparison ….……………………………………………………………………6-8

Modeler Data Row Comparison……….……………………………………………………………………8-11

Web Data…………………………………………………………………………………………………12-13

Important links…………………………………………………………………………………………………14

Learn…………………………………………………………………………………………………14

Discuss…………………………………………………………………………………………………14

Description:

This node allows you to create a word cloud from text data coming from one of three sources. The node can create a visualization for data in a modeler data object, it can read a single URL and parse the text using a CSS selector, or it can read a local .txt file. The extension has options for text cleaning, removing punctuation, numbers, and English stop words. It also has options for the displaying the word cloud.

Requirements:

- SPSS Modeler v16.0 or later
- R: http://www.r-project.org/
- 'R Essentials for SPSS Modeler' plugin: https://developer.ibm.com/predictiveanalytics/downloads/predictive-extensions

Installation:

Close SPSS Modeler. Save the .cfe file in the CDB folder of the IBM SPSS Modeler installation directory for Windows and Linux. The copy should reside in that same folder and not in a sub-folder.

For example, for Windows 7 and Modeler 17 the default location is "C:\ProgramData\IBM\SPSS\Modeler\17\CDB". If the ProgramData folder is hidden type the path manually.

Restart SPSS Modeler: the node will now appear in the Output palette.

R Packages used:

- 'wordcloud' package created by Ian Fellows https://cran.r-project.org/web/packages/wordcloud/

- 'tm' package created by Ingo Feinerer [aut, cre], Kurt Hornik [aut], Artifex Software, Inc. [ctb, cph] https://cran.r-project.org/web/packages/tm/

- 'rvest' package created by Hadley Wickham [aut, cre], RStudion [cph] https://cran.r-project.org/web/packages/rvest/

- 'RcolorBrewer' package created by Erich Neuwirth https://cran.r-project.org/web/packages/RColorBrewer/

User Interface

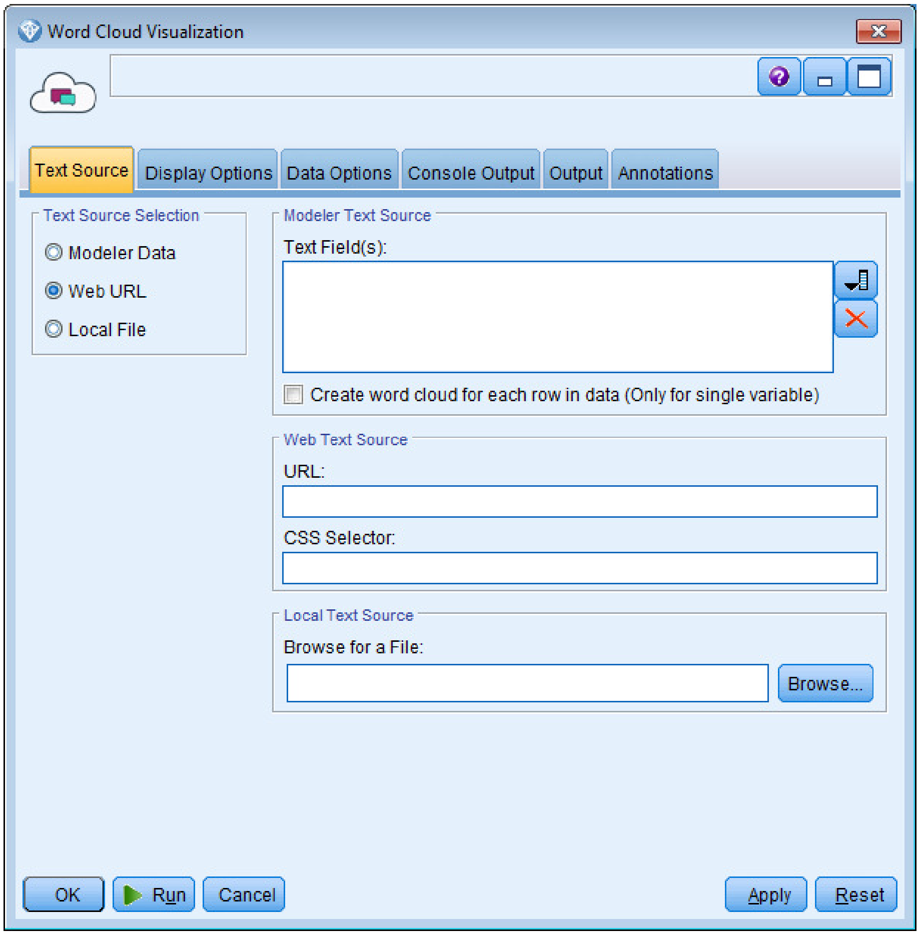
The first tab for this node is Text Source.

First chose the source of the text data, either from Modeler, Web, or Local file. This is selected using the radio button on the left side.

For Modeler Data, select one or more columns from the dataset containing text. If one column is selected, an option is available for creating a word cloud for each row of text in that column. This option is enabled by clicking the check box below the field selector. If that box is unchecked, or multiple columns are selected then the each column is concatenated into a single string and used to create a word cloud per column.

For Web Text, enter the URL containing the text for analysis and the appropriate CSS selector. If you are unfamiliar with CSS Selectors, I recommend a tool like http://selectorgadget.com/ to help find the correct HTML elements.

Local Text source files must be in a .txt format. One word cloud is generated for a text file.



Text Source Selection

○ Modeler Data
● Web URL
○ Local File

Modeler Text Source

Text Field(s):

□ Create word cloud for each row in data (Only for single variable)

Web Text Source

URL:

CSS Selector:

Local Text Source

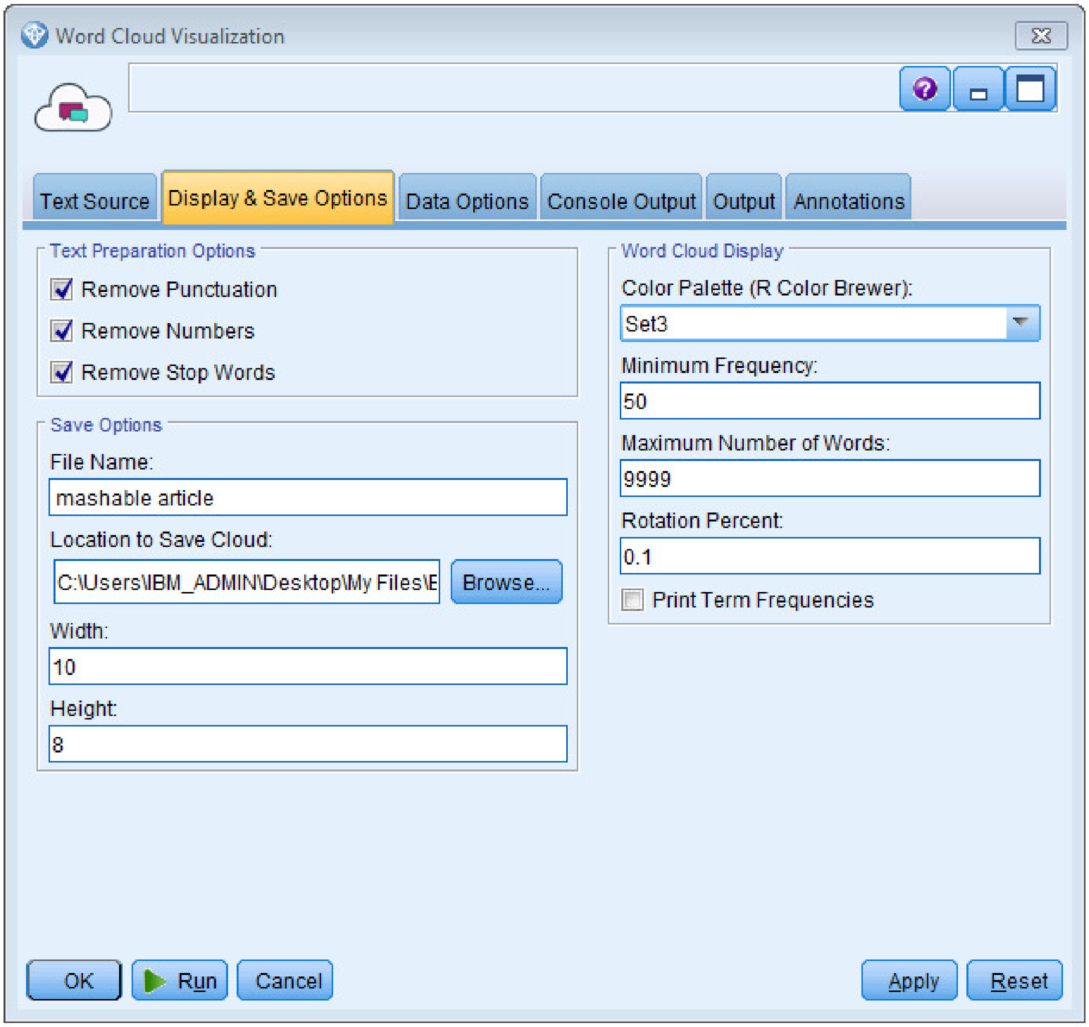
Browse for a File:

The second tab for this node is Display & Save Options.

The tab contains options for text preparation, defaulting to removing punctuation, numbers, and ('english') stop words from the text.

The word cloud display group of parameters adjust the colors used based on the R Color Brewer package. The minimum frequency of words to be used in the word cloud, the maximum number of words to display, and the rotation percent of words can be set in this section. You can also print the words with their respective frequencies by checking the box in this section.

The Save option will create a .png file containing the word cloud(s) generated by the node. If multiple files are created (for multiple word clouds) then a value at the end of the file name will increment for each file. The width and height values in this section are in inches.



Text Preparation Options

☑ Remove Punctuation

☑ Remove Numbers

☑ Remove Stop Words

Save Options

File Name: mashable article

Location to Save Cloud: C:\Users\IBM_ADMIN\Desktop\My Files\E

Width: 10

Height: 8

Word Cloud Display

Color Palette (R Color Brewer): Set3

Minimum Frequency: 50

Maximum Number of Words: 9999

Rotation Percent: 0.1

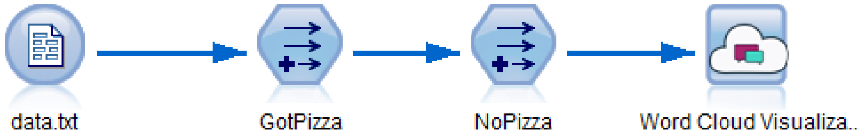
☐ Print Term Frequencies

Examples:

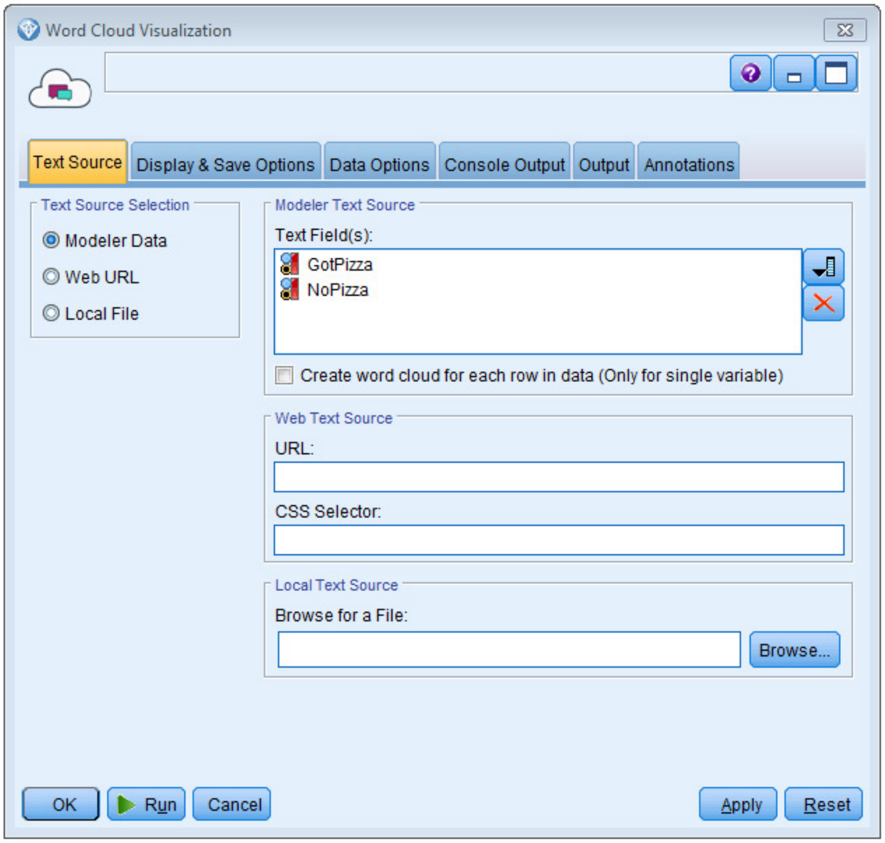
1. Modeler Data Column Comparison

In this example, we will create two columns of text and create a word cloud for each column. The dataset is included in this dataset and is a group of postings from the Random Acts of Pizza sub-Reddit. Each observation in this dataset is the original post made asking for a Pizza donation and a column with True or False corresponding to if the original poster (OP) received a pizza.

As a preprocessing step, I derived two new columns. One column only contains the original post if the outcome was True and one if it was False. The stream included in the folder includes these derive nodes.



Connect the Word Cloud Visualization node and choose the two new fields for the Modeler Text fields. The Modeler Data button should be selected in the “Text Source Selection” group, and the box for creating clouds for each row should be unchecked. We want this unchecked because we want to look at all the words in each column and create two word clouds that are representative of OPs that got pizza and OPs that did not.



Word Cloud Visualization

Text Source Selection

● Modeler Data
○ Web URL
○ Local File

Modeler Text Source

Text Field(s):

GotPizza
NoPizza

□ Create word cloud for each row in data (Only for single variable)

Web Text Source

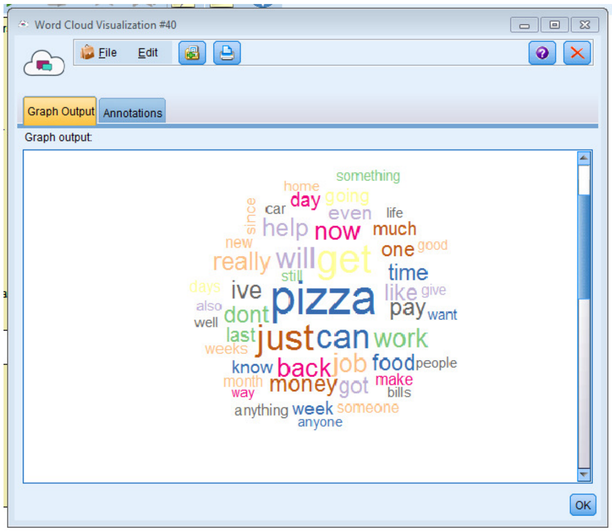
URL:

CSS Selector:

Local Text Source

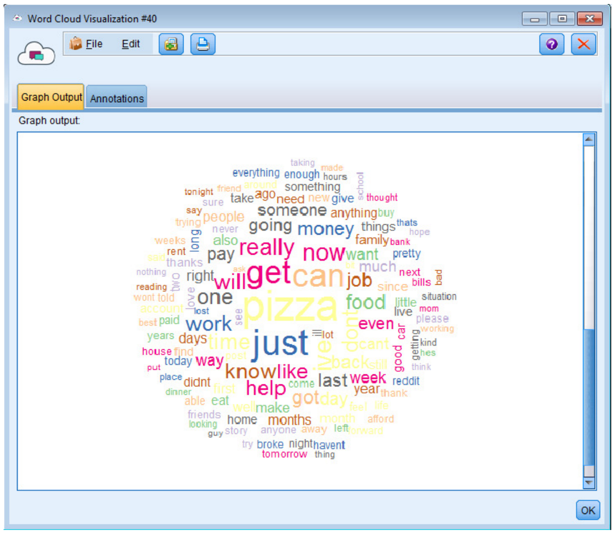
Browse for a File:

You can tweak the display and save options, but you should see a result with something like this:



Word Cloud Visualization #40

OP got Pizza



Word Cloud Visualization #40

File
Edit
Graph Output
Annotations
Graph output:
taking
everything enough made hours
tonight friend around something school
sure take ago need new give thought
say someone anything buy
trying people going money things thats
never hope
weeks long also family bank
rent really now want pretty
said thanks pay much
nothing ask bit next
right will get can job bad
reading two since bills
wont told love one situation
account lost pizza food little
best paid see live mom
work please
even working
years days time just lot dont car
house find cant getting kind
today ive good hes
put way post back still think
place knowlike
didnt last week reddit
dinner help come year
able eat got day thank
wellmake feel life
friends home months month afford
looking story anyone away left forward
guy
try broke night havent
tomorrow thing
OK

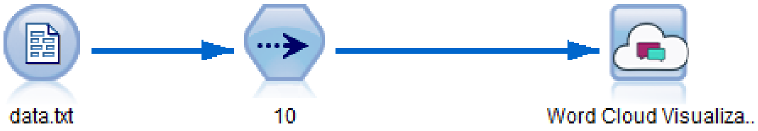
OP didn't get pizza

If you selected to save these word clouds you would see two .png files in the directory specified in the node. You may want to adjust settings like minimum word frequency and color to help fine tune the word clouds.

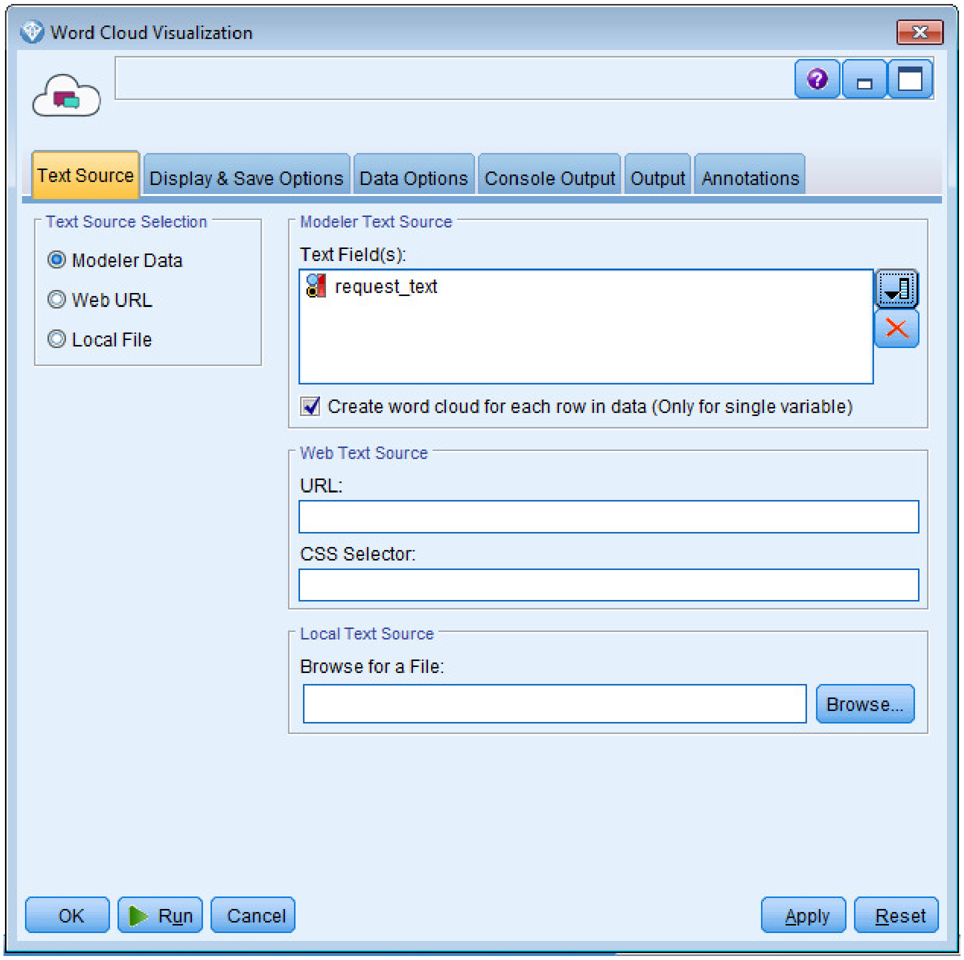
2. Modeler Data Row Comparison

For this example we will use the same dataset, but rather than comparing word clouds between columns, we will generate a separate cloud for each row. As you may imagine, this is only practical when each row has adequate text to generate a word cloud.

For this example we do not need to do any preprocessing. We will just sample the data to create 10 word clouds.



In the Word Cloud Visualization node, under Text Source Selection, Modeler Data should be selected. Select the field request_text in the Text field selector. This time, we also want to check the box to Create word cloud for each row. The interface should look like the screen shot below.



Word Cloud Visualization

Text Source Selection

● Modeler Data

○ Web URL

○ Local File

Modeler Text Source

Text Field(s): request_text

☑ Create word cloud for each row in data (Only for single variable)

Web Text Source

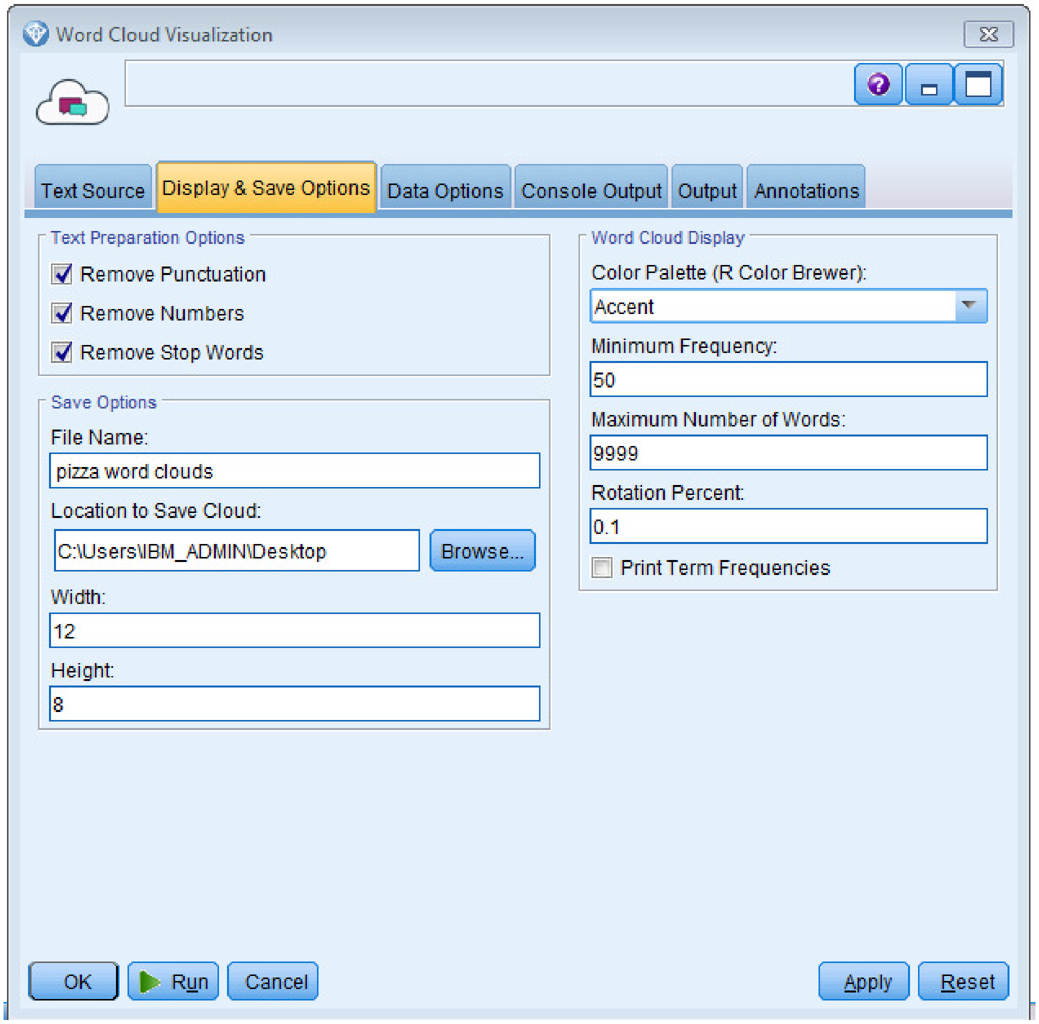
URL:

CSS Selector:

Local Text Source

Browse for a File:

This time let's also take a look at saving the word clouds. I'll be saving all 10 word clouds in my desktop. This can be done by using the following set up on the Display & Save Options tab.



Text Preparation Options

☑ Remove Punctuation

☑ Remove Numbers

☑ Remove Stop Words

Save Options

File Name: pizza word clouds

Location to Save Cloud: C:\Users\IBM_ADMIN\Desktop

Width: 12

Height: 8

Word Cloud Display

Color Palette (R Color Brewer): Accent

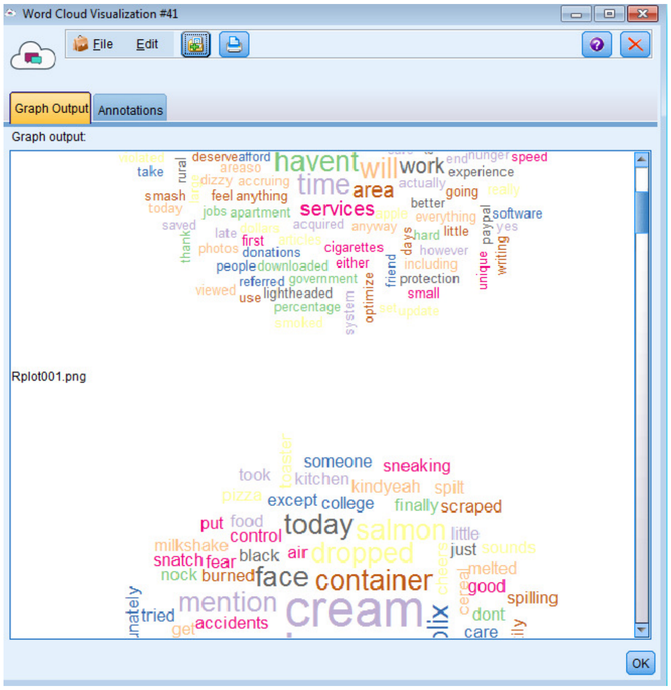
Minimum Frequency: 50

Maximum Number of Words: 9999

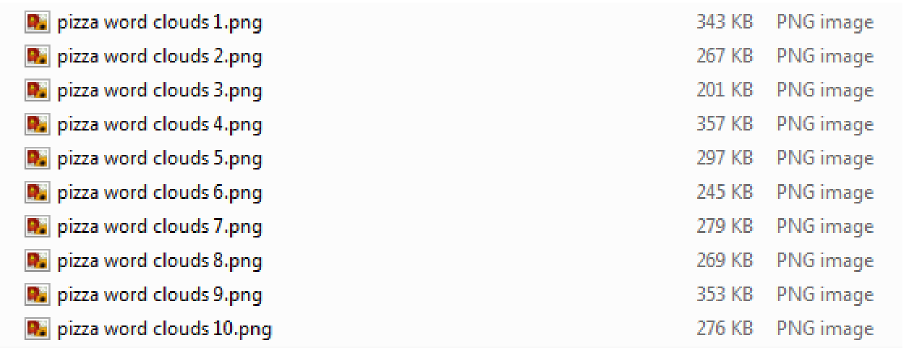
Rotation Percent: 0.1

☐ Print Term Frequencies

Output


Word Cloud Visualization #41
File
Edit
Graph Output
Annotations
Graph output:
wasted take rural deserve afford areaso havent will work end hunger speed experience
smash dizzy accruing time area actually going really
today feel anything better
saved jobs apartment services apple everything software
thank late dollars acquired anyway days paypal yes
photos first articles cigarettes hard little
donations however unique writing
people downloaded either friend including
referred government protection
viewed use lightheaded optimize small
percentage system sel update
smoked
Rplot001.png
toaster someone sneaking
took kitchen kindyeah spilt
pizza except college finally scraped
put food today salmon little
control just sounds
milkshake black air dropped
snatch fear cheers melted
nock burned face container cereal good
mention spilling
unately tried cream dont
get accidents ily care
OK

Output in Modeler



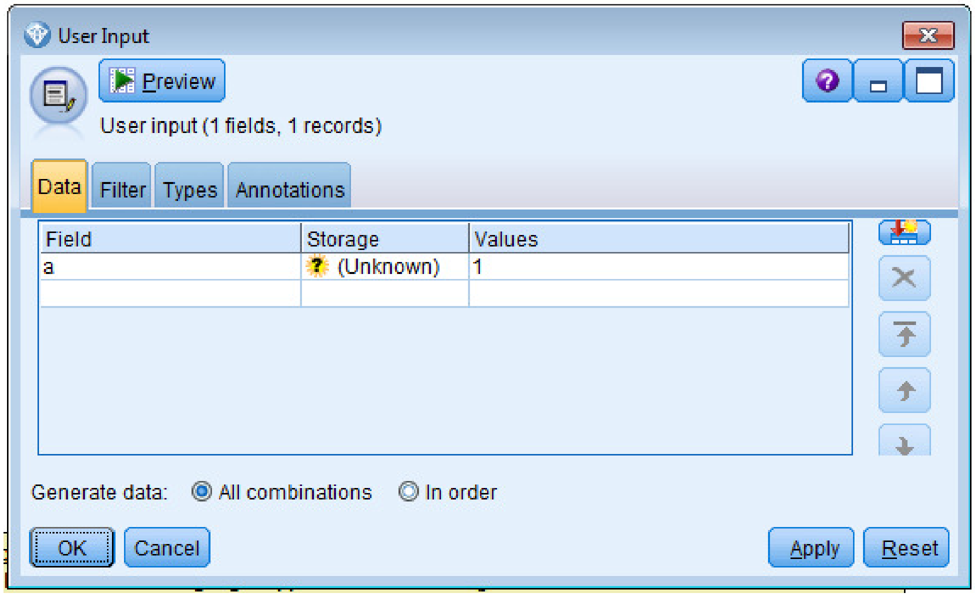
pizza word clouds 1.png	343 KB	PNG image
pizza word clouds 2.png	267 KB	PNG image
pizza word clouds 3.png	201 KB	PNG image
pizza word clouds 4.png	357 KB	PNG image
pizza word clouds 5.png	297 KB	PNG image
pizza word clouds 6.png	245 KB	PNG image
pizza word clouds 7.png	279 KB	PNG image
pizza word clouds 8.png	269 KB	PNG image
pizza word clouds 9.png	353 KB	PNG image
pizza word clouds 10.png	276 KB	PNG image

PNG output files

3. Web Data

As a final example to show another nice feature of this extension, let's make a word cloud for text from a web page.

For this example, we really only need the Word Cloud Visualization node, but R nodes can't be data sources for now in Modeler. To trick the system, add a user input node to the stream with dummy data as shown below:

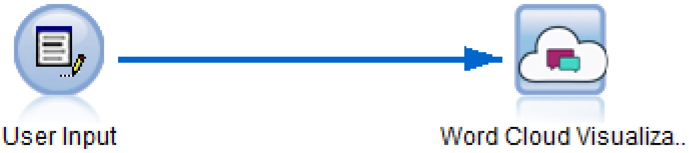


User input (1 fields, 1 records)

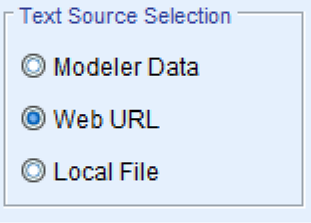
Field	Storage	Values
a	(Unknown)	1

Generate data: ● All combinations ○ In order

Now connect the word cloud node to this to create the stream below:



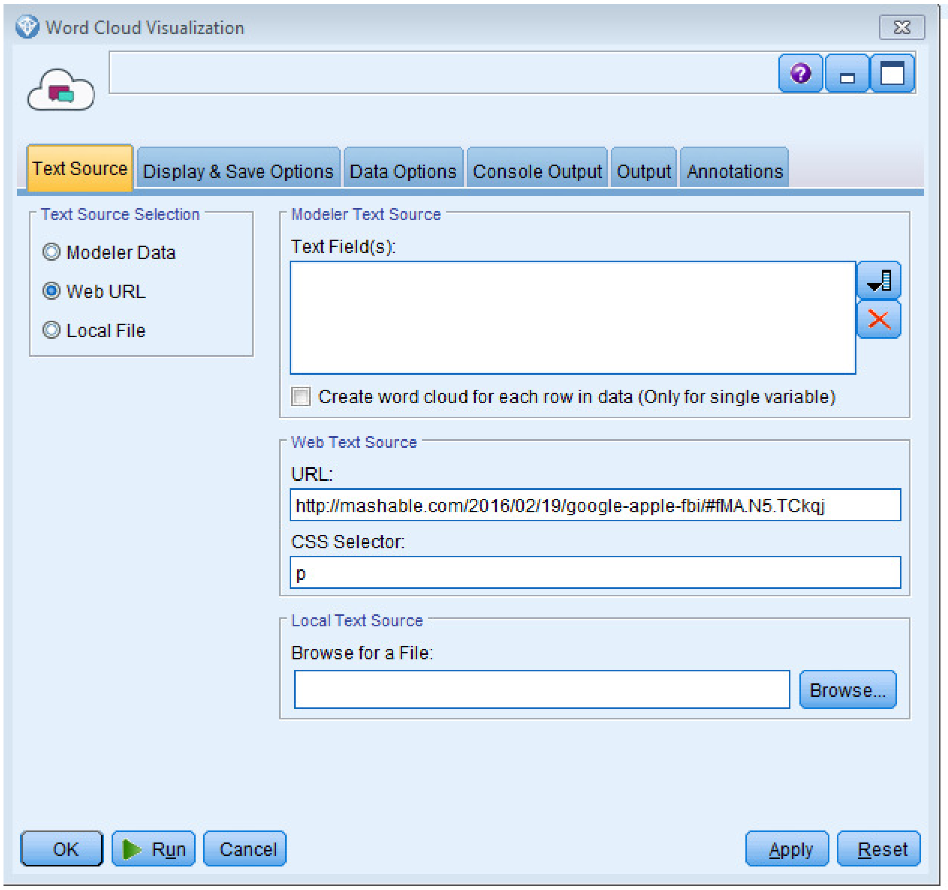
With this stream set up, we need to first change the Text Source Selection to be Web URL.



Text Source Selection

○ Modeler Data
● Web URL
○ Local File

Next we need to add the URL we want to make a word cloud for and enter the CSS selector we want text from. For this example we are using a Mashable article. This is a nice example because all the text is in the CSS selector p. The first tab of the node now should be completed to look like this:



Text Source Selection

○ Modeler Data
● Web URL
○ Local File

Modeler Text Source

Text Field(s):

□ Create word cloud for each row in data (Only for single variable)

Web Text Source

URL: http://mashable.com/2016/02/19/google-apple-fbi/#fMA.N5.TCkqj

CSS Selector: p

Local Text Source

Browse for a File:

Running this will open the web page, select the text based on the CSS Selector and combine it into a word cloud. The output will be a single word cloud that can be saved or manipulated just as in the previous examples.

Important Links

Learn

- Learn more about SPSS software.
- Visit developerWorks Business analytics for more technical analytics resources for developers.
- Create your Bluemix account here: https://console.ng.bluemix.net/solutions/watson
- The Comprehensive R Archive Network is the main site for the R project and each R package. The help pages and manuals that are associated with optimx, nlmrt, and Rcgmin are detailed. Numerous references are provided.
- Read "Do I need to learn R?" (Catherine Dalzell, developerWorks, September 2013) to learn why R is a valuable tool for data analytics that was expressly designed to reflect the way that statisticians think and work.
- "Calling R from SPSS" describes how to use R code inside IBM SPSS Modeler 16.
- Read “Create new nodes for IBM SPSS Modeler 16 using R” to learn how to create new extensions easily.

Discuss

- Visit the IBM SPSS Community to share tips and experiences with other IBM SPSS developers.
- Follow developerWorks on Twitter to be among the first to hear about new resources.