How to use

Self-service Sentiment Analysis

**Creation Date:** 2017-12-13

**Last Updated:** 2017-12-18

**Author: Denis Zuba**

**Revision: 1.0**

Copyright © 2017 Logitech Europe SA,

All Rights Reserved

|  |
| --- |
| **Document Approval** |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Approver** |  | **Signature** |  | **Date** | |  |  |  |  |  | |  |  |  |  |  | |  |  |  |  |  | |
|  |
| **Document Author** |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Author** |  | **Function** |  | **Date** | | Denis Zuba |  | Data Science |  | 18 Dec. 2017 | |  |  |  |  |  | |  |  |  |  |  | |

**Contents**

[1 Introduction 4](#_Toc501384760)

[2 Description of the application 4](#_Toc501384761)

[3 Concept 4](#_Toc501384762)

[4 Access rights (to do only once) 5](#_Toc501384763)

[5 Setup (to do only once) 5](#_Toc501384764)

[6 Steps to use the self-service sentiment analysis engine 6](#_Toc501384765)

[6.1 Prepare the Input file 6](#_Toc501384766)

[6.2 Output file 8](#_Toc501384767)

[6.3 Processed file 9](#_Toc501384768)

# Introduction

This document describes the various steps to use the self-service Sentiment Analysis engine.

# Description of the application

The Sentiment Analysis engine is a simple automated service, which provides the sentiment on sentences of a text provided by the user.

This service is automated and scheduled to run every 2 hours (even hours in PST, so odd hours in CET).

As an example, if you are in Paris or Lausanne and put a file in the Input folder at 4:05 PM your time, it will start being processed at 5PM your time.

The sentiment is returned to the user through a score taking any value between -1 (totally negative sentiment) and 1 (fully positive sentiment). A score of 0 means that the sentiment could be neutral or that it cannot be determined.

The sentiment is extracted through the following steps:

* Text data is split in sentences
* Each sentence is split in words
* Current words (e.g. the, a, and, …) are ignored
* For each sentence, a sentiment is calculated

Pre-requisites / assumptions:

* The following languages are covered:
  + English
  + German
  + French

# Concept

The user has to map the shared drive for sentiment analysis. Then he will be able to add his file to be treated and get the processed files as a result.

Each user is responsible for formatting his file in the required format (csv file containing 2 columns: one with an ID and one with the text) and for naming his files so they cannot be confused with files from other users.

These files are provided by copying them on a shared drive.

The engine outputs are:

- A CSV file with the sentences, and the sentiment (at the sentence level)

- Another CSV file with the sentences, the words, their frequency and the sentiment (at the sentence level)

This second file contains much more rows but could be used for preparing a word cloud for example.

The shared drive will contain the 3 following subfolders:

* input
* output
* processed

The folder structure is used as follow:

* Input: this directory is used for users to upload their input files to be treated.
* Output: this directory is the result place, where the engine will create the output file. This folder will also contain the error files in case one is created.
* Processed: once the file has been processed, it is moved from input to processed directory so it is not processed next time. This way, we can still access the previously processed files.

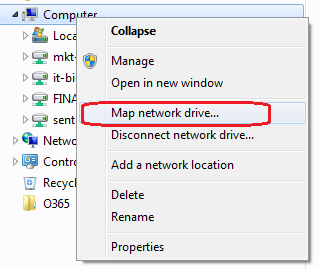
# Access rights (to do only once)

The user has to request write access to the following drive through a Jira ticket: \\ch01f15\sentiment

# Setup (to do only once)

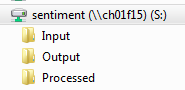
The user has to map the shared drive for sentiment analysis. Then he will be able to add his file to be treated and get the processed files as a result.

To do so, open your file explorer and right-click on “Computer”, then select “Map network drive…”



Then, select the drive you want (in the example, we select the S drive) and type in the folder (you can copy-past it from here): \\ch01f15\sentiment

You will then create the shared drive



# Steps to use the self-service sentiment analysis engine

## Prepare the Input file

The format of this file is csv:

* Delimiter is , (comma)
* Text should be enclosed in “ (double quotes) so a comma inside the text is not understood as a column delimiter

The file contains 2 columns:

* A unique identifier (it can be the row number)
* The text to analyze

Here is, below, an example of the csv file (edited with Notepad): you can see that the name of the columns (1st row) are not important, as the type of ID use is not important as well.

id,commentText

Ugy52g9lcIyEUOqUSQJ4AaABAg,i wish they had this but xxl

Ugz\_s3tbTrN8wmunS714AaABAg,"i just wish logitech created something in between the G900 and G403.. G403's grip is ideal for my hands (where as the g900 causes the joints in my index finger to hurt after a while) but the G900 has a better scroll wheel :c otherwise they're perfect. what would really be amazing is if they released a mouse with the Logitech G9/x's shell with the internals of the G403/603/900/903... i can dream right? XD (tbh, im actually thinking of buying a 603 and trying to take its internals and squish them into my old G9.)"

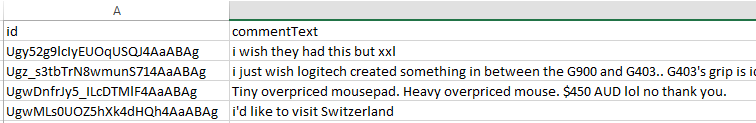
UgwDnfrJy5\_ILcDTMlF4AaABAg,Tiny overpriced mousepad. Heavy overpriced mouse. $450 AUD lol no thank you.

UgwMLs0UOZ5hXk4dHQh4AaABAg,i'd like to visit Switzerland

Only important things:

* 2 columns
* Delimited with a comma (“,”)

The same file opened by Excel shows:



The name of the input file (to be copied to **input** folder by the user) should be as follow:

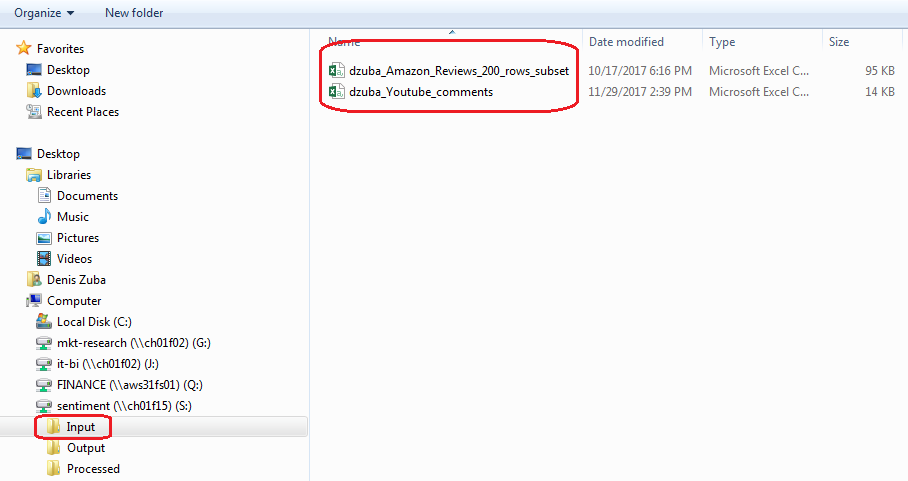
Username\_Filename.csv

Where username is the Logitech username (e.g. dzuba) and Filename can be any character chain that facilitate the identification of the data.

Example (as shown below):

dzuba\_youtube\_comments.csv

Here is an example where two files have been copied in the Input folder in order to be treated by the engine by the user dzuba:



## Output file

When the sentiment analysis succeed, it generates 2 output CSV files (in the output folder), which will provide the result of the sentiment analysis.

For example, where Input is dzuba\_youtube\_comments.csv as provided by the user (C.f. 5.1), the 2 result files will be:

* dzuba\_youtube\_comments\_output.csv
* dzuba\_youtube\_comments\_output\_words.csv

The second file contains the detail per word, so it has much more rows than the first file.

It can be used to prepare a word cloud for example.

The “output” file contains these columns:

* **textsentence\_id**: contains the original text ID, concatenated with the sentence ID (please see below).
* **text\_id**: contains the original id provided in the first column of the input file.
* **sentence\_id**: contains the number of the sentence in the text. This is incremental and starts from 0.
* **word\_count**: contains the number of words in the original sentence selected in the text.
* **sentiment**: contains the sentiment of the sentence selected in the text. The sentiment is a number ranging from -1 (extremely negative) to 1 (extremely positive). A 0 means that the sentence is neutral (or no sentiment could be calculated)
* **sentence**: this field contains the sentence that has been extracted from the text and on which the sentiment is calculated
* **language**: contains the language that has been detected by the process

The “output\_words” file contains these columns:

* **textsentence\_id**: contains the original text ID, concatenated with the sentence ID (please see below).
* **text\_id**: contains the original id provided in the first column of the input file.
* **sentence\_id**: contains the number of the sentence in the text. This is incremental and starts from 0.
* **word\_count**: contains the number of words in the original sentence selected in the text.
* **sentiment**: contains the sentiment of the sentence selected in the text. The sentiment is a number ranging from -1 (extremely negative) to 1 (extremely positive). A 0 means that the sentence is neutral (or no sentiment could be calculated)
* **sentence**: this field contains the sentence that has been extracted from the text and on which the sentiment is calculated
* **language**: contains the language that has been detected by the process
* **word**: contains the word that is currently
* **translated\_word**: contains the translation of the word in English of the word that is currently evaluated. The translation is not activated for the moment and it will simply copy the word from the previous field.
* **freq**: this field will contain the number of occurrence of the word. It always contains 1 and will be used by the end user to use the data (it can be used for the word size in a word cloud for example)

## Processed file

When the original file has been processed, it is placed into the “Processed” folder.

It is then easy to recover the original file and check how it was.