Manual For Data Parsing Program

of Fixed Compensation

# Description:

This is the describing manual for the file “**data\_parsing\_fixed\_with\_comment.ipynb**”, helping you to understand the overall structure and procedure of this program.

However you can still run python file “**data\_parsing\_fixed\_usable.py**” without understanding this, just change

“**input\_position**”, “**output\_position**” and “**date**” at first.

# Procedure:

Reading data from **input\_position**

Drop “index1” in existed file, change column names in new file, combine them.

Create new column **'split\_desc'** copied from “Right\_COMPENSATION\_DESC", this column will later be used at

1.deciding if splitting is needed with regular expression (regex).

2.storing split description.

General Function Preparation:

1. line\_spliter for splitting lines, create new lines into dataframe and split description into “split\_desc”.

2. money\_transfer for transferring money string into Decimal format.

### Data Parsing:

Overall idea is firstly parsing columns which will split lines, then dealing with remaining columns.

Moderately Different procedure for **Writer** and **Rights**.

1. (Not line splitting)

Copy and paste "Right\_COMPENSATION\_TYPE".

2. (Not line splitting)

**Remove** everything after "**Orig deal**". (Have not yet been in use for now, because in “new” there is no “original deal”. In future it can be used in “updated”. )

3. (line splitting)

Splitting lines if there is "**overage**". Before and after "overage", 1 line into 2 lines.

## **Other** than Writer and Rights **Writer**

4. (line splitting) Payment type

For "Compensation.Payment\_Type" and “"Compensation.%\_Total", like “20/60/10/10”, and also split lines by the number of parts.

4. (line splitting) Writer\_Payment\_Type\_spliter

For Writer’s "Compensation.Payment\_Type" and “"Compensation.%\_Total", only “50/50” and “25/25/25/25”, and then get the "Compensation.%\_Amount"

### (line splitting) per\_amount\_spliter

(In the code this function is after “start\_quali\_condi” because this function need to use start\_quali\_condi\_regex , but in the final procedure this runs before “start\_quali\_condi”.)

### For “Compensation.%\_Amount”, used when there are multiple quali condi and each has only 1 corresponding money amount.

(line splitting) writer\_comp\_spliter

(maybe this could be simplified.)

Whether the line in writer should be split is decided by how many times following 3 things have showed in description:

1**. writing step** (e.g. 1st draft)

2.**money amount** (e.g. $1000)

3.**qualifier and condition** (very mostly “On Commencement” and “On Delivery”)

Different situations have different ways to parse:

1.Only 1 writing step, just copy and paste.

2. Only 1 quali\_condi, parse it.

3. more than 1 writing step and no money showed, split by writing steps and the 1st has all the amount.

4. more than 1 pairs of writing step and money matched, no quali\_condi, copy and paste them.

5. more than 1 pairs of quali\_condi and money matched, no writing step, copy and paste them.

6. there are all writing step, money and quali\_condi, and they 3 may be not one-one mathced.

(\*This is not very accurate)

Cut into pieces by writing steps, then insert money and quali\_condi.

(Not line splitting) start\_quali\_condi

Catch if there are

For "Compensation.Start\_Qualifier" and "Compensation.Payment\_Start\_Condition" or "Compensation.Service\_Start\_Condition" (Director),

Like "Over Scheduled Principal Photography".

(Not line splitting)

**dura**\_num\_freq\_paster:

For 'Compensation.Duration\_#' & 'Compensation.Duration\_Freq'.

e.g. “5 weeks”.

(Not line splitting)

**rate**\_num\_freq\_paster

For 'Compensation.Rate' & 'Compensation.Rate\_Freq'.

e.g. “$500 per month”

(Not line splitting)

**start\_date**\_paster

For 'Compensation.Start\_Date'

e.g. from 11/02/2016

(Not line splitting)

**fee\_type**\_paster.

e.g. “Acting fee”. For most files, use some fee type as default.

(Not line splitting)

**service\_type**\_paster, only for Producer, similar to fee type.

Columns Only For **Rights**

(following **Rights** steps are after **1.dura\_num\_freq\_paster** -> **2.Payment\_Type\_spliter**

->**3.start\_quali\_condi\_paster** these 3 steps completed before, and meanwhile Rights has no “overage”.)

(Not line splitting)

total\_guarant\_paster

For “Compensatipon.Total\_Guaranteed\_Commitment”,

If the COMPENSATION\_TYPE is “**guaranteed**”, copy and paste "Right\_COMPENSATION\_AMOUNT".

(Not line splitting)

**check\_number**\_paster

Look for check number.

(Not line splitting)

**start\_expiry\_date**\_paster

For Start\_Date and Expiry\_Date

e.g. 11/12/1993 thru 11/11/1994

(Not line splitting)

**date\_paid**\_paster

For Date Paid

eg."\*\*Paid via wire on 3/30/17.\*\*"

Last step

**Write** parsed data into **excel** file in the “**output\_position**”.

**Reorder** them by corresponding order of columns

**Run** all the procedure for each file,

Loop by “Right\_COMPENSATION\_ID”, for if looping by index, there will be dead loop after insert new lines.