Shatam Structured Scoring (SSS)

(For measuring quality of Points of Interest (POI) and business directory data)

Dec 23 2017- Shatam Technologies, Nagpur

Summary:

We at Shatam Technologies have built expertise in mining data from the unstructured sources like web pages, PDF, Word etc. We extract this data and convert it in the structured form such as a database table or Excel file or a CSV file. As we work with our customers to identify the sources and write complex code to mine data from constantly changing web pages, we realized the basic problem of measuring quality. Looking at other providers of POI (Points of Interest), real estate, and business directory data, we couldn't find a good methodology to measure success. We see claims like 'Best data for businesses in Mexico', '22 Million records for India', 'High quality POI data for USA' etc. None of these claims are backed by any good quantification mechanism. As we work with our clients to collect and enrich Business Directory data for a specific country, we have decided to invent a method called '**Shatam Structured Scoring**' i.e. SSS Mechanism.

SSS allows us to score each record using a formula that considers weight of every field in the record. The average of all the scores in the dataset is the SSS score of that dataset. Our goal is to never let our data score drop between two deliveries.

SSS Method:

For all the fields in the structured dataset, we assign a weight which conveys the importance of the field for the value of each record. The weight is in the range of 5 to 100. 100 weight indicates the highest importance of that field for that record to be valuable. 5 indicates good-to-have. For example, we are using following weights for our fields. We may tweak these based on the value the customer of our data places on these fields.

ID=100	ZIP = 100	ANNUAL_SALES_VOLUME = 5
SIC_CATEGORY =100	PHONE = 50	NUMBER_OF_EMPLOYEES = 5
COMPANY_NAME=100	FAX=50	YEARS_IN_BUSINESS = 5
ADDRESS=100	URL=25	LONGITUDE=50
COLONIA=50	EMAIL = 50	LATITUDE=50
CITY=100	CONTACT_PERSON = 10	
STATE=100	TITLE = 10	

Score of a record = 100* (Sum of weights of all the fields that are not empty)/1060

Score of a dataset = Average of scores of all the records

* 1060 is the max score a record can get by having all the fields available.

Basic quality of data:

SSS method expects every dataset to follow basic rules of data sanity. It doesn't consider quantity of records as a measure of quality.

a) ID is always unique
b) All blanks are converted into nulls
c) All strings are trimmed
d) Data sanity checks are added for every field.
    a. ZIP code can't be more than X characters based on the country
    b. STATE and CITY must exist in that country
    c. EMAIL, FAX, and PHONE must have a valid format.
    d. LATITUDE and LONGITUDE must be within boundaries of that country and must be valid double numbers
    e. ANNUAL_SALES_VOLUME must be a valid number and within acceptable range for that currency and country.
    f. NUMBER_OF_EMPLOYEES and YEARS_IN_BUSINESS must be valid numbers and within acceptable range.
    g. SIC_CATEGORY must be a 4 or 6-digit number.
e) COMPANY_NAME + ADDRESS + CITY + STATE must be unique for that dataset.
f) COMPANY_NAME + PHONE (non-blank) must be unique for that dataset.
g) Every record must have a _SCORE value measured using SSS algorithm.