# Data Extraction from Free Text


## Abstract

Nowadays, the amount of information in the web is tremendous. Big part of it is presented as articles, descriptions, posts and comments i.e. free text in natural language and it is really hard to make use of it while it is in this format. Whereas, in the structured form it could be used for a lot of purposes. So, the main idea that this paper proposes is an approach for extracting data which is given as a free text in natural language into a structured data for example table. The structured information is easy to search and analyze. The structured data is quantitative, while the unstructured data is qualitative. Overall such tool that enables conversion of a text to a structured data will not only provide automatic mechanism for data extraction but will also save a lot of resources for processing and storing of the extracted data.


## Introduction

The amount of data grows exponentially. For the year 2020 the data amount is up to a 64.2 zettabytes [1]. Figure 1 shows the data from 2010 with projection for the next 4 years in zettabytes.

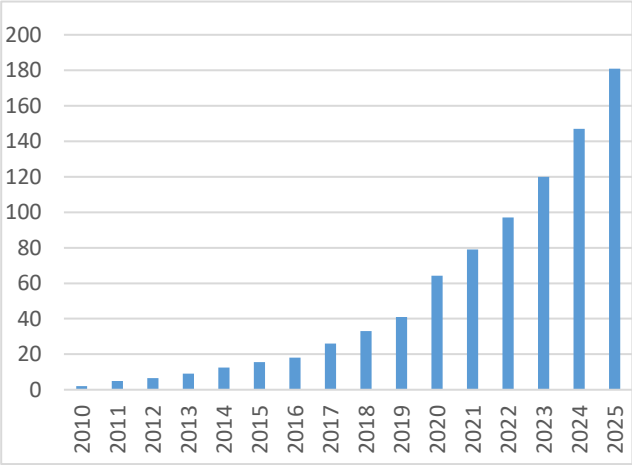


*Figure 1 Data amount projection*

Statistics shows that each person generates 1.7 megabytes of data in just a second. 80 - 90% of the generated data is in unstructured format [3]. Unstructured data could be any texts, emails, social media data, mobile data as text messages and locations, any MS office documents and other [4]. The conversion of the unstructured data will enable the usage of this vast data for analyzing purposes and training of artificial intelligence networks.

The suggested approach is to use and offers the opportunity to customize the key information which will be collected and transformed into records. The customization of the system implies extracting the information, only the one that is applicable for the specified context.

## Technical environment

### Python

The majority of modern solutions regarding natural language processing are created using Python. Python is high-level programming language that supports a variety of well-developed and widely used frameworks for language processing, including spaCy, NLTK and others. The solution presented in this paper is based primarily on Python 3.7, along with other Python libraries described below.

### RASA

RASA offers tools for building natural language processing. It consists of two independent modules. The first one Rasa natural language understanding (Rasa NLU) for language undemanding and Rasa Core for dialogue management. In the paper is observed the Rasa module. It combines a lot of natural language

processing modules and libraries for machine learning. [2]

The RASA platform provides model for intent recognition and key entities extraction. The model is trained with the sentences generated from step 1. They contain key entities with the corresponding indications. The platform allows regular expression search while recognition, which is very convenient for the numeric values in this case.

## Approach

The approach consists of several steps. Firstly, samples are generated. Then they are used for the training of a model for recognition. Once the model is trained it can recognize intents and entities. When entities are recognized they are put into structured records and saved. Those record could be later used as a base for generating training data.

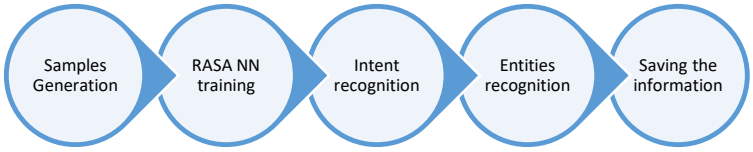


*Figure 2 Scheme of the approach*

**STEP 1 Samples Generation:** The requirements for this step are tables and database with description for each table. The table contains enumerable and numeric values. The enumerable column contains values of a finite set. Whilst the column area would contain a number therefore the column area is numeric column. The count of the items in the set of the distinct values in it is too large, so it cannot be covered as the enumerable. This type of information is contained in the additional database with the descriptions of the table that we mentioned earlier.

*Table 1 Part of the Real Estate training table.*

| EST TYPE | AREA | REGION | PRICE | B YEAR | CONSTR TYPE |
|---|---|---|---|---|---|
| maisonette | 105 | egyptian | 197551 | 2016 | wooden |
| studio | 116 | ukranian | 75964 | 2016 | brick |
| 3 bedroom | 69 | welsh | 59799 | 2007 | wooden |
| maisonette | 84 | scottish | 131159 | 1987 | brick |
| 2 bedroom | 134 | czech | 53766 | 1990 | panel |
| studio | 65 | swedish | 88844 | 1974 | brick |
| studio | 101 | japanese | 192306 | 1993 | brick |
| studio | 88 | angolan | 140748 | 2011 | wooden |
| 2 bedroom | 152 | mexican | 143697 | 1973 | brick |
| office | 160 | russian | 91114 | 1982 | wooden |
| … | … | … | … | … | … |

In addition to the types of the columns of the database also need to contain synonyms for each column. The synonyms are words with which the column could be named or described in natural language. Each column should have at least one synonym in order for the column to be called. Here are some examples for synonyms.

*Table 2 Part of table with synonyms for the real estate context.*

| EST TYPE | AREA | REGION | PRICE | B YEAR | CONSTR TYPE |
|---|---|---|---|---|---|
| type | area | region | price | construction year | construction type |
| property | space | locality | cost | construction | material |
| | field | city | amount | building | |

The encountered requirements are used for the generation of a large set of training data for the neural network which is used for the information extraction. Each sample is generated using a record from the table. Sentence generator implemented in python selects random columns to take part in the produced sentence. For each selected column is randomly selected chosen a word from its list with synonyms from the set of the synonyms belonging to the column. Examples of generated sentences used for the training of the network.

- **Area** *(area synonym)* and **price** *(price synonym)* for the **storehouse** *(estate*

*type key entity)* **property** *(estate type synonym)*, in **russian** *(region key entity)* **locality** *(region synonym)* are 106 *(number value)* and 62222*(number value)*.

- For the **3 bedroom** *(estate type key entity)* **property** *(estate type synonym)* in **swedish** (*region key entity)* **region** *(region synonym)* the **amount** *(price synonym)* is 137578*(number value)*.
- 45845 *(number value)* is the **amount** *(price synonym)* for the **2 bedroom** *(estate type key entity)* **type** *(estate type synonym)* in **belgian** *(region key entity)* **region** *(region synonym)* and **brick** *(construction type key entity)* **construction type** *(construction type synonym).*

The selected synonyms and values from the record are placed into one of the predefined sentence templates. Those templates are natural language sentence with blank placeholders. In the placeholders are put values and synonyms so the sentence describes the information from the record. This is how the natural language sentences for the training are generated.

**STEP 2 RASA NN training**: The RASA platform provides model for intent recognition and key entities extraction. The model is trained with the sentences generated from step 1. They contain key entities with the corresponding indications. The platform allows regular expression search while recognition, which is very convenient for the numeric values in this case.

Once the sentences from each table are generated they are fed to the RASA neural network. The network is also set up with additional settings for numeric values recognition via regular expression.

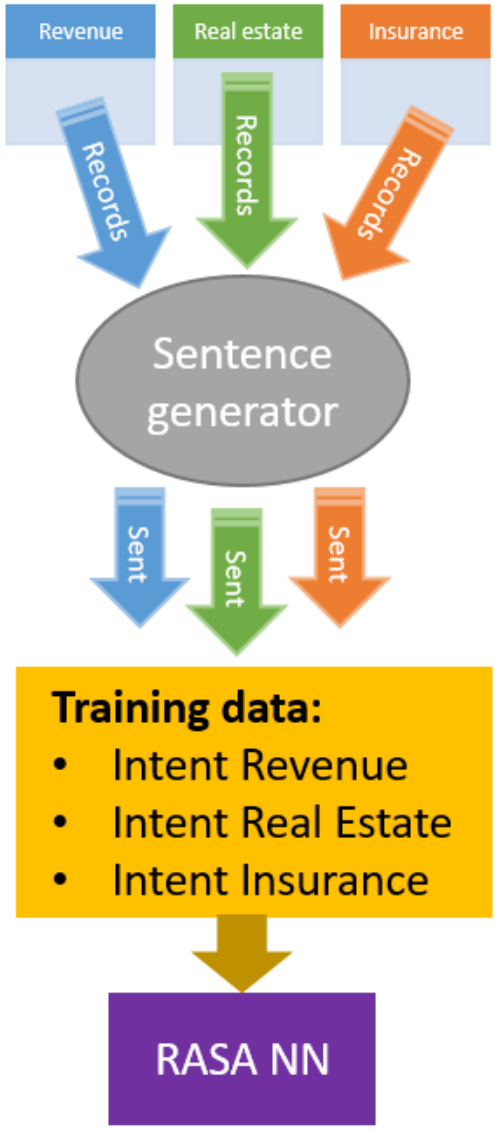


*Figure 3 Visual representation of step 1 and step 2*

**STEP 3 Intent recognition**: After the training the model is able to produce percentage of confidence for the input data to belong to each intent. The intent is the meaning of the sentence. The intent with highest percentage is taken into account. If the information is not related to the any of the topics all of the produces confidence indexes are approximately equal. In that case the information is considered not related to any of the observed topics. Once the intent is recognized the information extracted from the input will be referred to the corresponding table (topic).

**STEP 4 Entities recognition**: As well as the intent the network recognizes key entities. They represent the concrete values of the tables and synonyms for the columns of the referred from the intent OLAP table. Each entity contains name, value and confidence. The name and value are the affiliation of the entity and the fond concrete value. For example:

Input: The property type is 2 bedrooms

Result:

- Intent: real estate
- Entities:
  - Entity 1:
    - Name: estate type synonym
    - Value: property type
  - Entity 2:
    - Name: estate type key entity
    - Value: 2 bedrooms

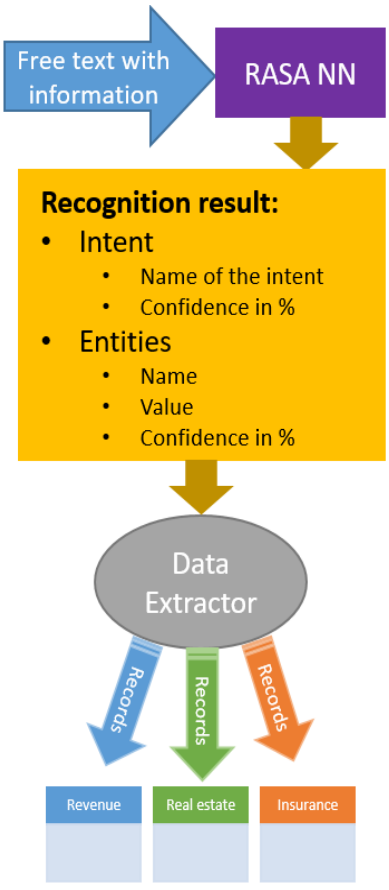


*Figure 4 Visual representation of step 3 and step 5.*

**STEP 5 Information saving**: Lastly the records are sent to Data Extractor also implemented in Python. The Data extractor produces records from the neural network results. The records are the structured view of the unstructured input data (sentences in natural language).

If there is addition information (like the underlined sentence in the input example) which is not related to the context and has no corresponding column in the table, it is not recognized as entity or in other words as value that could be included in the table so they are skipped.

## Results

For the test results are used three themes. Each has its own context that has nothing in common with the others. This shows that the system is independent from the context as soon as the needed information for the training is available. The environment used for the testing of the approach used 200 sentences for each context. They are generated by randomly selecting records three different from training tables each with 10000 rows.

The first test is made with the fallowing paragraph:

*For the revision year of 2019 the revenue in the swiss region is 600. The revenue in the german region for revision year 2018 is 550. The employees are 10.*

Found entities could be seen in table 3

*Table 3 Entities extracted from the first text.*

| Name | Value | Confidence |
|---|---|---|
| **revision_year** | revision | 0.949088742 |
| **measure_val** | 2019 | 0.994773702 |
| **revenue** | revenue | 0.875965164 |
| **region_val** | swiss | 0.889158256 |
| **region** | region | 0.891990163 |
| **measure_val** | 600 | 0.999656696 |
| **revenue** | revenue | 0.875963828 |
| **region_val** | german | 0.956621177 |

| | | |
|---|---|---|
| **region** | region | 0.953662739 |
| **revision_year** | revision | 0.992207866 |
| **measure_val** | 2018 | 0.999308417 |
| **measure_val** | 550 | 0.999674138 |
| **employees_num** | employees | 0.958521313 |
| **measure_val** | 10 | 0.999798283 |

The entities from table 3 are transformed into records that are shown in table 4

*Table 4 Records from the first text.*

| name | region | revenue | Revision year | Employees number |
|---|---|---|---|---|
| | swiss | 600 | 2019 | |
| | german | 2018 | 550 | 10 |

Another test is made with paragraph that refers to the real estate topic.

*The price is 28286.19 EUR. The property type is 2 bedrooms. The construction stage is under construction The number of floors is 1 The floor is first residential floor It is located in the German region. The area is 54 $m^2$. The type of construction is brick. The year of construction is 2022.*

The extracted information is organized into records in table 5

*Table 5 Records from the second text.*

| Estate type | area | region | price | B year | Constr type |
|---|---|---|---|---|---|
| 2 bedroom | 54 | German | 28286.19 | | brick |

## App

A simple web application was developed. It provides opportunities for changing some of the configurations of the system and to test the system for three predefined contexts.

### Data extraction

The home page which is show in figure 6 has text box for the unstructured data in text format. The data could be e either typed in the box or loaded from file. When the data is extracted the results table shows the records (structured data) that was generated from the text. There are three tabs each for a context. The contexts that network understands are companies' revenues, real estates and insurances. Those are used just for the demo version. The system can be set up for recognizing any context. That depends on the initial training data. In the entities table are shown the recognized entities. The name of the entity is the name of group it belongs to. For example, Swiss and German are values for the region column. The value is the concrete word matched as entity and confidence is the probability of the value to belong to the group displayed in the first column (name). The entities with lower confidence that a predefined threshold value are not taken into account when creating the records.

### Network configurations

The editable network configuration settings could be found in figure 7. The training and synonyms tables is used for the generation of the sentences used for the training of the model. Editing the training table will reflect the distinct values that the network recognizes. Changes in the synonyms table will also reflect the network recognition ability. The synonyms refer to the columns of the training table. They are quite important for the recognition of the numeric entities and placing them into the results records. The network will work with the words displayed as synonyms and distinct values.

## Conclusion:

When the extracted information is voluminous enough it can be used as a new table for the generation of the sentences. This operation will improve the accuracy of the training network. But for its execution is needed a large set of collected records

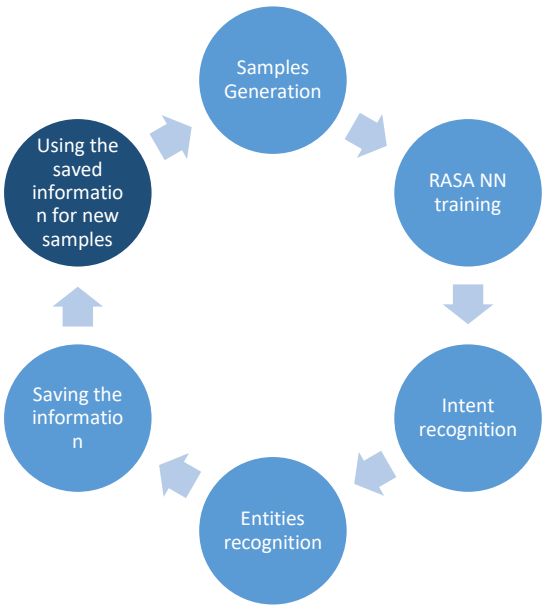


*Figure 5 Complete circularity of the steps in the process.*

## References:

[1] Holst A. (2021, June 30) *"Amount of data created, consumed, and stored 2010-2025",* https://www.statista.com/statistics/871513/worldwide-data-created/

[2] T. Bocklisch, J. Faulkner, N. Pawlowski и A. Nichol, *„Rasa: Open Source Language Understanding and Dialogue Management,"*

[3] Petrov. C. (2021, June 30) *"25+ Impressive Big Data Statistics for 2021",* https://techjury.net/blog/big-data-statistics/#gref

[4] Taylor. C. (2021, June 30) *"Structured vs. Unstructured Data"* https://www.datamation.com/big-data/structured-vs-unstructured-data/

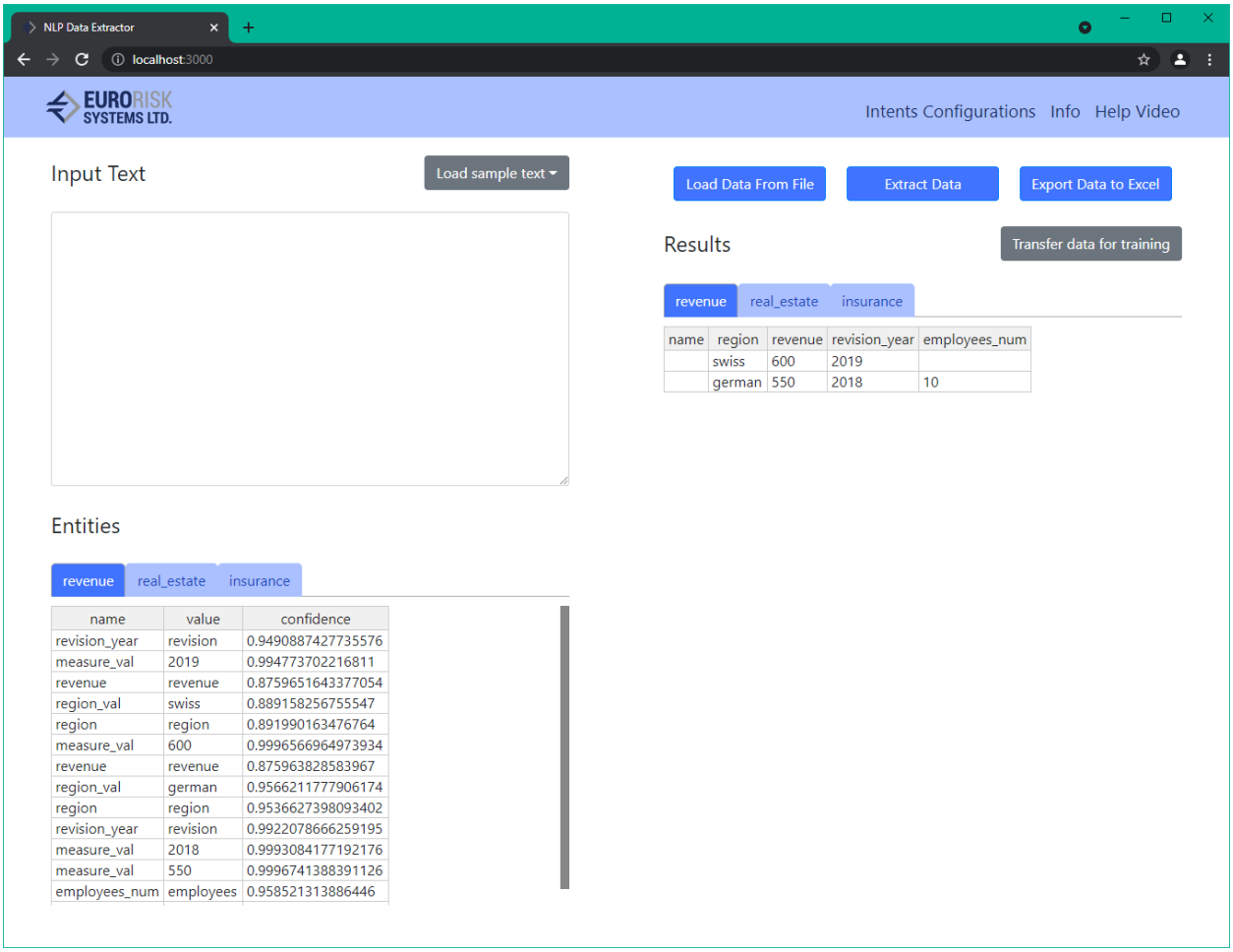


*Figure 6 Home page of the application.*

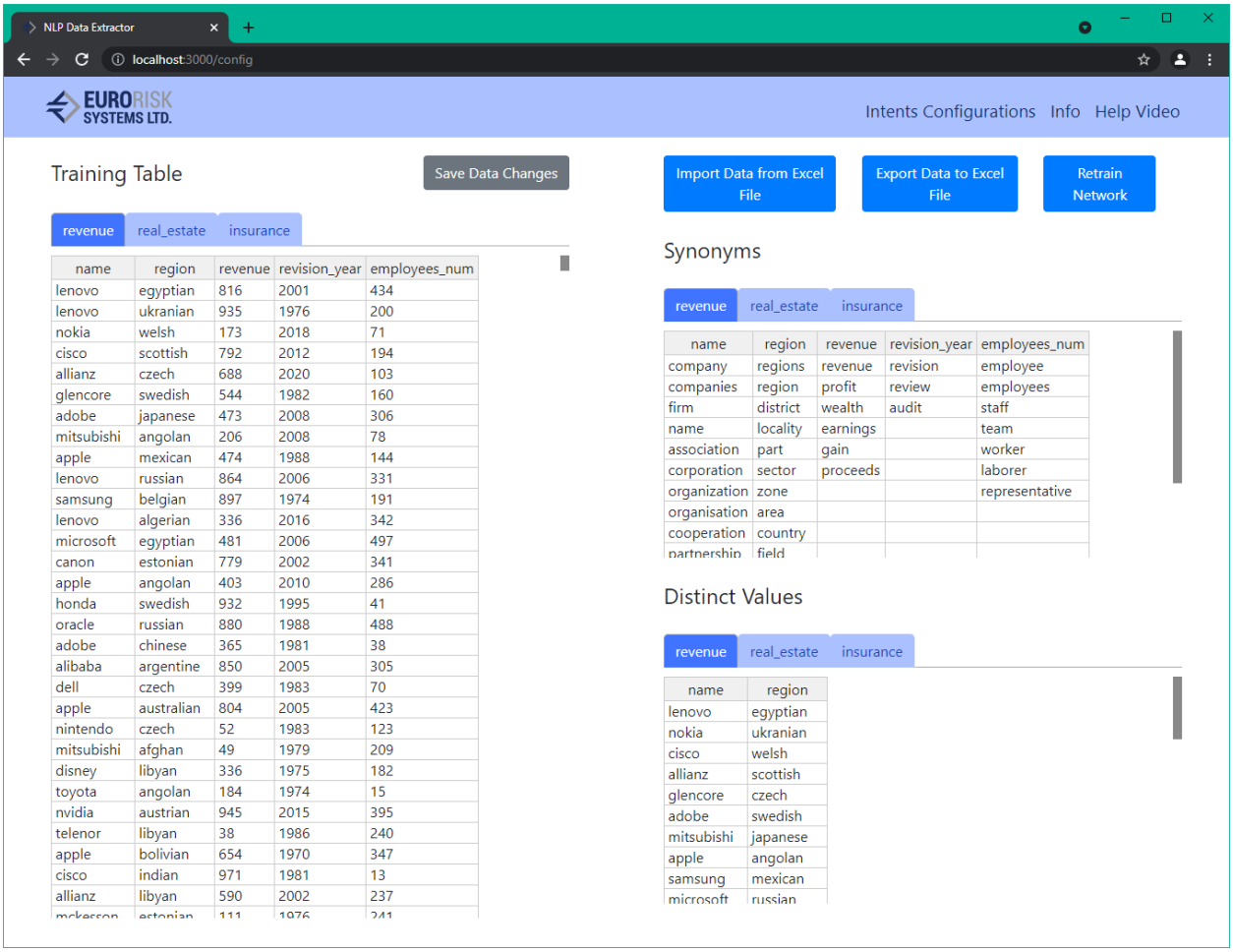


*Figure 7 Configuration setting page of the application.*