Information Extraction

Anushree Tomar

11 October 2019

Natural Language Processing

NLP or text analytics is the process of extracting meaningful information from the text document. This technology is highly applicable in Social media analysis, Product Reviews analysis, Customer support using chatbots, and Content-based Recommendation engine, etc. To accomplish this task first we need to split the document into sentences then tokenize words from each sentence and to extract useful information We will use parts of speech (POS) tags and Entities Recognition.

Problem Statement:

You are given a bank statement (PDF) of a customer, and you will need to write a program to extract the following information.

1. Name of customer
2. Address of customer
3. Bank account number
4. Statement Date
5. List of transactions with following information (Date, Description, Amount, transaction type - Debit or Credit)

Solution

Extract text from pdf

```
## Table: [6 x 4]
##
##   page_id element_id.1 element_id.2 text
## 1 1       1            1            DBS Bank Ltd
## 2 1       2            2            12 Marina Boulevard, Marina Bay Financia
## 3 1       3            3            www.dbs.com.sg | www.posb.com.sg
## 4 1       4            4            S/N: EN05310800573439
## 5 1       5            5            JOHNKARTHIK
## 6 1       6            6            GALI
## . ...     ...          ...          ...
```

Sentence and Word token annotations

Here we will create annotators for sentences and words.

```
##  id type     start end  features
##   1 sentence     1  102 constituents=<<integer,18>>
##   2 sentence   104  120 constituents=<<integer,3>>
##   3 sentence   122  298 constituents=<<integer,13>>
##   4 sentence   300  303 constituents=57
##   5 sentence   305  917 constituents=<<integer,62>>
##   6 sentence  1047 1054 constituents=120
```

POS tags

In this step, we will tag each word in the document as nouns, pronouns, verbs, adverbs etc.so that we can understand the uses of words in the document in a particular sentence.

```
##   text.word. postags
## 1        DBS     NNP
## 2       Bank     NNP
## 3        Ltd     NNP
## 4         12      CD
## 5     Marina     NNP
## 6  Boulevard     NNP
```

Named Entity Recognition

NER is the first step in Information Extraction and is used to identify named entities in the document such as person names, Locations, Organizations, etc.

Entity recognition for an organization

```
## [1] "DBS Bank Ltd 12"
## [2] "Financial Centre Tower"
## [3] "CONSOLIDATED STATEMENT JOHNKARTHIK GALI Joh   RIEGER KUMAR KTGIF SINGAPORE PTE."
## [4] "DBS Co. Reg"
## [5] "POSB Biz Reg No."
## [6] "TRANSIT LINK PTE LTD 30"
## [7] "S & S LINKERS PTE LTD 31"
## [8] "PTE LTD Total"
```

In the Above output we are able to extract list of organizations from the bank statement.Although some irrelevant informations are also extracted.

Entity recognition for location

```
## [1] "Singapore"
```

Extract zip code

```
## [1] SINGAPORE 058571
## Levels:  SINGAPORE 058571
```

Above zip code is related to the customer address.

Extract Bank Account Number

```
##      Accountnum
## 1
## 21 010-025010-2
## 22  12-145753-2
```

Extract Bank Statement Date

```
##    Statement_date
## 1
## 16    31 Aug 2018
```

List of Transaction

	Date	Description	Withdrawl
1		CURRENCY: SINGAPORE DOLLAR	NA
2		Balance Brought Forward	NA
3	28 Aug	Quick Cheque Deposit	NA
4	30 Aug	Point-of-Sale Transaction	20.0
5		TRANSIT LINK PTE LTD	NA
6	30 Aug	Point-of-Sale Transaction	465.0
7		S & S LINKERS PTE LTD	NA
8	31 Aug	Point-of-Sale Transaction	26.5
9		GAYATRI RESTAURANT	NA
10	31 Aug	Point-of-Sale Transaction	16.0