The Title

By

Alice Irankunda (alice.irankunda@aims.ac.rw)
Supervised by Dr.Yabebal FANTAYE

June 2017

AN ESSAY PRESENTED TO AIMS RWANDA IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN MATHEMATICAL SCIENCES

AIMS
African Institute for
Mathematical Sciences
RWANDA

DECLARATION

ACKNOWLEDGEMENTS

DEDICATION

Abstract

Contents

1. Introduction of the Research

In every organization there is a way to communicate. One of the most popular way to transmit the information is to produce a written report which explains how different activities of the organization are going. For large organizations there are huge number of reports and it is so challenging to go through each and every report manually. This research has an aim of providing an easy way of visualizing and extracting the important information locked in reports from NGO and large organisations.

In 1919, the International Federation of Red Cross and Red Crescent societies (IFRC) has been founded, it has some millions of reports related to humanitarian support, How to know automatically the number of people who suffered from a disease? How to know the fraction of fund spent on shelter?

In this research, we tried to use a combination of statistics formulae and techniques of Natural Language Processing (NLP) to find the solution for the extracting entities, Big data and Machine learning for analysing the huge data by using statistical and computing algorithms. Entity can be defined as an instance of existence of something, for example what is the activity done on what place when and how ?

Document modelling by extracting entities is one of the way to deal with natural big data linguistic problems where entity can be considered as a single unit of data like location, people, organization and so on. Entities can be classified based on their relationship.

These are key procedures to be performed for extracting entities:

- The sentences which compose a report must be parsed.
- Entities must be identified in the report and classified.
- Relationship between entities must be modelled.

A report is composed by paragraphs, each paragraph is made by sentences. Natural Language Processing techniques deal with sentences and content based analysis by splitting the sentences into tokens then remove the common words and work with corpus to get entities. The meaning of a word can depend on its surroundings as well as it can be independent. For extracting significant entities, the context of a word is one of the points to be considered carefully.

2. Literature review

In today's life, many organizations are generating unstructured data while they are communicating, there are plenty of entities to be extracted. In this research, n this research, all reports we considered are written in English.

2.1 Corpus Preprocessing

To label the boundaries of sentences is one of the important prerequisite steps in Natural Language Processing. The punctuation marks cause some ambiguity (**?**) for example it is challenging to differentiate the point in abbreviations and a full stop. To handle this ambiguity some systems use the special purpose-regular expression grammar, exception rule method etc.

David Palmer and Marti A. Hearst worked on the problem of punctuation marks. (**?**), they developed an efficient system with high accuracy in automatic labelling the boundaries of the sentence by using the feed forwarding neural-networks where the input was the POS probabilities of all tokens which are surrounding the punctuation and output was found as the label to be assigned to the token.This work was able to correct up to 98.5% for punctuation of sentence-boundaries. A proposed new approach was how to represent the context of punctuation marks without ambiguities.

This research will also look at how neural networks can be used to label different tokens.

Capitalization can be used in different ways such as the beginning of the proper noun, the abbreviation, the post of high level profile people etc. Considering the English language text, if we are given a particular token it is not by chance to determine whether it is a name or not. Some of the approaches to indicate a name are to use capitalization, detection of sentence boundaries and dictionaries (**?**).

2.2 Parse Tree

One of the sentences that compose our sample report says: "Assessment reports indicated 117 deaths, 544 people injured, 12,794 homes damaged and 7,384 houses destroyed", Suppose that this sentence is called "S"

There are two mains steps which can be performed to get the entities from this sentence:

- **Tokenizing**: This is a procedure of taking a sentence and extract the composing atomic linguistic elements e.i. words, verbs, punctuations, adjectives etc . S has the following tokens: ['Assessment', 'reports', 'indicated', '117', 'deaths', ',', '544', 'people', 'injured', ',', '12,794', 'homes', 'damaged', 'and', '7,384', 'houses', 'destroyed']
- **POS**: part-of-speech is a process of attaching to every linguistic element of the sentence

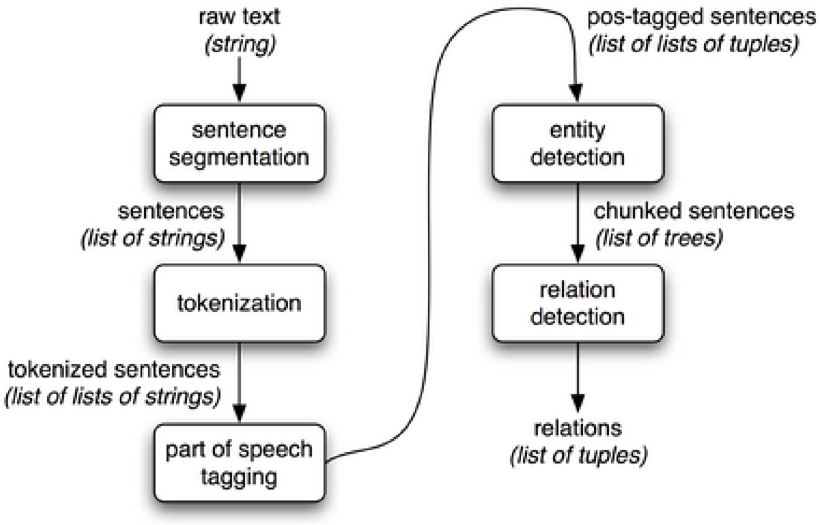
a corresponding tagg based on grammar rules. The POS of S are: [('Assessment', 'JJ'), ('reports', 'NNS'), ('indicated', 'VBD'), ('117', 'CD'), ('deaths', 'NNS'), (',', ','), ('544', 'CD'), ('people', 'NNS'), ('injured', 'VBN'), (',', ','), ('12,794', 'CD'), ('homes', 'NNS'), ('damaged', 'VBN'), ('and', 'CC'), ('7,384', 'CD'), ('houses', 'NNS'), ('destroyed', 'VBD')]

The meanings of the used tags for S:

- JJ: **Adjective**: 'Assessment'
- NNS: **Noun,plural**: 'reports', 'deaths', 'people','houses'
- VBD :**Verbs,past tense**:'indicated', 'injured','damaged','destroyed'
- CD :**Cardinal Number**: '117', '544', '12,794','7,384',
- CC :**Coordinate Conjugation**: 'and'

The parse tree is formed based on the POS, the classification of word and the way words are arranged in a sentence show a kind of relationship between words.

Figure 2.1: Relations extraction from



2.3 Named Entity Recognation and Classification NERC

The term "Named entity" has been coined in 1996 in "sixth Message understanding Conference"(MUC-6 R. Grishman and Sundheim 1996). Entity can be referred as a task, the entity is "named" when it is restricted to one or many rigid designators (**?**), example: persons, location, product are the named entities.

Based on the classification of Standard Generalizes Markup Language (SGML) a task can be divided into three subtasks:

- ENAMEX: location,product,country ,organization
- NUMEX : percentage,quantity
- TIMEX : time, date

The entities from different reports. For extracting entities in a report there are different models which can be used:

2.4 Hidden Markov Model

This model is based on Bayesian probability inference which has been initiated in 18th century. HMM is the earliest applied model for Natural Entities Recognition for English language.The way to perform these tasks is to find the most likely sequence of tagged names(TN) given a sequence of words(SW).

$$P(TN|SW) = \frac{P(SW|TN)P(TN)}{P(SW)} \tag{2.4.1}$$

The equation (2.4.1) is conditional probability, $P(TN|SW)$ can be called posterior and it is the probability of an event Sequence of word occurring given Tagged names has observed. $P(SW|TN)$ is also called likelihood e.i. it is the probability of observing the sequence of words(SW) when the given hypothesis tagged name (TN) is true. On another hand P(TN) doesn't depend on the evidences, P(TN) is called prior e.i. that it is true even if there is no given evidence at all(masters thesis). We can be ignored P(SW) and the remaining objective is to maximise the probability of getting the sequence of tagged names when sequence of words is given.

$$Max\left[P(TN|SW)\right] \tag{2.4.2}$$

From the equation (2.4.2) of the maximization, the following estimation can be made.

$$P(TN) \approx \prod_{i=1}^{n} P(TN_i|TN_{i-1}) \tag{2.4.3}$$

Where TN_i is a tag in the sequence of names (TN), for the likelihood probability can be estimated as :

$$P(SW|TN) \approx \prod_{i=1}^{n} P(SW_i|TN_i) \tag{2.4.4}$$

The above estimations was for a small sequence where TN_i is a tag in the sequence of names (TN) and SW_i is a tag at index i in a sequence words (SW). For the large training corpus, the needed step is estimate based on the number of times the tag occurs and the position of the tag in a given corpus.

$$P(T_i|T_{i-1}) = \frac{K(T_{i-1}, T_i)}{K(T_{i-1})} \tag{2.4.5}$$

Based on the training corpus, $K(T_{i-1}, T_i)$ is referred as a how many times the tag T_i occurs after the tag T_{i-1}. In the corpus, $K(T_{i-1})$ is considered as the number of occurrences for the tag T_{i-1}.

Therefore the estimation can be performed as follow:

$$P(C_i|T_i) = \frac{K(T_i, C_i)}{K(T_i)} \tag{2.4.6}$$

From the equation (2.4.6), the term $K(T_i, C_i)$ is referred as the sum of the times that a word "C_i" has a tag $T_i$in the training corpus. The process of computing the posterior using the above steps is called Markov model.

It is one of the most powerful statistical and machine learning (ML) techniques in modelling and high qualified in entities extraction. When the researcher is willing to train new data, HMM is very robust and efficient in computations. One of the limitations of HMM is that the researcher must have the notion of model topology and statistical techniques on how to deal with large amount of training data.

2.5 Supporting Vector Machine based model(To be edited for non linear data)

This model has an aims of classifying the named entities by using the linear support vector machine which separate input train documents into two categories, a document must be categorized as either positive or negative and be represented in two dimensional graph. Hyperplane is for separating train documents based on their categories and "w" is a weight vector which is perpendicular to hyperplane is represented by the following equation:

$$w.x - b = 0 \tag{2.5.1}$$

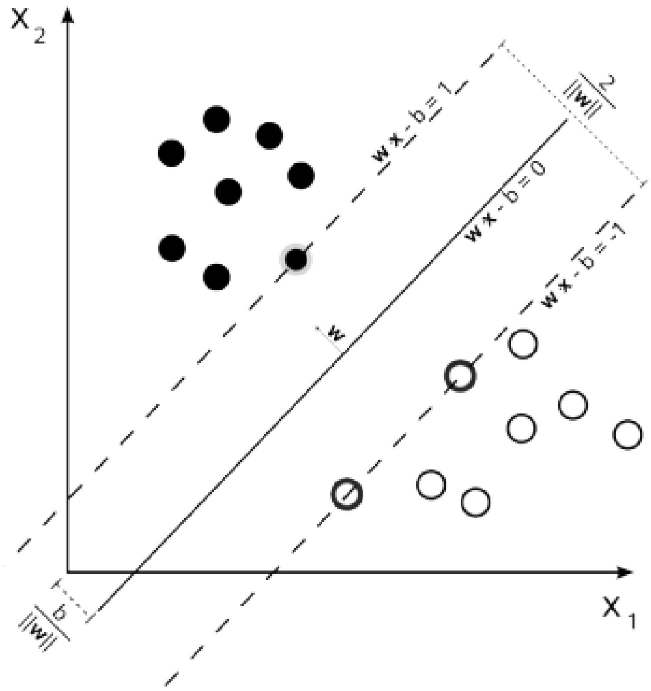
From the (2.5.1), the offset of the hyperplane is $\frac{b}{\|w\|}$

The target is to maximize the the margin between the the points which represent two categories. Remember that the vectors which pass through each of the point representative is perpendicular to the w, suppose that there will be an imaginary line which join two borders points h_{-} and h_{+}. Supporting vectors which are demonstrated by the dashed lines on the figure above are formed by :

$$w.x - b = 1 \qquad \text{and also} \tag{2.5.2}$$

$$w.x - b = -1 \tag{2.5.3}$$

Figure 2.2: SVM in hyperplane representation



There are many algorithms with different approaches to optimization problems but all tends to the same solution says that minimize $njwn$ automatically maximize the margin between h_{-} and h_{+} where the boundary is a half way.

2.6 Big data and Machine learning

The natural language processing is not enough to handle the sophistication and ubiquity of textual data reason why deep learning using machine learning techniques has been introduced. Text analysis using machine learning follow these steps:

- NLP For short
- Lemmatization
- NER
- POS

Now,add another constraint for each document category from the equations (2.6.1) and (2.6.2), in order to hit the target

$$w.x - b \geq 1 \qquad \text{and also} \tag{2.6.1}$$

$$w.x - b \leq -1 \tag{2.6.2}$$

Disadvantages of SVM

The classification of particular documents is not easy to be performed by SVM without destroying the constructed weights but with hand-written rule model. Machine learning uses decision tree procedure than SVM. In addition the decision tree has a detailed boolean-like model which is more popular to user.

2.7 Some Terminologies

Hand-written rule

It is one of the standard approaches of NER and IE, it has been used for extracting the patterns from automated pages such as amazon, NLP is so useful for unstructured humman-written text by delivering part-of-speech (POS), syntactic parsing and categories of semantic words.

Rule /pattern based extraction

Many IE systems uses rule/pattern to extract words and also phrases by looking to the context of those words or based on the their surroundings.(**?**). Some system decided if the procedure of extracting the words should rely on the meaning of each word independently or on the context of their surroundings in a phrase. The limitation of this method is that some words do not have a closer mining to their surroundings that is why Patwardhan Siddharth with help of Ellen Rilo in workshop called "ACL 2006" presented another approach which was generating an automated IE system to learn patterns from a large fixed data set within a specific domain (**?**)

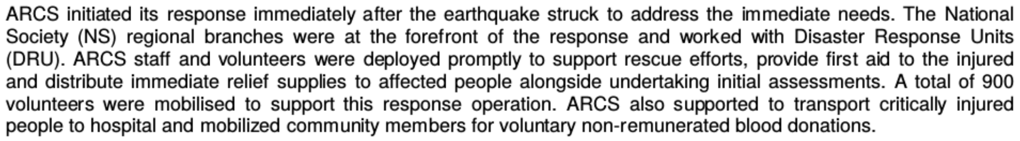
Our research deals with reports generated through a template, compared to the work of (**?**) templates usages is a limitation.

2.1.3. Text classification and Naive Bayes

It is one of the most important algorithm in text classification by using base rule and bag of words to classify the entities (**?**).The user instead of going through the report and start posing many queries, text classification algorithm transient the need information. Its aims is to build a function θ which takes the bag of words and returns the class of sentiment C either positive or negative.

$$\theta$$

$$\Updownarrow$$



ARCS initiated its response immediately after the earthquake struck to address the immediate needs. The National Society (NS) regional branches were at the forefront of the response and worked with Disaster Response Units (DRU). ARCS staff and volunteers were deployed promptly to support rescue efforts, provide first aid to the injured and distribute immediate relief supplies to affected people alongside undertaking initial assessments. A total of 900 volunteers were mobilised to support this response operation. ARCS also supported to transport critically injured people to hospital and mobilized community members for voluntary non-remunerated blood donations.

$$\Updownarrow$$

$$\mathsf{C}$$

The procedure is to look for all words and retrieve those which form the subsets. bag of words are formed after throwing away all words except the subsets.The use of the function θ is for attributing to each item of the bag of words a sentiment.

2.8 Machine learning for Named Entities

The natural language processing is not enough to handle the sophistication and ubiquity of textual data. Deep learning using machine learning techniques has been introduced to solve this problems. The advantages of machine learning for Named Entities:

- Manual extraction of entities is too expensive.
- Fast processes of extraction.
- extraction done by learning algorithms and Natural language tools.
- No limitation of languages.

Frequently ML perform extraction of entities into two phases:

1. Entity boundaries recognition
2. Entities classification

3. Research methodology

3.1 Data and tools

We downloaded the reports about appeals from the IFRC website. We used On the website of IFRC, R-scripts form our co-supervisor professor Xavier.

We downloaded 1262 reports which have been submitted between 01^{st} January 2015 and 31^{st} December 2016. To differentiate the reports, each report has a report Id but different reports can refer to the same appeal Id. As an international organization which insights on the largest humanitarian activities in the world, IFRC reports we have talk about disasters and cash transfer program. Cash Transfer Program (CTP) describes the money used by IFRC to buy food, shelter, etc.

After downloading the data, we transformed the format from pdf to txt in order to manipulate the reports easily, we were able to transform 1260 PDF reports which became our dataset. For analysing the data, we used python programming language.

3.2 Supervised vs Unsupervised Machine Learning

- **Supervised** is a machine learning part which deals with "labelled data", data are categorized and classified. We have a csv document which summarize the reports. We have a csv document which summarize the appeals what we have. The shape of this document is 25 columns and 3997 rows. The "CTP" feature indicate if the appeal is classified as a Cash transfer document or not. Among 3997 appeals, 404 are CTP. Due to limited time, We did not extract

- **Unsupervised** can be defined as a way machine learning processes "unlabelled data". the data are unstructured, uncategorised and unclassified. The reports we have are good example of unlabelled data.

  - **clustering** is a technique for analysing data by making them together into called "clusters", and association rules. Clustering is an unsupervised data-analysis technique used to identify hidden patterns in data. Clustering is also part of exploratory analysis, used to understand data and its properties and to identify any outliers that exist. But primarily it is used for identifying hidden groups in a data set.

3.3 Corpus from Natural Languague Tolkit

Corpus is a set of large data which are semi-structured, to extract entities is simple than to deal with unstructured data. To get the corpus we filtered the data by using unicode of utf-8.

To get compatible data, we have to filter using the Unicode provides canonical and compatible equivalence.

- **Regular Expressions**: in Python, regular expression has operations and modules like "re.py" as so on. they are used to manipulate characters in strings. regular expressions use a backslash ("\") to indicate a special form without invoking the meaning of the special form. There are many regular expressions functions but some of what we used the most are :
    - re.split(): this function split par pattern and return the list of string.
    - re.search(): it returns match objects.
    - the match object ".end()": in a search string, it returns the end position of the match.
    - the match object: in a search string, it returns the start position of the match.
- **Python string Strip() method**: strip method helped us to remove unwanted characters from the beginning and end of the string. to indicate the position of the character to be stripped we use left(l.strip()) which removes the character at the beginning of a string or right(r.strip()) to remove the character at the end of the string.

After getting semi-structured documents, we removed the Stop Words which are defined as unnecessary words for the meaning of the reports. we ignored them because they return vast amount of unwanted information. Some example of English Stop Words: a, almost, details, during, upon and so on.

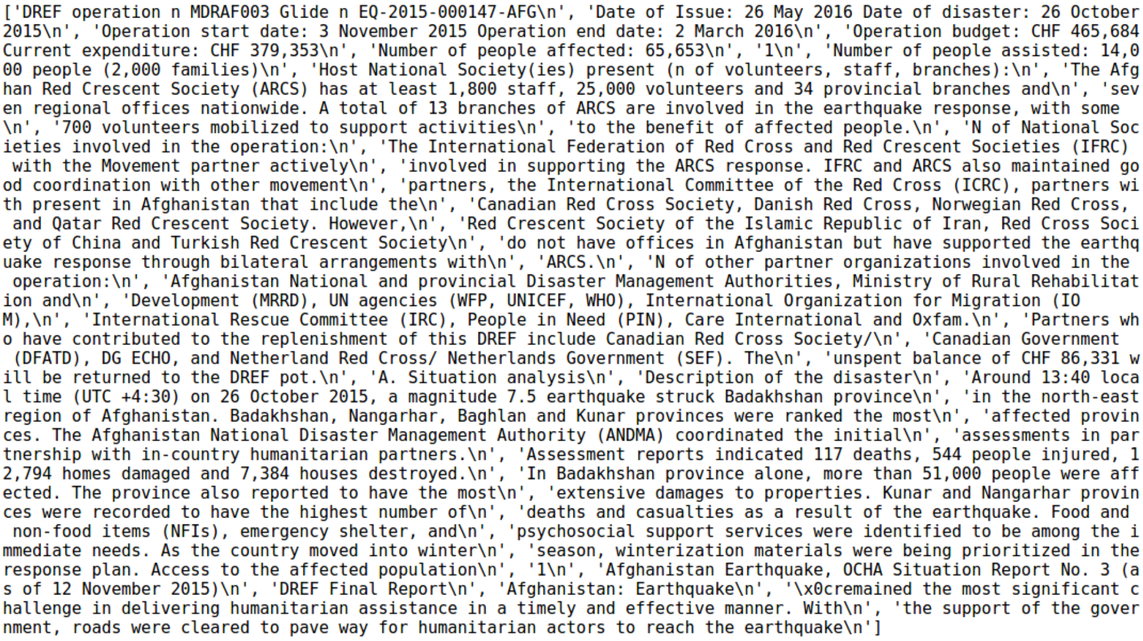
Now we can check how for all of 1260 documents and count the Stop Words to be removed from vocabularies of corpus. We trained corpus by nltk package called FreqDist which uses frequency distribution of each word occurs in corpus, then the module of nltk technique called "nltk.corpus.PlaintextCorpusReader" helped us to get total 58104 stop words over the whole 7796263 vocabularies.

3.4 Extraction of Entities

To extract entities we used default dictionary built in collection package of nltk. Our dataset now is a folder containing 1260 corpus files, we used nltk chruncker to get sets of sentences of corpus. let have a look for our sample document the way sentences are split.

The figure above show the 45 first lines of the sample document. each each line is ended by $'/n'$ Now we can use the Stanford named entities recognizer for tagging and extracting the named entities.

Figure 3.1: Set of sentences



```
['DREF operation n MDRAF003 Glide n EQ-2015-000147-AFG\n', 'Date of Issue: 26 May 2016 Date of disaster: 26 October
2015\n', 'Operation start date: 3 November 2015 Operation end date: 2 March 2016\n', 'Operation budget: CHF 465,684
Current expenditure: CHF 379,353\n', 'Number of people affected: 65,653\n', '1\n', 'Number of people assisted: 14,0
00 people (2,000 families)\n', 'Host National Society(ies) present (n of volunteers, staff, branches):\n', 'The Afg
han Red Crescent Society (ARCS) has at least 1,800 staff, 25,000 volunteers and 34 provincial branches and\n', 'sev
en regional offices nationwide. A total of 13 branches of ARCS are involved in the earthquake response, with some
\n', '700 volunteers mobilized to support activities\n', 'to the benefit of affected people.\n', 'N of National Soc
ieties involved in the operation:\n', 'The International Federation of Red Cross and Red Crescent Societies (IFRC)
 with the Movement partner actively\n', 'involved in supporting the ARCS response. IFRC and ARCS also maintained go
od coordination with other movement\n', 'partners, the International Committee of the Red Cross (ICRC), partners wi
th present in Afghanistan that include the\n', 'Canadian Red Cross Society, Danish Red Cross, Norwegian Red Cross,
 and Qatar Red Crescent Society. However,\n', 'Red Crescent Society of the Islamic Republic of Iran, Red Cross Soci
ety of China and Turkish Red Crescent Society\n', 'do not have offices in Afghanistan but have supported the earthq
uake response through bilateral arrangements with\n', 'ARCS.\n', 'N of other partner organizations involved in the
 operation:\n', 'Afghanistan National and provincial Disaster Management Authorities, Ministry of Rural Rehabilitat
ion and\n', 'Development (MRRD), UN agencies (WFP, UNICEF, WHO), International Organization for Migration (IO
M),\n', 'International Rescue Committee (IRC), People in Need (PIN), Care International and Oxfam.\n', 'Partners wh
o have contributed to the replenishment of this DREF include Canadian Red Cross Society/\n', 'Canadian Government
 (DFATD), DG ECHO, and Netherland Red Cross/ Netherlands Government (SEF). The\n', 'unspent balance of CHF 86,331 w
ill be returned to the DREF pot.\n', 'A. Situation analysis\n', 'Description of the disaster\n', 'Around 13:40 loca
l time (UTC +4:30) on 26 October 2015, a magnitude 7.5 earthquake struck Badakhshan province\n', 'in the north-east
region of Afghanistan. Badakhshan, Nangarhar, Baghlan and Kunar provinces were ranked the most\n', 'affected provin
ces. The Afghanistan National Disaster Management Authority (ANDMA) coordinated the initial\n', 'assessments in par
tnership with in-country humanitarian partners.\n', 'Assessment reports indicated 117 deaths, 544 people injured, 1
2,794 homes damaged and 7,384 houses destroyed.\n', 'In Badakhshan province alone, more than 51,000 people were aff
ected. The province also reported to have the most\n', 'extensive damages to properties. Kunar and Nangarhar provin
ces were recorded to have the highest number of\n', 'deaths and casualties as a result of the earthquake. Food and
 non-food items (NFIs), emergency shelter, and\n', 'psychosocial support services were identified to be among the i
mmediate needs. As the country moved into winter\n', 'season, winterization materials were being prioritized in the
response plan. Access to the affected population\n', '1\n', 'Afghanistan Earthquake, OCHA Situation Report No. 3 (a
s of 12 November 2015)\n', 'DREF Final Report\n', 'Afghanistan: Earthquake\n', '\x0cremained the most significant c
hallenge in delivering humanitarian assistance in a timely and effective manner. With\n', 'the support of the gover
nment, roads were cleared to pave way for humanitarian actors to reach the earthquake\n']
```

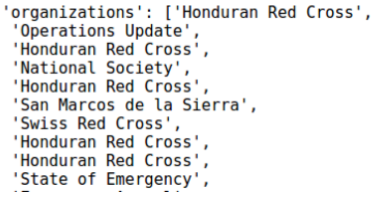
3.2.1 Entities Recognition and Classification

Stanford named entities recognition is a able to correctly identify the named recogniser which labels sequences of words in a text. The next step is to split the sentences into set of words called tokens. By using the Stanford NER tokenizer where token can be tagged.

- **Stanford NER Tagger**: The process of classifying tokens with the taggs. A tagg can be defined as one of classes significant words like nouns, adjectives etc. we used the package Stanford POS Tagger to classify the words.
- **Stanford NER Models**: Stanford has different models such as "stanford-corenlp-full-2016-10-31", "stanford-ner-2014-01-04" which is the version we used.

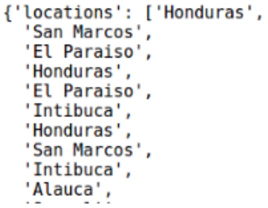
We specified the named entities that we wanted to extract. we classified them into the four categories locations, organizations, persons who participated in reported activities and others. The last category called "other" combined all numerical entities such as time, amount of money, number of people, percentage, etc.

(a) Sample image of extracted locations



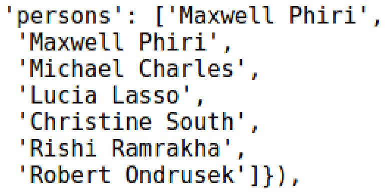
```
'organizations': ['Honduran Red Cross',
 'Operations Update',
 'Honduran Red Cross',
 'National Society',
 'Honduran Red Cross',
 'San Marcos de la Sierra',
 'Swiss Red Cross',
 'Honduran Red Cross',
 'Honduran Red Cross',
 'State of Emergency',
```

(b) Sample image of extracted locations



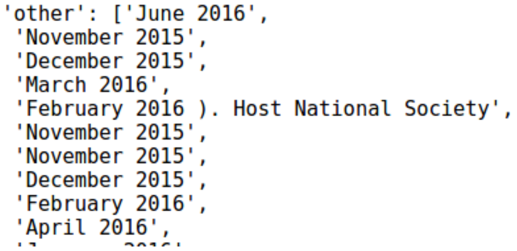
```
{'locations': ['Honduras',
  'San Marcos',
  'El Paraiso',
  'Honduras',
  'El Paraiso',
  'Intibuca',
  'Honduras',
  'San Marcos',
  'Intibuca',
  'Alauca',
```

(c) Sample image of extracted locations



```
'persons': ['Maxwell Phiri',
 'Maxwell Phiri',
 'Michael Charles',
 'Lucia Lasso',
 'Christine South',
 'Rishi Ramrakha',
 'Robert Ondrusek']}),
```

(d) Sample image of extracted locations



```
'other': ['June 2016',
 'November 2015',
 'December 2015',
 'March 2016',
 'February 2016 ). Host National Society',
 'November 2015',
 'November 2015',
 'December 2015',
 'February 2016',
 'April 2016',
```

References

Alan Adolphson, Steven Sperber, and Marvin Tretkoff, editors. *p-adic Methods in Number Theory and Algebraic Geometry*. Number 133 in Contemporary Mathematics. American Mathematical Society, Providence, RI, 1992.

Alan Beardon. From problem solving to research, 2006. Unpublished manuscript.

Matthew Davey. *Error-correction using Low-Density Parity-Check Codes*. Phd, University of Cambridge, 1999.

Leslie Lamport. *LaTeX: A Document Preparation System*. Addison-Wesley, 1986.

D. J. C. MacKay and R. M. Neal. Good codes based on very sparse matrices. Available from www.inference.phy.cam.ac.uk, 1995.

David MacKay. Statistical testing of high precision digitisers. Technical Report 3971, Royal Signals and Radar Establishment, Malvern, Worcester. WR14 3PS, 1986a.

David MacKay. A free energy minimization framework for inference problems in modulo 2 arithmetic. In B. Preneel, editor, *Fast Software Encryption (Proceedings of 1994 K.U. Leuven Workshop on Cryptographic Algorithms)*, number 1008 in Lecture Notes in Computer Science Series, pages 179–195. Springer, 1995b.

Claude Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423, 623–656, 1948.

Claude Shannon. The best detection of pulses. In N. J. A. Sloane and A. D. Wyner, editors, *Collected Papers of Claude Shannon*, pages 148–150. IEEE Press, New York, 1993.

Web12. Commercial mobile robot simulation software. Webots, www.cyberbotics.com, Accessed April 2013.

Wik12. Black scholes. Wikipedia, the Free Encyclopedia, http://en.wikipedia.org/wiki/Black%E2%80%93Scholes, Accessed April 2012.