

Natural Language Processing: An Analysis of Open Source Named Entity Recognition Tools

R. Janani¹, Dr.S. Vijayarani²
Ph.D. Research Scholar¹, Assistant Professor²
Department of Computer Science
Bharathiar University, Coimbatore – 46.

Abstract: To extract the important information from natural text, recognition of exact entities such as person, place, organization, concept and locations is very useful. Learning to extract names in natural language text is called Named Entity Recognition (NER) task. It is used to solve the problems in the area of Information Retrieval, Machine Translation, Text Summarization and Web Search. For NER, the number of open source tools are available. The main aim of this research work is to analyze the performance of nine open source NER tools. They are Dandelion API, spaCy, Stanford Named Entity Tagger, ParallelDots, Text Analysis API, displaCy Named Entity Visualizer, TextRazor, Cognitive Computing Group and NLTK. Based on the results, it is observed that the Dandelion API tool gives the better performance when compared to other tools.

Keywords: NLP, Named Entity, Entity Recognition, Open Source Tools, Information Extraction, Information Retrieval

I. INTRODUCTION

Named entity is a phrase or word which exactly recognizes the single entity from the collection of documents that has the related characteristics. Named Entity is the term which was introduced in the sixth Message Understanding Conference (MUC-6). In fact, the MUC conferences were the events that have provided a significant way to the research of this area. It has delivered the benchmark for named entity systems that achieved a variety of information extraction tasks [1].

In information extraction, the Named Entity Recognition (NER) is the sub problem and it is used to identify the entities such as, name, place, concept and organization [2]. It encompasses the processing of structured and unstructured documents. In Natural Language Processing (NLP) system, the NER is the essential task and this process is the core of NLP. Mainly, the NER involves two important tasks, first the extraction of proper entities in the text and second the classification of extracted entities into the set of predefined classes or categories person names, organizations (companies, government organizations, committees, etc.), locations (cities, countries, rivers, etc.), date and time expressions [3].

1.1 NEED FOR NER

Nowadays, the amount of digital information will propagate by the factor of 44 and to manage that information, the investment and the staff will grow by the factor of 1.4. Hence, there is a need for handling and searching the concise information from the structured and unstructured data [4]. Named Entity Recognition, which significances to recognize the semantics in the unstructured texts and it is serving as the basis for several other critical areas to achieve the information such as, text mining, information extraction, semantic annotation, question answering, ontology population and opinion mining [5].

Sample Input

Request-oriented classification may be a classification that is targeted towards a particular audience or user group. For example, a library or a database for feminist studies may classify/index documents differently when compared to a historical library. It is probably better, however, to understand request-oriented classification as policy-based classification: The classification is done according to some ideas and reflects the purpose of the library or database doing the classification, Stanford University. In this way, it is not necessarily a kind of classification or indexing based on user studies, John, 1985, London.

Result:

Request-oriented classification may be a classification that is targeted towards a particular audience or user group. For example, a library or a database for feminist studies may classify/index documents differently when compared to a historical library. It is probably better, however, to understand request-oriented classification as policy-based classification: The classification is done according to some ideas and reflects the purpose of the library or database doing the classification, Stanford University. In this way, it is not necessarily a kind of classification or indexing based on user studies, John, 1985, London.

The entities are extracted from the input natural text as it denotes as Concepts, Organization, Person, Place, Date.

This paper is organized as follows, section 2 gives the details about open source NER tools, its functions and its results. The performance analysis of different open source NER tools are given in Section 3. Section 4 explains the conclusion of this comparative analysis.

II. NER TOOLS

For NER, many open source tools are available. The tools are listed as follows,

- Dandelion API
- spaCy
- Stanford Named Entity Tagger
- ParallelDots
- Text Analysis API
- displaCy Named Entity Visualizer
- TextRazor
- Cognitive Computing Group
- NLTK

In order to perform the analysis, the same input is provided and the input is processed by these tools and it produced the output. The output of each tool is considered for the analysis. Each tool has recognized the different types of entities for the same input.

2.1 Dandelion API

This is a named entity extraction & linking API which performs very well for short and large text. This tool currently works on texts in English, French, German, Italian, Portuguese, Russian, Spanish and many other languages. With this API user will be able to automatically tag their texts, extracting Wikipedia entities and enriching their data [6]. Figure 1 and 2 shows the input and output of Dandelion API.

The screenshot shows the Dandelion API web interface. At the top, there is a text input field containing the paragraph: "Request-oriented classification may be a classification that is targeted towards a particular audience or user group. For example, a library or a database for feminist studies may classify/index documents differently when compared to a historical library. It is probably better, however, to understand request-oriented classification as policy-based classification: The classification is done according to some ideas and reflects the purpose of the library or database doing the classification, Stanford University. In this way, it is not necessarily a kind of classification or indexing based on user studies, John, 1985, London." Below the text field, there is a "Language:" dropdown menu set to "Autodetected". To the right of the dropdown is a "More Tags" slider control, currently positioned towards "More Precision". Below these controls is a red button labeled "Extract Entities".

Fig 1. Dandelion Input

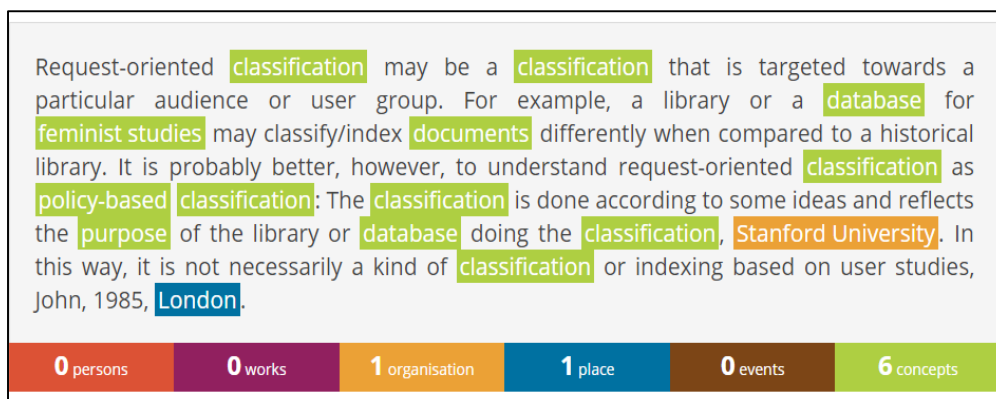


Fig 2. Dandelion Output

2.2 spaCy

spaCy is a library for industrial-strength natural language processing in Python and Cython. It features state-of-the-art speed and accuracy, a concise API, and great documentation [7]. Figure 3 and 4 shows the input and output of spaCy NER tool.

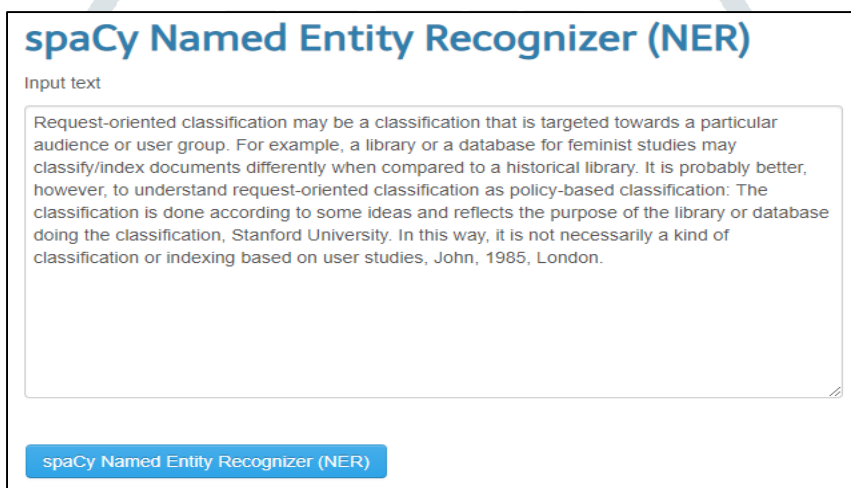


Fig 3. spaCy NER Input

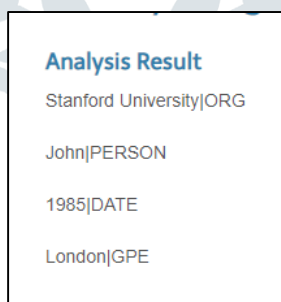


Fig 4. spaCy NER Output

2.3 Stanford Named Entity Tagger

Stanford NER is the java based important tool for Named Entity Recognizer. Named Entity Recognition (NER) tags the categorizations of words in a text which are the names of things, such as person and company names, or gene and protein names. It comes with well-engineered feature extractors for Named Entity Recognition, and many options for defining feature extractors [8]. Figure 5 and 6 shows the input and output of Stanford NER tool.

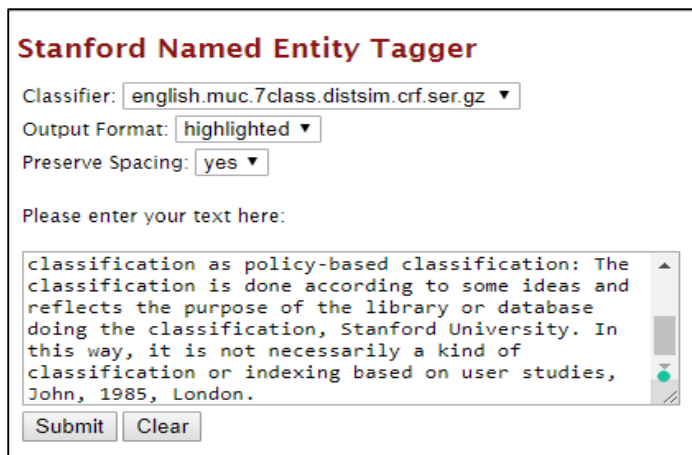


Fig 5. Stanford NER Input

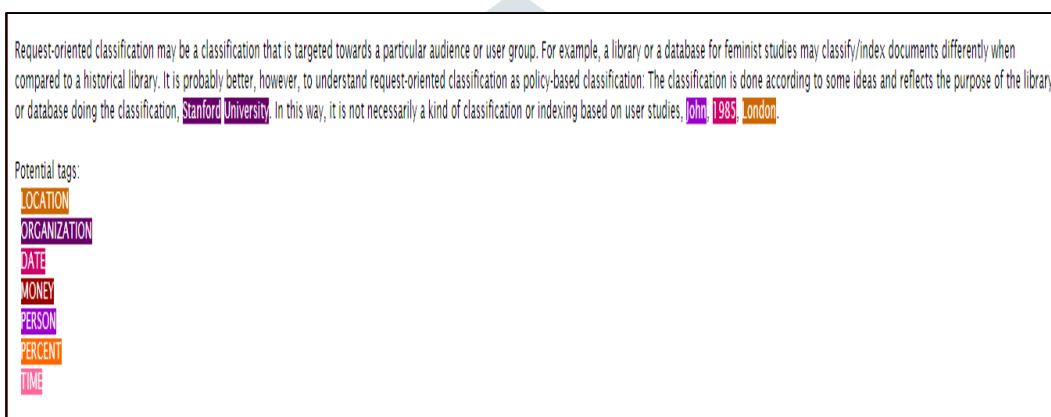


Fig 6. Stanford NER Output

2.4 ParallelDots

This is the best applied AI research groups in the world. They work with enterprises globally to tackle challenging business problems and create products that bring real value to real people. They also provide AI consulting services to explore the what, why, how and who about deploying AI in businesses. Named Entity Recognition can identify individuals, companies, places, organization, cities and other various type of entities. API can extract this information from any type of text, web page or social media network [9]. Figure 7 and 8 shows the input and output of ParallelDots NER tool.

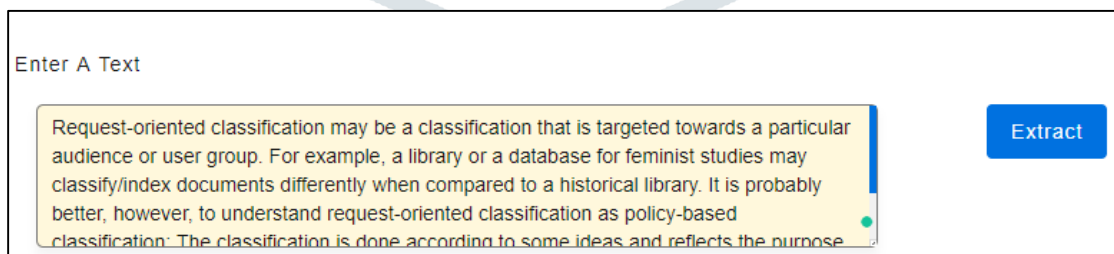


Fig 7. ParallelDots NER Input

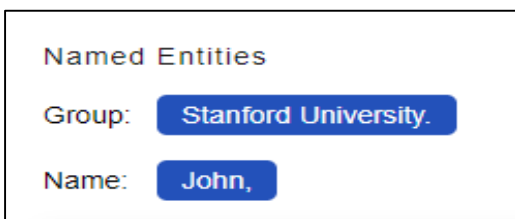


Fig 8. ParallelDots NER Output

2.5 Text Analysis API

An easy-to-use API used to perform a variety of complex NLP tasks on documents, reviews, social comments, or any other type of text. It analyzes the sentiment towards entities found in text. Extracts mentions of named entities (Person, Organization, Location), associates a type and links them to DBpedia (where possible), and evaluates sentiment towards each of the entities. This endpoint includes some of the functionality of the Entity Extraction and Concept Extraction endpoints [10]. Figure 9 and 10 shows the input and output of Text Analysis API tool.

The screenshot shows the input interface of the Text Analysis API. At the top left, there is a label 'Input' with a hamburger menu icon and the word 'Text'. Below this is a large text area containing the following text: "Request-oriented classification may be a classification that is targeted towards a particular audience or user group. For example, a library or a database for feminist studies may classify/index documents differently when compared to a historical library. It is probably better, however, to understand request-oriented classification as policy-based classification: The classification is done according to some ideas and reflects the purpose of the library or database doing the classification, Stanford University. In this way, it is not necessarily a kind of classification or indexing based on user studies, John, 1985, London." Below the text area is a 'Language:' label with a dropdown menu currently showing 'English'. At the bottom right, there is a blue button with a gear icon and the text 'Analyze'.

Fig 9. Text Analysis API Input

Entity	Overall Sentiment	Type	Mentions
Stanford University	Neutral 0.50	Organization	1
John	Neutral 0.51	Person	1
London	Neutral 0.46	Location	1

Fig 10. Text Analysis API Output

2.6 displaCy Named Entity Visualizer

Explosion AI is a digital studio specializing in Artificial Intelligence and Natural Language Processing. They have design custom algorithms, applications and data assets. They are the creators of spaCy, the leading open-source library for advanced NLP and Prodigy, a new annotation tool for radically efficient machine teaching [11]. Figure 11 and 12 shows the input and output of displaCy Named Entity tool.

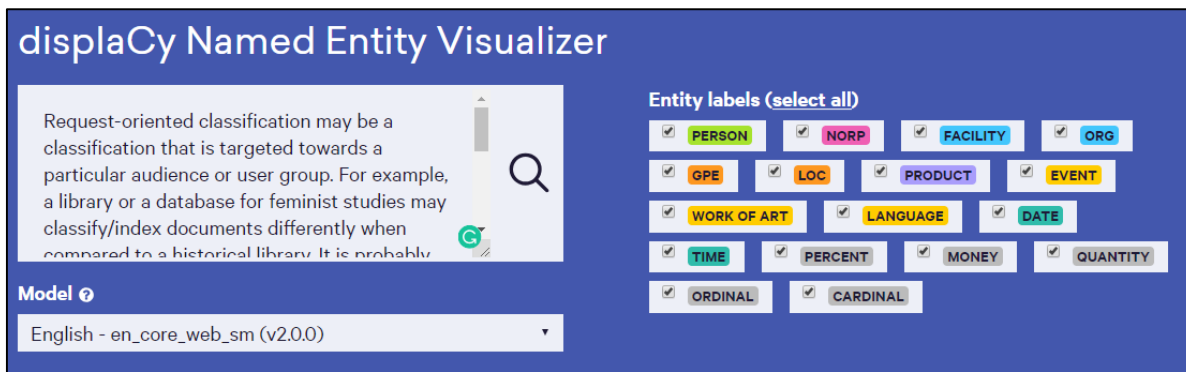


Fig 11. displaCy Named Entity Input

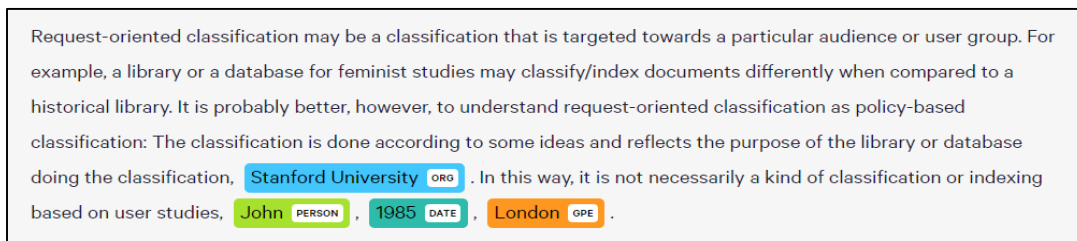


Fig 12. displaCy Named Entity Output

2.7 TextRazor

TextRazor was designed to make any text classification or extraction project easy. Here they will go over a few simple use cases to give you a starting point. Examples are given using our Python SDK for convenience, but the same concepts equally apply to other languages or the REST API [12]. Figure 13 and 14 shows the input and output of TextRazor tool.

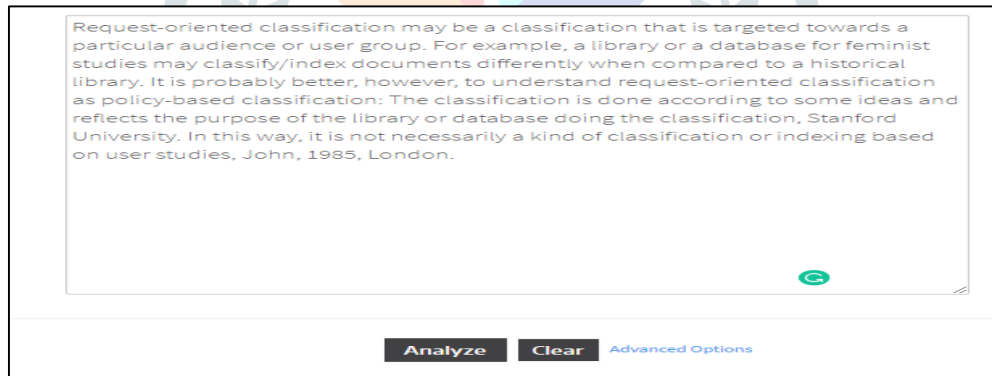


Fig 13. TextRazor Input

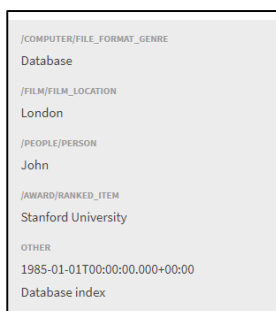


Fig 14. TextRazor Output

2.8 Cognitive Computing Group

Understanding text to the level that we can extract information from it in an intelligent way and answer questions with respect to it requires the ability to identify different types of entities and categories in text. E.g., this phrase represents a name of a person, an organization, a location and other semantic categories. This is a context sensitive problem ("Washington" is a location in one context and a person in another) and machine learning techniques are used to resolve this and determine the appropriate semantic category of entities [13]. Figure 15 and 16 shows the input and output of Cognitive Computing NER tool.

Request-oriented classification may be a classification that is targeted towards a particular audience or user group. For example, a library or a database for feminist studies may classify/index documents differently when compared to a historical library. It is probably better, however, to understand request-oriented classification as policy-based classification: The classification is done according to some ideas and reflects the purpose of the library or database doing the classification, Stanford University. In this way, it is not necessarily a kind of classification or indexing based on user studies, John, 1985, London.

Submit

Fig 15. Cognitive Computing NER Input

Request-oriented classification may be a classification that is targeted towards a particular audience or user group. For example, a library or a database for feminist studies may classify/index documents differently when compared to a historical library. It is probably better, however, to understand request-oriented classification as policy-based classification: The classification is done according to some ideas and reflects the purpose of the library or database doing the classification, [ORG Stanford University] . In this way, it is not necessarily a kind of classification or indexing based on user studies, [PER John] , 1985, [LOC London] .

Key

PER	Person
ORG	Organization
LOC	Location
MISC	Miscellaneous

Fig 16. Cognitive Computing NER Output

2.9 NLTK

This is a demonstration of NLTK part of speech taggers and NLTK chunkers using NLTK 2.0.4. These taggers can assign part-of-speech tags to each word in your text. They can also identify certain phrases/chunks and named entities [14]. Figure 17 and 18 shows the input and output of NLTK NER tool.

Tag and Chunk Text

Choose tagger/chunker
Default Tagger & NE Chunker

Enter text

Request-oriented classification may be a classification that is targeted towards a particular audience or user group. For example, a library or a database for feminist studies may classify/index documents differently when compared to a historical library. It is probably better, however, to understand request-oriented classification as policy-based classification: The classification is done according to some ideas and

Enter up to 50000 characters

Tag & Chunk

Fig 17. NLTK NER Input

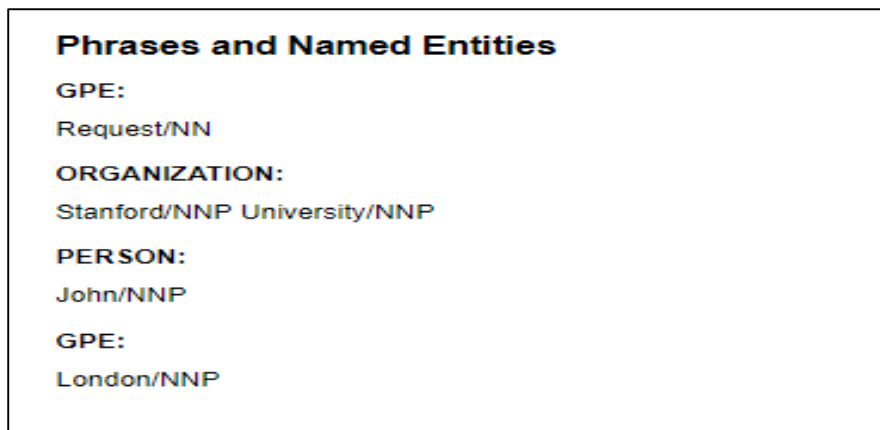


Fig 18. NLTK NER Output

III. PERFORMANCE ANALYSIS

In order to perform the comparative analysis of the NER tools, there are two performance measures are used; limitations and output. Limitations describe the type of classes. Output helps to find how the tool extracted the entities from the natural text. Each tool has produced different output for the same input document. Table 1 provides the performance of the nine open source tokenization tools.

Table 1: Performance Analysis of NER tools

Tool Name	Limitation	Output
Dandelion API	This tool is used to extract more types of entities like person, concept etc., Compare to other tools, this will extract the concepts alone	
spaCy	spaCy is used to identify the entities from the text but the entities will alone it will display.	
Stanford Named Entity Tagger	This tool will also gives the better results like Dandelion but, there is no concept class.	

<p>ParallelDots</p>	<p>This will extract only the group and name entities from the text.</p>	<div data-bbox="703 176 1091 405"> <p>Named Entities</p> <p>Group: Stanford University.</p> <p>Name: John,</p> </div>																
<p>Text Analysis API</p>	<p>It will analyze the sentiment then it will extract the entities based on sentiment.</p>	<table border="1" data-bbox="703 443 1417 680"> <thead> <tr> <th>Entity</th> <th>Overall Sentiment</th> <th>Type</th> <th>Mentions</th> </tr> </thead> <tbody> <tr> <td>Stanford University</td> <td>Neutral 0.50</td> <td>Organization</td> <td>1</td> </tr> <tr> <td>John</td> <td>Neutral 0.51</td> <td>Person</td> <td>1</td> </tr> <tr> <td>London</td> <td>Neutral 0.46</td> <td>Location</td> <td>1</td> </tr> </tbody> </table>	Entity	Overall Sentiment	Type	Mentions	Stanford University	Neutral 0.50	Organization	1	John	Neutral 0.51	Person	1	London	Neutral 0.46	Location	1
Entity	Overall Sentiment	Type	Mentions															
Stanford University	Neutral 0.50	Organization	1															
John	Neutral 0.51	Person	1															
London	Neutral 0.46	Location	1															
<p>displaCy Named Entity Visualizer</p>	<p>This tool will also give the better results like Dandelion but, there is no concept class.</p>	<div data-bbox="703 714 1417 958"> <p>Request-oriented classification may be a classification that is targeted towards a particular audience or user group. For example, a library or a database for feminist studies may classify/index documents differently when compared to a historical library. It is probably better, however, to understand request-oriented classification as policy-based classification: The classification is done according to some ideas and reflects the purpose of the library or database doing the classification, Stanford University ORG. In this way, it is not necessarily a kind of classification or indexing based on user studies, John PERSON, 1985 DATE, London GPE.</p> </div>																
<p>TextRazor</p>	<p>This tool will also give the better results like Dandelion but, there is no concept class.</p>	<div data-bbox="703 994 1225 1346"> <pre> /COMPUTER/FILE_FORMAT_GENRE Database /FILM/FILM_LOCATION London /PEOPLE/PERSON John /AWARD/RANKED_ITEM Stanford University OTHER 1985-01-01T00:00:00.000+00:00 Database index </pre> </div>																
<p>Cognitive Computing Group</p>	<p>It is used to extract the four entities from the text.</p>	<div data-bbox="703 1379 1417 1785"> <p>Request-oriented classification may be a classification that is targeted towards a particular audience or user group. For example, a library or a database for feminist studies may classify/index documents differently when compared to a historical library. It is probably better, however, to understand request-oriented classification as policy-based classification. The classification is done according to some ideas and reflects the purpose of the library or database doing the classification, [org Stanford University]. In this way, it is not necessarily a kind of classification or indexing based on user studies, [per John], 1985, [loc London].</p> <p>Key</p> <table border="1" data-bbox="703 1547 1417 1785"> <tr> <td style="background-color: #ff0000; color: white; padding: 2px;">PER</td> <td>Person</td> </tr> <tr> <td style="background-color: #0000ff; color: white; padding: 2px;">ORG</td> <td>Organization</td> </tr> <tr> <td style="background-color: #0000ff; color: white; padding: 2px;">LOC</td> <td>Location</td> </tr> <tr> <td style="background-color: #0000ff; color: white; padding: 2px;">MISC</td> <td>Miscellaneous</td> </tr> </table> </div>	PER	Person	ORG	Organization	LOC	Location	MISC	Miscellaneous								
PER	Person																	
ORG	Organization																	
LOC	Location																	
MISC	Miscellaneous																	
<p>NLTK</p>	<p>This tool will also give the better results like Dandelion but, there is no concept class.</p>	<div data-bbox="703 1818 1169 2047"> <p>Phrases and Named Entities</p> <p>GPE: Request/NN</p> <p>ORGANIZATION: Stanford/NNP University/NNP</p> <p>PERSON: John/NNP</p> <p>GPE: London/NNP</p> </div>																

IV. CONCLUSION

In information extraction, the Named Entity Recognition (NER) is the sub problem and it is used to identify the entities such as, name, place, concept and organization. It encompasses the processing of structured and unstructured documents. In Natural Language Processing (NLP) system, the NER is the essential task and this process is the core of NLP. This research analyses the performance of nine open source NER tools. Some of the tools will extract the limited number of entities. By analyzing all the measures, the Dandelion API will give the better results when compared to other NER tools. In future, there is need to develop a NER tool for all languages and for more entities.

REFERENCES

- [1]. R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history," in 16th Conference on Computational linguistics, 1996, pp. 466-471.
- [2]. O. Borrega, M. Taulé, and M. A. Martí, "What do we mean when we speak about Named Entities," in Conference on Corpus Linguistics, 2007.
- [3]. H. Cunningham, "Information extraction, automatic," in Encyclopedia of Language and Linguistics, 2nd ed., Elsevier, 2005, pp. 665-677.
- [4]. J. Gantz and D. Reinsel, "The Digital Universe Decade, Are You Ready?," 2010.
- [5]. V. Uren et al., "Semantic annotation for knowledge management: Requirements and a survey of the state of the art," Journal of Web Semantics, vol. 4, no. 1, pp. 14-28, 2006.
- [6]. <http://nlp.stanford.edu:8080/ner/process>
- [7]. <https://www.paralldots.com/named-entity-recognition>
- [8]. <https://developer.aylien.com/text-api-demo?text=http%3A%2F%2Fwww.businessinsider.com%2Fcarl-icahn-open-letter-to-apple-2014-1&tab=entities&run=1>
- [9]. <https://text-processing.com/demo/tag/>
- [10]. http://cogcomp.org/page/demo_view/ner
- [11]. <https://www.textrazor.com/demo>
- [12]. <https://explosion.ai/demos/displacy-ent>
- [13]. <https://textminingonline.com/text-analysis-online-no-longer-provides-nltk-stanford-nlp-api-interface>
- [14]. <https://dandelion.eu/semantic-text/entity-extraction-demo/>