COMPARATIVE STUDY OF ANNOTATION TOOLS AND TECHNIQUES

A Thesis Submitted to the Department of Computer Science

At the African University of Science and Technology

In Partial Fulfilment of the Requirements for the Degree of

MASTER of Computer Science

By

Musabeyezu, Fortunee

Abuja, Nigeria

June, 2019

Certification

This is to certify that the thesis titled ‘Comparative Study of Annotation Tools and Techniques’ submitted to the school of postgraduate studies, African University of Science and Technology (AUST), Abuja, Nigeria for award of Master’s degree is a record of original research carried out by MUSABEYEZU FORTUNEE In the Department of Computer Science.

African University of Science and Technology [AUST]

Knowledge is Freedom

APPROVAL BY

Supervisor

Surname: DAVID

First name: Amos

Signature:

The Head of Department

Surname: DAVID

First name: Amos

Signature:

Abstract

A huge amount of data is generated on daily basis. The generated data can be both structured and unstructured data. The sources from which most of the unstructured data are found are the dailies, social networks (posts from Facebook, tweeter, etc.), event reporting (for example recounting an accident), etc.

One of the biggest challenges in Big Data analysis is the use of unstructured data. There is need to structure the corpus so as to permit analysis and one of the approaches for structuring unstructured data is the technique of annotation. Annotation could be fully automatic, semi-automatic, or fully manual (human).

The technique of annotation has been of important help in different domains and sectors (machine learning, education, health, commerce, etc.). For example, in machine learning especially for supervised learning where annotation is used in the training phase to label data.

In this research we studied and analysed different annotation tools and techniques. The studied tools were tested and their most important features that should be taken into consideration when choosing a tool were used for the comparison.

Keywords: unstructured data, annotation, annotation tools, annotation techniques, ontology, semantic web, semantic annotation

Acknowledgement

I deeply appreciate Prof. Amos David who doubles as my supervisor and HOD for his tireless guidance and motivation throughout the period of this research, especially the faith he has in me.

I thank other faculty members of Computer Science, AUST, for impacting their wealth of knowledge and experience during coursework.

I am indebted to the Acting AUST President, Prof. C.E. Chidume, the entire management and members of AUST community for the opportunity to study on full scholarship and their gesture of love. I thank you all immensely for all the hospitality you offered me.

I also acknowledge all my classmates for all the academic and moral support I received. Your patience and dedication towards my success in AUST will never be forgotten.

Finally, I appreciate my family for their fervent prayers, emotional and financial support.

Table of content

List of tables

List of figures

List of Abbreviations

IT	Information Technology
URL	Uniform Resource Locator
PHP	Hypertext Pre-processor
EI	Economic Intelligence
S-Cream	Semi-automatic Creation of metadata
PDF	Portable Document Format
HTML	Hypertext Mark-up Language
XML	Extensible Markup Language
RDF	Resource Description Framework
NPL	Natural Processing Language
CSV	Comma Separated Values
HTTP	HyperText Transfer Protocol
CoNLL	Computational Natural Language Learning
JSON	Java Script Object Notation
JSONL	Java Script Object Notation Lines
JDK	Java Development Kit
AWS	Amazon Web Services
W3C	World Wide Web Consortium

Chapter One

Introduction

1.1 Research and Background

In recent years, big data has become a salient research topic in the IT industry. A huge volume of data is generated on daily basis, owing to the fact that numerous numbers of devices are connected to the internet. The huge volume of data come from different sources, such as the World Wide Web, Ecommerce, and social media platforms like Facebook, twitter etc. Big data can be of two types, structured and unstructured data; structured data are the usual data that can be stored in a relational database, the query of these data is easy and analysis on them can be seamlessly done (Kiefer, 2016). The unstructured data, are usually data from email, pictures, document, video files, audio and other sources. It is difficult to process these unstructured data with relational databases, hence there is the challenge of managing such data. In this research the technique of annotation will be studied; various annotation tools and techniques have been proposed, that help in the analysis of big data.

The technique of annotation has played a considerable role in different sectors (education, health, commerce etc.) for example in machine learning, annotation tool can be used to train data sets.

Annotation has different definition, depending on the context of its use. Annotation is a way of interpreting document (Okunoye, Oladejo, & Odumuyiwa, 2010). There are different annotation forms such as labelling an object, making a comment, tagging images, audio and videos etc. Annotating a document makes the document to be more detailed, informative and also makes the document to be easily queried, hence it adds value to a document.

Annotation may be of two types, that is implicit and explicit annotation (Okunoye et al., 2010). Implicit annotation is one which is assumed to be only understood by the maker. Unlike implicit annotation, explicit annotation is one that the meaning of the annotation is assumed to be known by a group, team or users of the same field of study (Okunoye et al., 2010).

Annotation as object is defined as an intentional and topical value-adding note linked to an extant information object (Bodain & Robert, 2007). Annotation is also defined as “any object (annotation) that is associated with another object (document) by some relationship” (Brusilovsky, 2005). The definition of annotation by (Brusilovsky, 2005) does not only consider annotation as object but also as an action involving anchoring the object with the concerned document. Annotation as action is defined as an act of interpreting a document (Robert, 2007). It is a process of creating annotation as object and anchoring it to the document object (i.e. information source being annotated).

Annotation is also defined as a way of attaching extra information (metadata) to a database record to provide better understanding and connective to the related information. Annotation can be manual, semi-automatic or automatic.

- **Automatic annotation**: makes use of computerized automated tools to annotate a document.
- **Semi-automatic**: makes use of computerized automated tools to annotate a document but also requires human intervention.
- **Manual annotation**: the annotation is totally done by human annotator.

1.2 Problem statement

There is some useful information that can be got from unstructured data when they are processed and analysed. Many annotation tools and techniques can be used to structure unstructured data, so because of numerous numbers of the available annotation tools and techniques out there it may be difficult to choose a particular tool which will be suitable for a pertaining data or available operating environment. There is need to study and analyse different annotation tools and techniques, to know the usefulness, usability, strength and the type of data they are most suitable for.

This thesis proposes a comparative study of different annotation tools and techniques, to facilitate in decision making for the best tool to employ when faced with a problem solvable by annotation.

1.3 Research aim and objective

1.3.1 Aim

A comprehensive review of existing proposals in the field of annotation (theses, articles, software). The result should be in form of a comparative table of existing techniques and tools. The tools discovered should be tested.

1.3.2 Objective

- Study and analyse different existing annotation tools and techniques.
- Test the discovered tools and techniques.
- Compare the studied annotation tools and techniques.
- Make a comprehensive conclusion and suggestion on the best tools suitable for different context of use.

1.4 Limitation of study

This research study is limited to comparing a few existing annotation tools and techniques that can be used in processing unstructured data.

1.5 Research outline

The entire research is divided into five chapters. Each chapter highlighted different topics and subtopics as follows:

Chapter 2 discusses the basic concepts and literature review related to annotation tools and techniques. Chapter 3 studies, analyses and compares different annotation tools. Chapter 4 presents the comparison results and discussion. Lastly, Chapter 5 contains summary, conclusion, recommendations, and future work.

Chapter Two

Literature review

This chapter presents some basic terminologies on the subject matter with a review of works done related to the research.

2.1 What is annotation?

Annotation has been one of important topics under research for the past years and it has played a considerable role in data management.

Annotation is a way of interpreting a document or commenting on a document.it adds extra information to it and makes it easier to browse, search, retrieve, categorize and analyse (Okunoye et al., 2010). It can also be understood as a way of creating semantic labels within a document (Pernelle & Nathalie, 2017).

Annotation structures data and facilitates easy storage and access of data. It helps in information sharing and knowledge elicitation by reading one's perception on a particular document of interest.

2.2 Types of annotations

The formats of the document to be annotated determines the type of annotation. The annotated document can be of different formats: text, image, videos.

2.2.1 Text annotation

Text annotation is an act of adding extra note to a text. This can be achieved by adding marks, putting footnotes, highlighting or underlining a text of interested. Annotating a text adds value to it, makes it more informative and allow user own to integrate his/her interpretation with the text (Gosal, 2015).

2.2.2 Image annotation

Image annotation is a process of adding descriptive captions or tags (location, time, etc.) or keywords to the image to make it more accessible and more descriptive so that it can easily be stored, searched, categorized, retrieved and recognized (Hanbury, 2008).

2.2.3 Video annotation

Video annotation is a process of adding captions or keywords that add extra information to the video in order to facilitate videos access and retrieval from a large video database (Dasiopoulou, Giannakidou, Litos, Malasioti, & Kompatsiaris, 2011). It also helps in storing, browsing, searching, categorization of videos.

2.3 Annotation techniques

Based on the type of data, time and annotation accuracy that one wants to achieve, one may choose among the three annotation techniques. Annotation can either be manual, semi-automatic or automatic.

2.3.1 Manual annotation

Manual annotation is a way of transforming the extant syntactic resources into interconnected structures buy putting extra information to a document or part of document which forms metadata (Pernelle & Nathalie, 2017). Manual annotation is more accurate than automatic annotation but is expensive and does not take into consideration multiple perspectives (Pernelle & Nathalie, 2017).

2.3.2 Automatic annotation

Automatic annotation has become a solution to big data and large datasets problem, where manual annotation cannot be applied (Pernelle & Nathalie, 2017) . Despite that, the automatic annotation is not accurate and full of error-prone. Thus, semi- automatic annotation has been created to solve the problems of manual and automatic annotations and is used in most of the current annotation tools (Pernelle & Nathalie, 2017).

There are many proposed automated annotation methods such as supervised machine learning based methods which is composed by two phases: training and annotation. In the annotation phase, entities and different relations between entities are identified (Pernelle & Nathalie, 2017).

The main idea is to build an ontology based pattern, then the built ontology is used to automatically extract the desired information from the data (Li, Tang, Li, & Luo, 2009).

2.3.3 Semi-automatic annotation

The mixed method of manual and automatic annotation called semi-automatic annotation, combines automatic recognition technologies and human intervention (Pernelle & Nathalie, 2017). This set of annotation systems can differ in different features such as architecture, performance, methods and tools of information extraction, the manual work amount required to achieve annotation, storage management and other features (Slimani, 2013).

2.3.4 Comparison of annotation techniques

Annotation techniques	Manual	Semi-automatic	automatic
Human task	Defining and putting descriptive keywords	Review and make corrections on the automated annotations	Verify the outputs
Computer task	Provide space for annotations	Automatically generate annotations with help of recognition technologies	Automatically provide descriptive keywords using recognition technology
Advantages	High accuracy of annotations	High quality of annotations due to the interaction between human and machine	Less time consuming
Disadvantages	Expensive and time consuming	expensive	Inaccurate annotations due to error-prone

Table 1: comparison of annotation techniques

2.4 Advantages of annotation

Annotation plays a considerable role in different domains; here are some benefits of annotation:

- Easy access of data
- It structures data for data analysis and storage prepares data for data visualization
- It is used in machine learning to train data sets
- It makes a document more informative and understandable
- It makes the search and retrieval of data easier
- It helps in information sharing and knowledge elicitation

2.5 Metadata

Data about other data are called metadata. For example, ISBN and author's name are data about a novel. In a database, the data types are the metadata describing data. Metadata is also data; the difference lies in the intended usage of the data and in the subject of metadata. It's metadata that is used to search and discover the information. Annotation can also be seen as metadata or data added to raw data (Passin, 2004).

2. 6 Semantic Web

Semantic shows that the meaning of data on the web can be discovered by both people and computers.(Passin, 2004). In contrast, most of today's data meaning on the web is deduced by people who read web pages and the labels of hyperlinks, and other people who implement specialized software to work with data.

The term “Semantic Web” stands for a vision in which computers software as well as people can find, read, understand, and use data over the World Wide Web to accomplish useful goals for.

2.7 Semantic annotation

Semantic annotation is a way of generating metadata and makes use of schema to enable new information access methods and to extend the existing ones (Kiryakov, Popov, Terziev, Manov, & Ognyanoff, 2004). Automatic semantic annotations facilitate many new forms of annotations: highlighting, indexing and retrieval, categorization, generation of more advanced metadata, smooth traversal between unstructured text and available relevant knowledge (Kiryakov et al., 2004). Semantic annotation can be applied for any text format (non-web documents, web documents, text fields in databases, etc. Further, knowledge acquisition can be performed on the basis of the extraction of more complex dependencies analysis of relationships between entities, event and situation descriptions (Kiryakov et al., 2004).

2.8 Ontology

Ontology stands for a typical knowledge representation by a set of concepts describing a domain of interest and relationships between those concepts. Depending on the precision of this specification, the notion of ontology encompasses several data or conceptual models, e.g., classifications, database schemas, fully axiomatized theories(Euzenat & Shvaiko, 2013).

2.9 Related work

(Manthalu, 2014) proposed an annotation tool for web search results. He stated that the semantic web is not structured, the web search results can include varying types of information relating to the same query. Such results cannot directly be analyzed to meet specific interpretation. It is necessary the web search result records for so that they can be processed by machine. Annotating web search results will add value to the search result record, for them to be stored for further analysis and for them to be read and understood easily (Manthalu, 2014) .The main purpose of his work was to automatically annotate and restructure the web search result record to allow data visualization for users, with help of a web crawler that was used to fetch websites related to public health domain. A search engine was used to search information from the crawled websites and when the search result records are being displayed, they are automatically put into different categories based on the theme of content in the document and they are automatically annotated by assigning labels to web pages. Table annotation was used and different features that can help to identify a webpage and assign an appropriate label to it such as: URL of a website, title of a website, file type, etc. In implementing the proposed tool, a web crawler was used for fetching and indexing documents, a search engine and a classifier for categorizing search result records. The software was implemented in PHP with MySQL as the backend database.

(Okunoye et al., 2010) Proposed annotation tool that helps in knowledge elicitation. In their work, they stated that Economic Intelligence is the current leading approaches that are used to solve organization's decision making. This approach involves interaction among Economic Intelligence Actors.

Many companies are making investment on tools that can help in sharing knowledge among their employees (Okunoye et al., 2010). In their work, proposed an annotation model that can facilitate knowledge elicitation among Economic Intelligence Actors. The knowledge elicitation or collaboration among EI Actors is achieved by the fact that the annotator or the person who makes annotation captures his/her interpretation and expresses his/her knowledge on a document of interest in such a way that other EI Actors will be able to understand the annotator's perception.

This could be done either synchronously where all actors have to be online at the same time, or asynchronously where they don't need to be online but they are notified of a pending message (Okunoye et al., 2010). The collaboration among different actors helps in decision making. One actor should be able to express his/ her own perception in such a way that another actor should be able to analyze and understand it. Unlike most of the annotation tools that users can only add values to pre-defined attributes, the proposed system presents annotation as attribute-value pair, which is user defined. With this annotation tool, the user is able to define his/her own attribute to represent the semantic of his/her annotation and associate a value to the defined attribute (Okunoye et al., 2010). The approach of managing interactions was achieved by enabling group awareness among actors. Awareness as defined by Dourch: is an understanding of other's activities which provides context for your own activities. (Gosal, 2015) Surveyed the semantic annotation of text. He stated that when the semantic annotation is done manually, is a very expensive activity and often does not take into consideration the multiple perspectives of a data source. There is need to make use of automatic annotation the in order to make the annotation of the existing document scalable and reduce the burden of annotating new documents considering we have to deal with large collections of data.

The automatic annotations bring with them the benefits of improved information retrieval and enhanced interoperability (Gosal, 2015). They studied the utilization of semantic annotation some issues related to automatic representation.

Some important text annotation tools were reviewed and analyzed. The reviewed text annotation tools are: Annotea, Cohse, MnM and S-Cream (Gosal, 2015).

(Hanbury, 2008) Did a survey on image annotation methods. He stated that, the use of image annotation in the creation of ground truth for the evolution of object recognition and automated image algorithms were discussed and three main a methods that are used to annotate images were reviewed (Hanbury, 2008). The three methods that are mostly used to annotate images are: Free text annotation: user uses descriptive text to annotate the image; Keyword annotation: arbitrary keywords or predefined keywords are added to the image and Ontology based annotation: this is specialization of conceptualization. The generation of keyword vocabulary in automated image evaluation was discussed.

(Khurana & Chandak, 2013) Reviewed different video annotation techniques. They stated that Annotating a video plays a meaningful role in improving the way a video can be searched, retrieved and accessed (Khurana & Chandak, 2013). Different video annotation techniques were studied. Some of annotation techniques were discussed. Some of the techniques used to annotate videos were discussed such as free textual description: which is a text that describes a video that is added to it to make its search and retrieval easier; Based on text in the video: the text that appears enclosed within the video especially captions, they provide very useful information that describes the video; Based on rule learning.

Different machine learning techniques like Bayesian network, clustering, support vector machine, similarity and metric learning are used to extract low-level features; Based on rule learning: there exists a gap between the low-level visual feature of a video and the interpretation that the same data has for a user.

Rules are set to deduce a set of high-level concepts from low-level descriptors; Based on graph: graph-based learning is used for annotation of different modularity Based on ontology (Khurana & Chandak, 2013): Ontology is a large classification that classifies different aspect of life into hierarchical categories. Using ontology-based technique, helps in semantic annotation. Machine learning along with ontology can give better annotations compared to other techniques (Khurana & Chandak, 2013).

(Abahai, 2015) proposed automated document annotation for effective data sharing system. In this system, annotation was done by adding additional information to the document when it is being created. The attribute to be added to the document are recommended based on content score and query score (Abahai, 2015). The system is composed by two phases: annotation phase and search phase. In annotation phase, the author uploads the document into the repository and the system recommends the appropriate annotation attribute depending on high querying score and high content score (Abahai, 2015). Query score is the frequency of occurrence of an attribute in the query collection; and Content score is the frequency of occurrence of an attribute in the document (Abahai, 2015). However, the system can allow the author to create his/her own attribute for annotation. This system can be used in data mining where you need to extract important data from a huge amount of information.

(Bipboy, 2017) Reviewed annotation Tools to enhance learning.

Social annotation tools are used more often in contemporary learning processes in order to boost learning. Nowadays, plenty models; types and applications have been created that facilitate annotation to online documents, web pages, even multimedia (Bipboy, 2017).

In his work, twenty-four annotation tools that help in improving reading comprehension were reviewed, analyzed and compared based on three criteria: knowledge dissemination, usability of the tool and effectiveness of the tool. The reviewed annotation tools were compared based on their features and properties. The parameters used to compare the reviewed annotation tools are: annotated target, form of annotated data, searching option of annotated data, sharing option of annotations, social annotation or not, private annotation or not and commenting option for other's annotation.

(Nixon & Troncy, 2014) Surveyed semantic media annotation tools. In their work, they set the linked media principles which media annotation approaches should conform to. The current media annotation tools that do not conform to those principles were reviewed and the two emerging toolsets that support linked media principles conformant annotation were presented. A survey of the existing annotation tools showed that their implementation doesn’t conform to those linked media principles.

(Raut & Sawarkar, 2016) Surveyed unstructured document annotation using content and query-based values. They stated that in many organizations, most of the generated textual data contain a big amount of useful information which underlies in the unstructured text (Raut & Sawarkar, 2016).

Document annotation is a process of attaching metadata on the document which helps in information extraction(Raut & Sawarkar, 2016). An adaptive technique that automatically generate data input forms was presented.

They also suggested relevant and recommended attributes that can be used to annotate unstructured textual documents, such that the use of the uploaded data is maximized, given the user information needs. The properties that were used to suggest attribute for annotation are query value with respect to query workload and the content value with respect to the document.

They presented Probabilistic methods and algorithms that integrate the information from the query workload into the data annotation process, in order to generate metadata that are not just relevant to the annotated document, but also useful to the users querying the database (Raut & Sawarkar, 2016).

Chapter Three

Discussion of annotation Tools

Many annotation tools have been developed and they are available to satisfy user needs. Most of the annotation tools are accessible on cloud and have free trial versions for testing purpose, whereas others require buying license and custom installation or configuration on your local infrastructure.

In this chapter six annotation tools are studied and tested. This testing focuses on some parameters that can be used to evaluate the performance of an annotation tool.

3.1. Machine learning and annotation

Machine learning models are sometimes employed in automatic annotation by some annotation tools, and in return annotation tools are used in the training phase of some machine learning models, this is usually found in supervised learning, where the inputs(X) and output(y) need to be labelled.

3.2. Categories of annotation tools

Constrained on the target data types that are to be annotated by an annotation tool, we can categorize the annotation tools into three main categories:

- Text annotation tools
- Image annotation tools
- Video annotation tools

However, this study is restricted to six text annotation tools, and these tools are tested and their functionalities are discussed.

3.3 Textual data

Textual data can be defined as a collection of material that are systematically collected, consisting of written, printed, or electronically published words, usually either intentionally written or transcribed from speech; Textual data are any form of document that can be supported by a text annotation tool editor. Examples of textual data are PDF document (.pdf), a text document (.doc), a web pages, an RDF document (.rdf), HTML document (.html), XML document (.xml), etc.

3.4 Annotating text

Text annotation is an act of extracting text of interest from a document. this can be done by highlighting a text of interest, adding a comment or a note, defining entities of interest, creating relations between the entities, labelling documents within the corpus, labelling the entities, etc.

3.5 Text annotation tools

Manual annotation of text is costly and time consuming which led people to think of developing software and application platforms that can make the annotation task easier and more efficient. There are many text annotation tools and frameworks available that provide user friendly annotation interface. Studying and discussing about annotation tools, we look at the parameters that can be used to evaluate an annotation tool.

The following are the features and properties that can be taken into consideration to study and evaluate the discovered annotation tools:

- Availability of the tool
- Ease of use
- Target data that can be annotated by the tool

- Imported data format
- Exported data format
- Annotations types
- The annotation techniques
- Annotations accuracy
- Collaboration among annotators
- Operating platform and requirements
- Advantages of using the tools

The main criteria that lead us to choose the following tools to be studied and tested is the availability and accessibility of the tool for testing purpose. Our choice of the tools to be studied prioritizes the annotation tools that are available for free to test.

3.5.1 Tagtog annotation tool

Tagtog (Cejuela et al., 2014) is a web-based tool for annotating text, that supports Artificial Intelligence. It provides machine learning models that have been already trained to automatically annotate an uploaded document. Tagtog also allows the user to train his/her own machine learning model that can be used to automatically annotate documents in the future.

Tagtog is available on cloud as well as on-premises. There is no installation required for a user to start using Tagtog on cloud, all that is required is to register as a new user and login with the login details (username and password). Tagtog on-premises requires custom installation on your computer or any public cloud (Amazon, Google, Azure, etc.). Using Tagtog annotation tool on cloud could offer easy access of the tool at low cost, but it is always preferable to use it on-premises in case the user wants to achieve high privacy regulation requirements.

Tagtog has been known in the past years for its good annotation performance. Here are the main features presented by the tool:

- The tool allows the user to work on more than one project.
- Collaborative annotation.
- Entities can be normalized.
- Active learning, as the machine learning model is being trained as the annotator is annotating manually and the user can make corrections or wrong predictions of a machine learning model.
- Provides a searching tool that searches by concepts.
- The supported import file formats are: pdf, csv, plain text, source code, html, etc.
- Multilingual support: English, French, Swedish, Swahili, Chinese, Arabic, etc. Unicode support

3.5.1.1 Functionalities of Tagtog annotation tool

The following are annotations that can be defined in Tagtog:

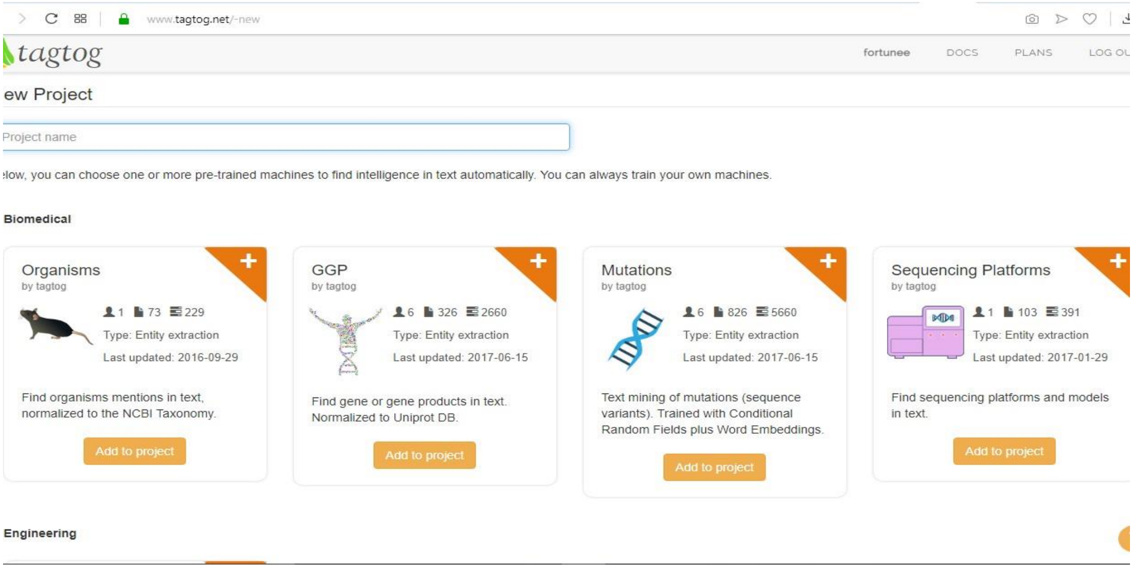
- **Entity creation:** Tagtog allows the user to create entities of interest. It can also automatically provide some real-world entities by default (person, organization, location, etc.).
- **Relation definition:** Tagtog annotation tool allows the user to define relationship between defined entities.
- **Entity labelling:** Tagtog annotation tool presents a feature that helps to label entities based on entity groups.

- **Normalization:** the normalization feature of Tagtog supports the import of resources from external databases (Wikipedia, google) related to the annotated span text.
- **Document labelling:** labels assigned to documents in the corpus helps in text classification.

With Tagtog annotation tool, annotation can be manual, semi-automatic or fully automatic.

3.5.1.1.1 Manual annotation in Tagtog annotation tool

With Tagtog, manual annotation is performed by human annotator in Tagtog annotation editor window. Tagtog online trial (“AI-enabled Text Annotation Tool | PDF, Markdown, CSV, html, tweets, &amp; many more types of Documents,” n.d.), which is available for free ,allows anybody who is willing to use it for testing purpose to register as a new user and create only one project.



ɘlow, you can choose one or more pre-trained machines to find intelligence in text automatically. You can always train your own machines.

Organisms — Type: Entity extraction — Last updated: 2016-09-29

Find organisms mentions in text, normalized to the NCBI Taxonomy.

GGP — Type: Entity extraction — Last updated: 2017-06-15

Find gene or gene products in text. Normalized to Uniprot DB.

Mutations — Type: Entity extraction — Last updated: 2017-06-15

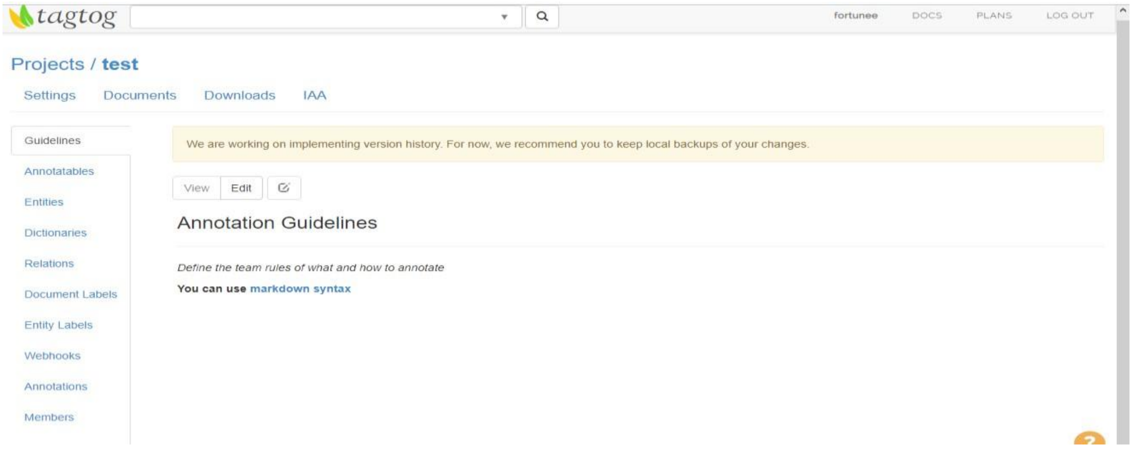
Text mining of mutations (sequence variants). Trained with Conditional Random Fields plus Word Embeddings.

Sequencing Platforms — Type: Entity extraction — Last updated: 2017-01-29

Find sequencing platforms and models in text.

Figure 1: Tagtog interface

In the process of creating a new project, for the manual annotations, the annotator does not need to choose a machine learning model from the displayed machine learning models, he/she just create a new project and give it a name. After the person must have created a project, he/she opens it by clicking on the project's name. In figure 2 below, the project name is test.



We are working on implementing version history. For now, we recommend you to keep local backups of your changes.

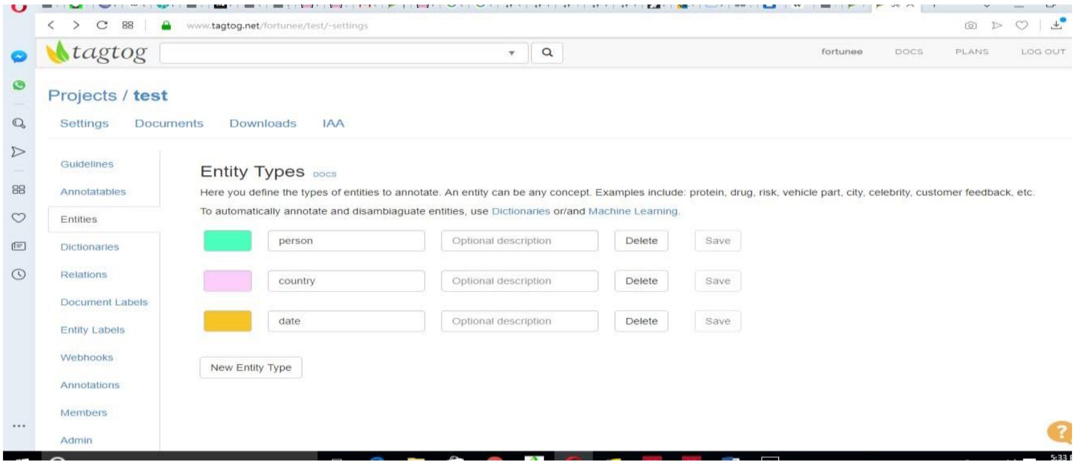
Annotation Guidelines

Define the team rules of what and how to annotate

You can use markdown syntax

Figure 2: the displayed window when a project is opened

On the displayed window, go to setting to see different text annotations that can be defined are displayed. You can create entities, relations, entity labels, document labels, normalization, etc. for the testing purpose, with the online trial, a project named test is created.



Entity Types DOCS

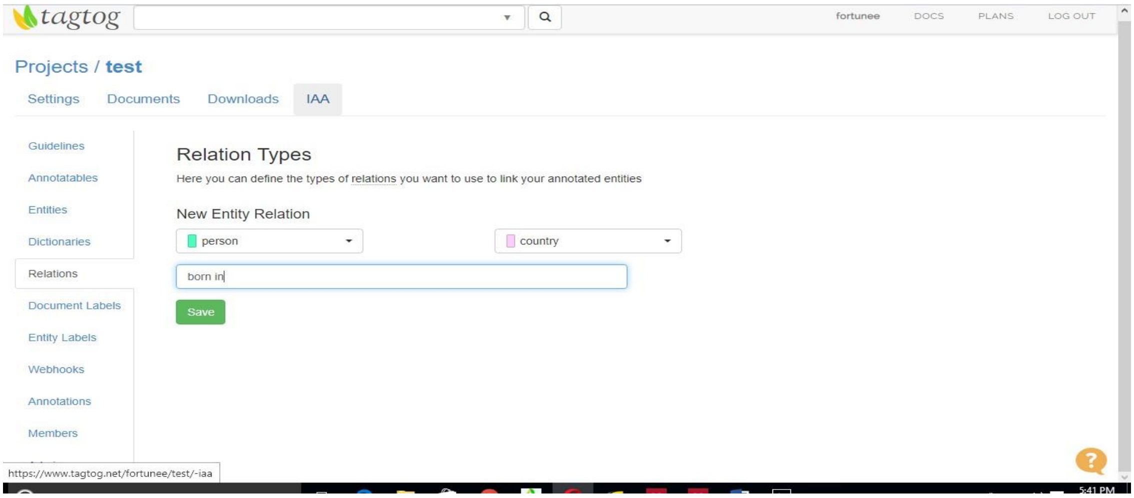
Here you define the types of entities to annotate. An entity can be any concept. Examples include: protein, drug, risk, vehicle part, city, celebrity, customer feedback, etc.

To automatically annotate and disambiaguate entities, use Dictionaries or/and Machine Learning.

🟩	person	Optional description	Delete	Save
🟪	country	Optional description	Delete	Save
🟨	date	Optional description	Delete	Save

New Entity Type

Figure 3: interface to create or delete an entity



Relation Types

Here you can define the types of relations you want to use to link your annotated entities

New Entity Relation

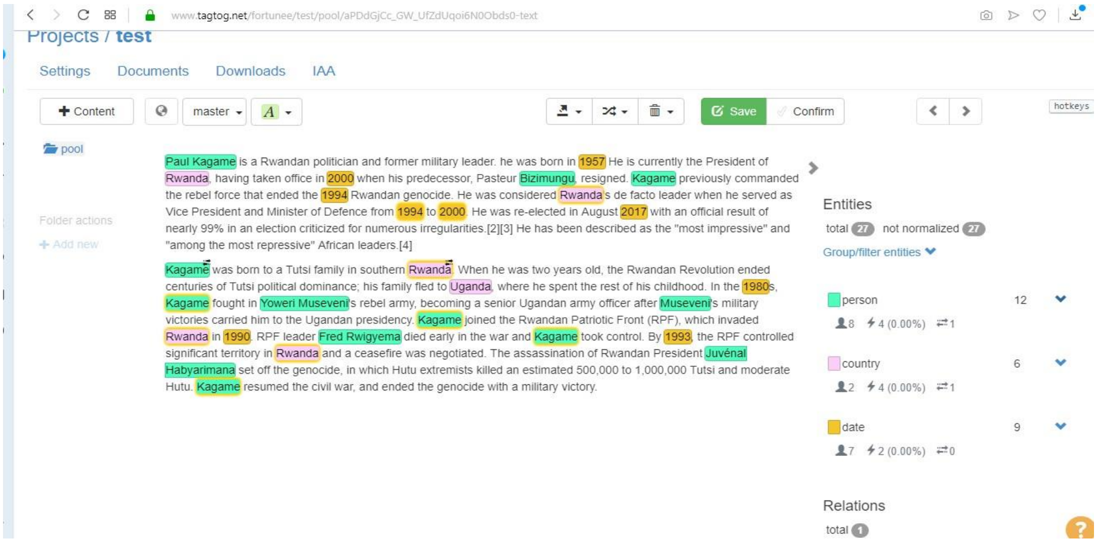
🟩 person

country

born in

Figure 4: Relation creation interface

Figure 4 above, shows how a relation named born in created between entity person and entity country. After creating entities and relations, the next step is to upload a text document that is to be annotated. To upload, the user clicks on the Documents tab, which lets he/she browse through files in the local machine, selects and upload the intended file.



Paul Kagame is a Rwandan politician and former military leader. he was born in 1957 He is currently the President of Rwanda, having taken office in 2000 when his predecessor, Pasteur Bizimungu, resigned. Kagame previously commanded the rebel force that ended the 1994 Rwandan genocide. He was considered Rwanda's de facto leader when he served as Vice President and Minister of Defence from 1994 to 2000. He was re-elected in August 2017 with an official result of nearly 99% in an election criticized for numerous irregularities.[2][3] He has been described as the "most impressive" and "among the most repressive" African leaders.[4]

Kagame was born to a Tutsi family in southern Rwanda. When he was two years old, the Rwandan Revolution ended centuries of Tutsi political dominance; his family fled to Uganda, where he spent the rest of his childhood. In the 1980s, Kagame fought in Yoweri Museveni's rebel army, becoming a senior Ugandan army officer after Museveni's military victories carried him to the Ugandan presidency. Kagame joined the Rwandan Patriotic Front (RPF), which invaded Rwanda in 1990. RPF leader Fred Rwigyema died early in the war and Kagame took control. By 1993, the RPF controlled significant territory in Rwanda and a ceasefire was negotiated. The assassination of Rwandan President Juvénal Habyarimana set off the genocide, in which Hutu extremists killed an estimated 500,000 to 1,000,000 Tutsi and moderate Hutu. Kagame resumed the civil war, and ended the genocide with a military victory.

Entities

total 27 not normalized 27

🟩 person 12
8 4 (0.00%) 1

country 6
2 4 (0.00%) 1

🟨 date 9
7 2 (0.00%) 0

Relations

total 1

Figure 5: Extraction of entities and relations of interest

After the document is uploaded you can open it by clicking on it, then select the text of interest and choose the related entity, and then you can add relation.

When you are done annotating, you save your annotated document on the cloud with the online account created in Tagtog website. The annotated document can later be viewed by accessing your account remotely.

3.5.1.1.2 Semi- automatic annotation in Tagtog annotation tool

For semi-automatic annotation, click on the annotations tab, activate machine learning and then, tick the option to annotate automatically using machine learning. Once this is done this, go ahead with manual annotation following the same process seen above in manual annotation using Tagtog. After saving the annotations, the user clicks on the confirm button to train your machine learning model. The model will learn from your manual annotations and next time that a document to be annotated is uploaded, it will be automatically annotated using the last trained machine learning model, and all the annotator will have to do is to make corrections on the wrong predictions.

3.5.1.1.3 Automatic annotation in Tagtog annotation tool

The tool provides already trained machine learning models. The automatic annotation is done by selecting one or more already trained models of your interest, and upload the text document to be annotated and the text document will be automatically annotated following the chosen machine learning model(s).

3.5.2 Annotea annotation tool

Annotea is an open annotation framework instantiated in different tools including Amaya, annozilla and vannotea. In this study we will only Amaya is used to test Annotea. Annotea uses w3c standard and supports annotation of RDF (Resource Descriptive Framework) documents which is based on annotations defined in a schema (Kahan & Koivunen, 2001). The annotations are stored in RDF databases and they can be accessed through HTTP (Hypertext Transfer Protocol) servers.

The annotations are defined in RDF schema and Xpointer are used to locate the annotations in the document. The supported document format is HTML or XML.

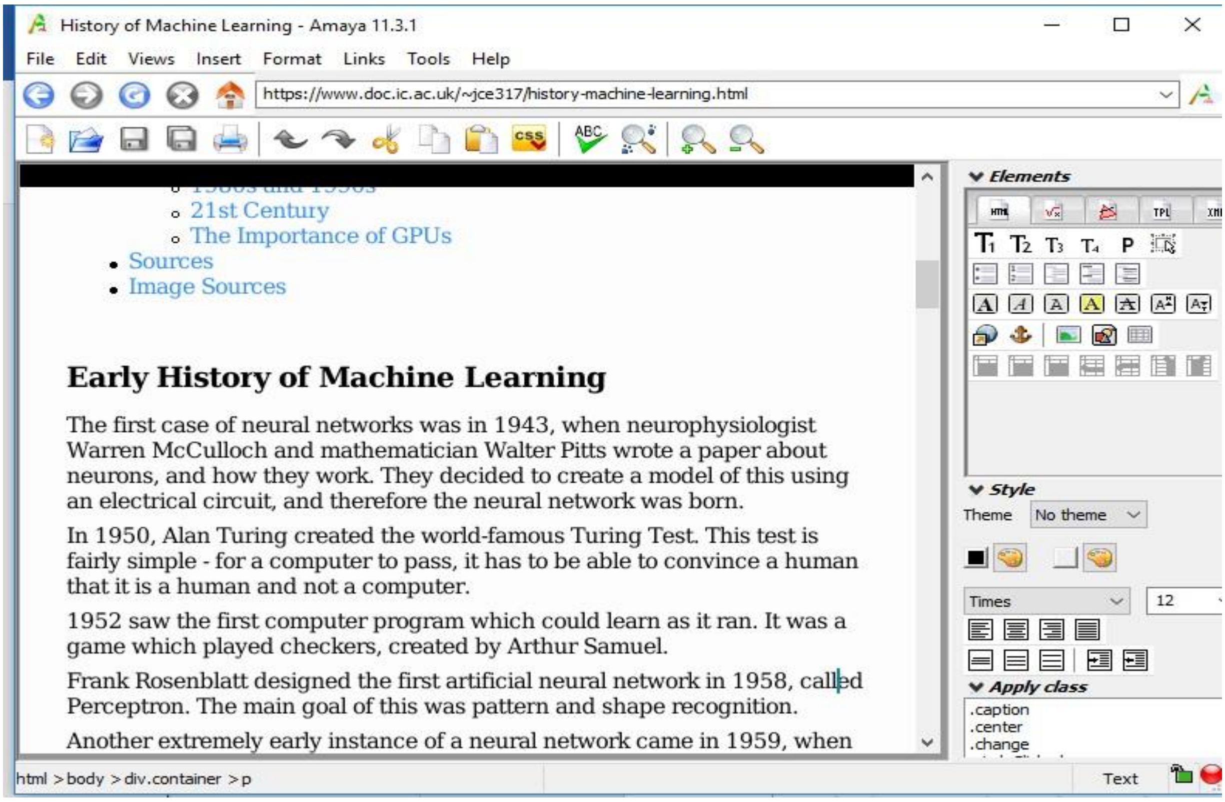
When the user uploads a document to be annotated, he/she can also upload the annotation related to it by selecting the annotation server or different annotation servers.

The tool allows collaboration among annotators which allows a user to see other users' annotations. With this ability to view other users' annotations, a user can reply or comment on annotations done by other.

3.5.2.1 Functionality of Annotea

The tool's approach concentrates on a semi-formal style of annotation, in which annotations are free text statements about documents (Uren, Hall, & Keynes, n.d.). These statements must have metadata (author, creation time etc.) and may be typed according to user-defined RDF schema of arbitrary complexity (Uren et al., n.d.).

The annotation is done manually and the first step to start annotating with Annotea, is to launch Amaya web browser. Once the Amaya editor is launched, the user or the annotator browses on the document to be annotated using the document's URL.


History of Machine Learning - Amaya 11.3.1
File Edit Views Insert Format Links Tools Help
https://www.doc.ic.ac.uk/~jce317/history-machine-learning.html
Elements
Style
Theme No theme
Times
12
Apply class
.caption
.center
.change
html > body > div.container > p
Text

- 21st Century
- The Importance of GPUs
- Sources
- Image Sources

Early History of Machine Learning

The first case of neural networks was in 1943, when neurophysiologist Warren McCulloch and mathematician Walter Pitts wrote a paper about neurons, and how they work. They decided to create a model of this using an electrical circuit, and therefore the neural network was born.

In 1950, Alan Turing created the world-famous Turing Test. This test is fairly simple - for a computer to pass, it has to be able to convince a human that it is a human and not a computer.

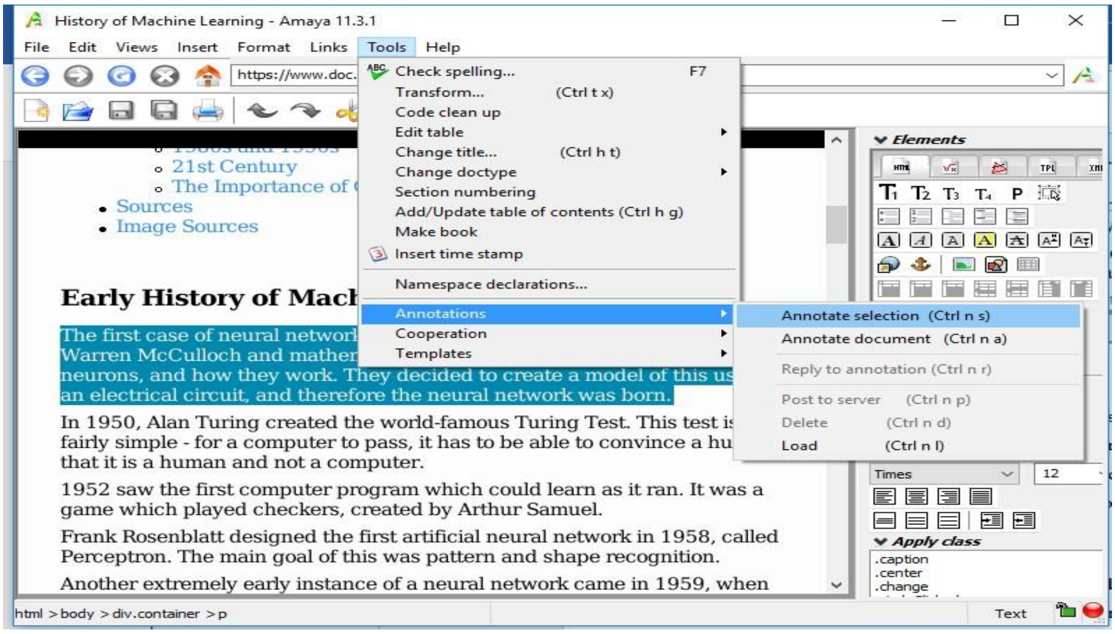
1952 saw the first computer program which could learn as it ran. It was a game which played checkers, created by Arthur Samuel.

Frank Rosenblatt designed the first artificial neural network in 1958, called Perceptron. The main goal of this was pattern and shape recognition.

Another extremely early instance of a neural network came in 1959, when

Figure 6: Amaya annotation editor interface

When the document is uploaded in the annotation editor, the user clicks on Tools on the menu bar and selects annotations. Once the user clicks on annotations, he/she is given an option to either choose to annotate document (in case the user wants to annotate the full document) or to annotate selection (in case the user wants to annotate some texts in the document).


History of Machine Learning - Amaya 11.3.1
File Edit Views Insert Format Links Tools Help
https://www.doc.
Check spelling... F7
Transform... (Ctrl t x)
Code clean up
Edit table
Change title... (Ctrl h t)
Change doctype
Section numbering
Add/Update table of contents (Ctrl h g)
Make book
Insert time stamp
Namespace declarations...
Annotations
Cooperation
Templates
Annotate selection (Ctrl n s)
Annotate document (Ctrl n a)
Reply to annotation (Ctrl n r)
Post to server (Ctrl n p)
Delete (Ctrl n d)
Load (Ctrl n l)
Elements
Times
12
Apply class
.caption
.center
.change
html > body > div.container > p
Text

- 21st Century
- The Importance of
- Sources
- Image Sources

Early History of Mach

The first case of neural networ
Warren McCulloch and mather
neurons, and how they work. They decided to create a model of this us
an electrical circuit, and therefore the neural network was born.

In 1950, Alan Turing created the world-famous Turing Test. This test is fairly simple - for a computer to pass, it has to be able to convince a hu that it is a human and not a computer.

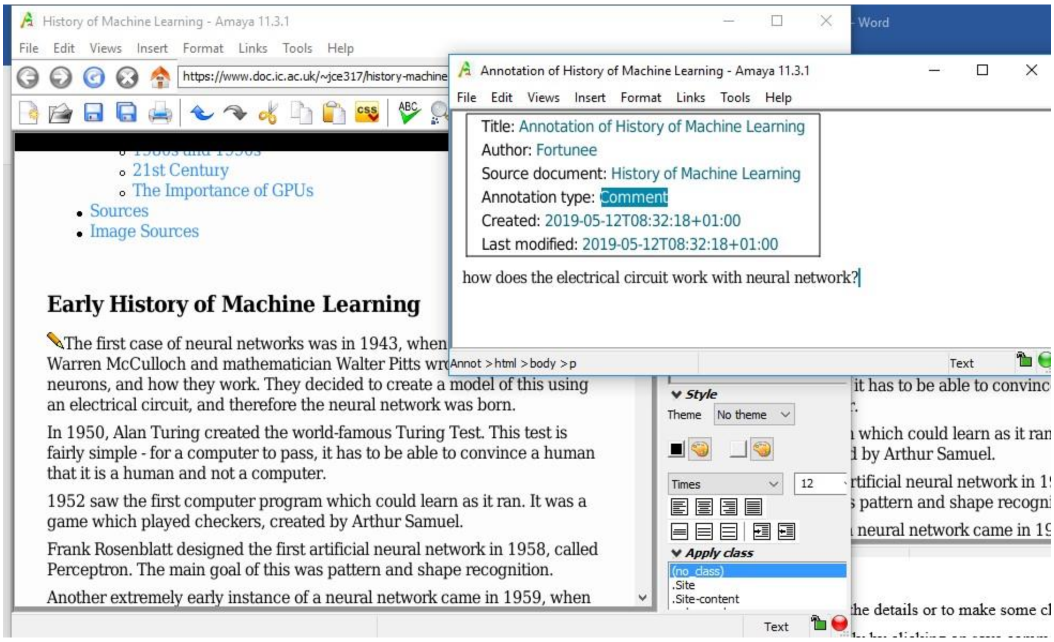
1952 saw the first computer program which could learn as it ran. It was a game which played checkers, created by Arthur Samuel.

Frank Rosenblatt designed the first artificial neural network in 1958, called Perceptron. The main goal of this was pattern and shape recognition.

Another extremely early instance of a neural network came in 1959, when

Figure 7: Selection of text span in the document to be annotated

After choosing the annotation option, the annotation window is displayed containing initial metadata like title of the annotation, author's name, document source, annotation type and date of creation and last modified date.


History of Machine Learning - Amaya 11.3.1
File Edit Views Insert Format Links Tools Help
https://www.doc.ic.ac.uk/~jce317/history-machine
Word
Annotation of History of Machine Learning - Amaya 11.3.1
File Edit Views Insert Format Links Tools Help
Annot > html > body > p
Text
Style
Theme No theme
Times
12
Apply class
(no_class)
.Site
.Site-content
Text

Title: Annotation of History of Machine Learning
Author: Fortunee
Source document: History of Machine Learning
Annotation type: Comment
Created: 2019-05-12T08:32:18+01:00
Last modified: 2019-05-12T08:32:18+01:00

how does the electrical circuit work with neural network?

- 21st Century
- The Importance of GPUs
- Sources
- Image Sources

Early History of Machine Learning

The first case of neural networks was in 1943, when Warren McCulloch and mathematician Walter Pitts wr neurons, and how they work. They decided to create a model of this using an electrical circuit, and therefore the neural network was born.

In 1950, Alan Turing created the world-famous Turing Test. This test is fairly simple - for a computer to pass, it has to be able to convince a human that it is a human and not a computer.

1952 saw the first computer program which could learn as it ran. It was a game which played checkers, created by Arthur Samuel.

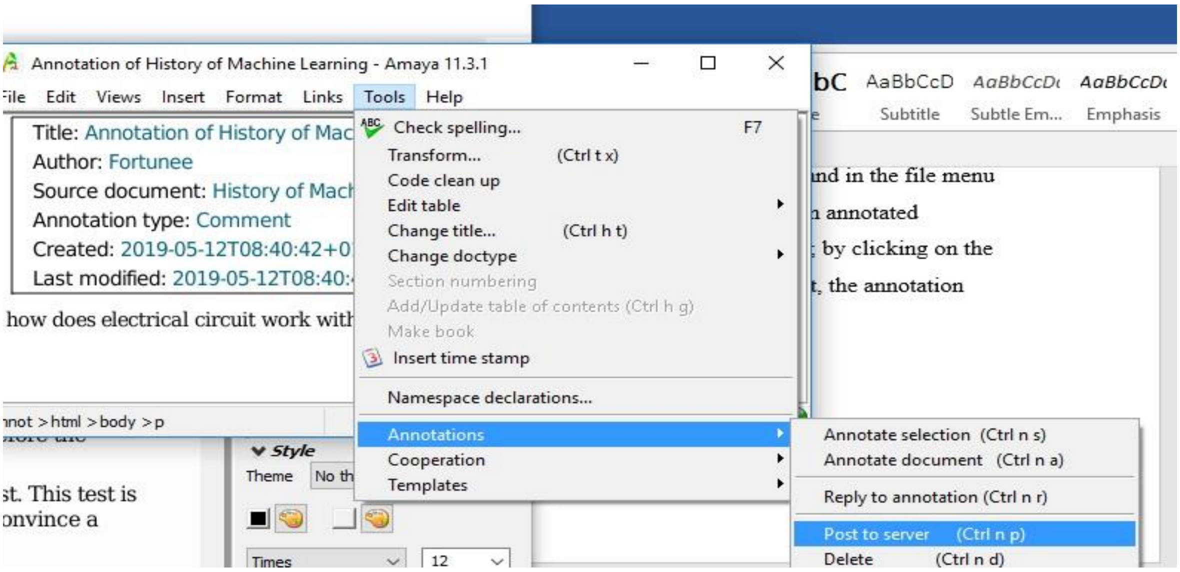
Frank Rosenblatt designed the first artificial neural network in 1958, called Perceptron. The main goal of this was pattern and shape recognition.

Another extremely early instance of a neural network came in 1959, when

Figure 8: A display annotation window

Double clicks on a metadata to view the details or to make some changes. The annotations can be saved locally by clicking on save command in the file menu or you can save it remotely by choosing post to server.

When a user opens an annotated document, the annotation within the document is identified by a pencil icon; by clicking on the pencil icon, the annotated text is selected and by clicking on the annotated text, its annotation window is displayed.

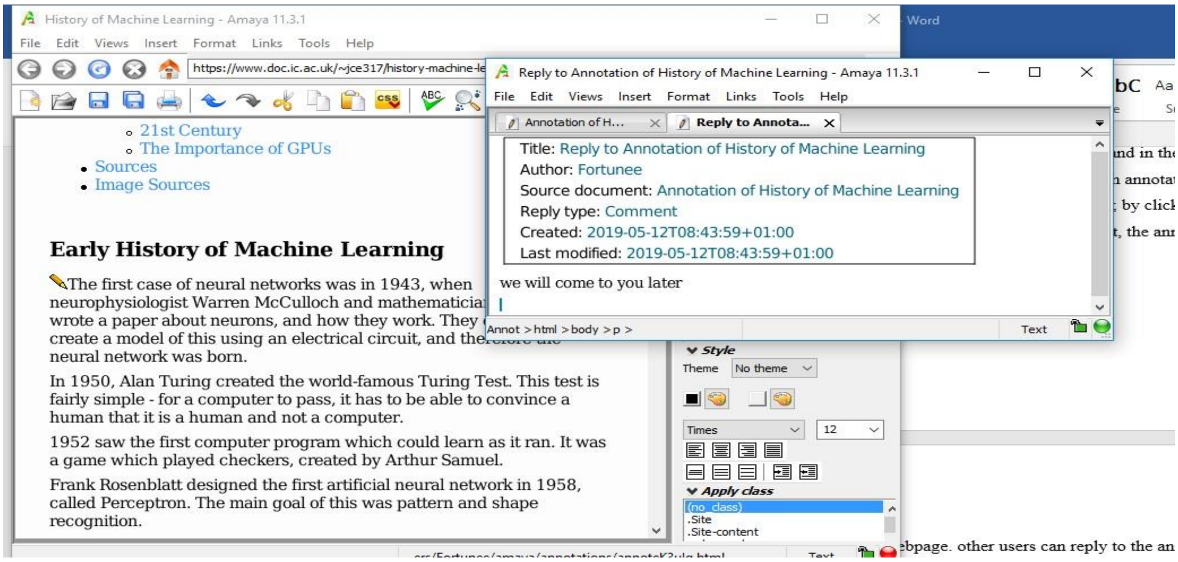


Title: Annotation of History of Mac
Author: Fortune
Source document: History of Mach
Annotation type: Comment
Created: 2019-05-12T08:40:42+0
Last modified: 2019-05-12T08:40:

how does electrical circuit work with

Annotation of History of Machine Learning - Amaya 11.3.1
File Edit Views Insert Format Links Tools Help
Check spelling... F7
Transform... (Ctrl t x)
Code clean up
Edit table
Change title... (Ctrl h t)
Change doctype
Section numbering
Add/Update table of contents (Ctrl h g)
Make book
Insert time stamp
Namespace declarations...
Annotations
Cooperation
Templates
Annotate selection (Ctrl n s)
Annotate document (Ctrl n a)
Reply to annotation (Ctrl n r)
Post to server (Ctrl n p)
Delete (Ctrl n d)
nnot > html > body > p
Style
Theme
Times
12
Subtitle
Subtle Em...
Emphasis

Figure 9: Posting the annotations to the server



Title: Reply to Annotation of History of Machine Learning
Author: Fortunee
Source document: Annotation of History of Machine Learning
Reply type: Comment
Created: 2019-05-12T08:43:59+01:00
Last modified: 2019-05-12T08:43:59+01:00

we will come to you later

- 21st Century
- The Importance of GPUs
- Sources
- Image Sources

Early History of Machine Learning

The first case of neural networks was in 1943, when neurophysiologist Warren McCulloch and mathematicia wrote a paper about neurons, and how they work. They create a model of this using an electrical circuit, and the neural network was born.

In 1950, Alan Turing created the world-famous Turing Test. This test is fairly simple - for a computer to pass, it has to be able to convince a human that it is a human and not a computer.

1952 saw the first computer program which could learn as it ran. It was a game which played checkers, created by Arthur Samuel.

Frank Rosenblatt designed the first artificial neural network in 1958, called Perceptron. The main goal of this was pattern and shape recognition.

History of Machine Learning - Amaya 11.3.1
File Edit Views Insert Format Links Tools Help
https://www.doc.ic.ac.uk/~jce317/history-machine-le
Reply to Annotation of History of Machine Learning - Amaya 11.3.1
Annotation of H...
Reply to Annota...
Annot > html > body > p >
Text
Style
Theme No theme
Times 12
Apply class
(no_class)
.Site
.Site-content
Word

Figure 10: Replying to the annotation

Annotations are seen as comments on the webpage and other users can reply to the annotations. An annotation on the current document can also be deleted using delete command on tool/annotations menu.

3.5.3 Diigo annotation tool

Diigo (Tyler Manolovitz, 2005) is a web browser extension annotation tool that annotates the browed PDFs.

It is mostly used by researchers and teachers for knowledge elicitation purpose (Tyler Manolovitz, 2005).

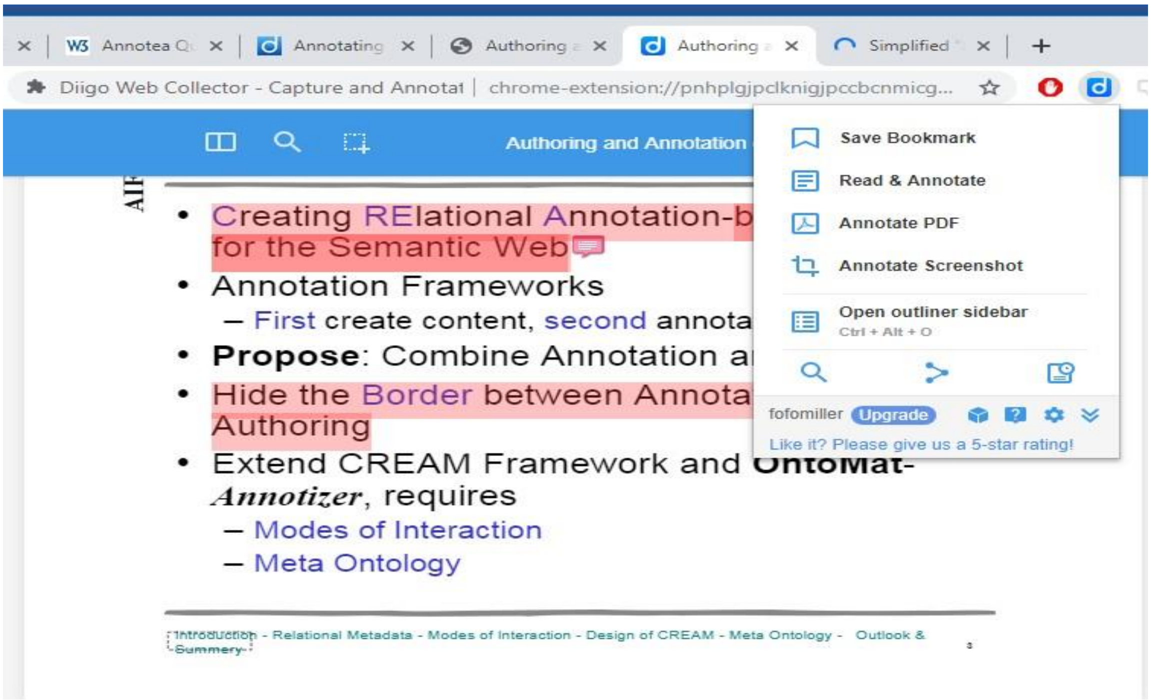
3.5.3.1 Functionality of Diigo annotation tool

This tool is mostly known as a free social bookmarking, research, and knowledge sharing tool that was created for the purpose of mimicking the ease of taking notes while providing a network for sharing and discovering information(Tyler Manolovitz, 2005) page. The user can also bookmark and tag the web pages (Tyler Manolovitz, 2005).

3.5.3.2 Diigo features

- **Highlighting:** Diigo permits the user to highlight a text on a web page while reading and the annotation is automatically saved.
- **Adding notes:** Diigo allows the user to comment on a text on the web page by adding sticky notes. This type of annotation is seen in the text as unobtrusive balloon and if you click on it, all the comments are displayed.
- **Bookmarking:** Diigo allows the user to bookmark and save the web page. As you bookmark the page, Diigo allows the ability to add tags to help keep

everything organized and provide easy access and retrieval, and sharing your annotations with a selected group of users (Tyler Manolovitz, 2005).



- Creating RElational Annotation-b
  for the Semantic Web
- Annotation Frameworks
  - First create content, second annota
- **Propose**: Combine Annotation a
- Hide the Border between Annota
  Authoring
- Extend CREAM Framework and **OntoMat-*Annotizer***, requires
  - Modes of Interaction
  - Meta Ontology

Introduction - Relational Metadata - Modes of Interaction - Design of CREAM - Meta Ontology - Outlook & Summery

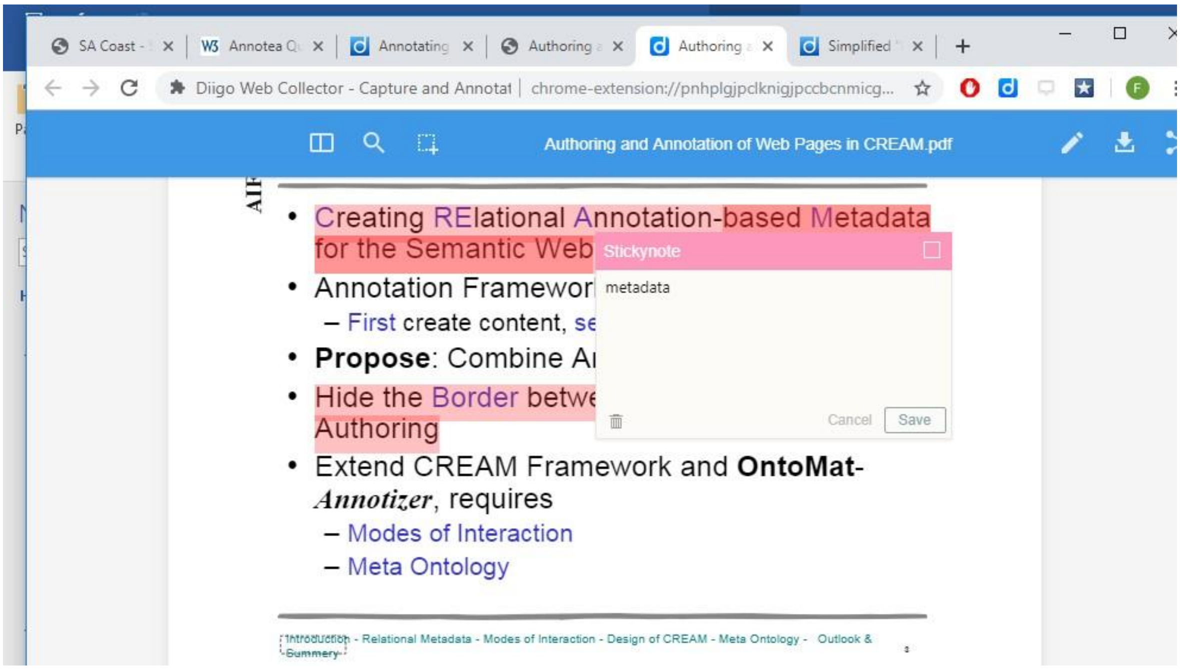
Figure 11: Annotations provided by Diigo

Bookmarking a web page gives a user an option to share his annotations with other users within the same group.

The human annotator annotates a web page and Diigo automatically save the annotations; hence, the annotation done is semi-automatic.

There are two ways to annotate with Diigo. The first way is to upload the pdf document to the Diigo interface in the user's library then open the document and select text of interest to be annotated ("Tools," n.d.).

The other way is to add Diigo annotation tool to the supported browser as a web browser extension. Once a PDF document is loaded, the user can annotate it by clicking on annotate then select the text of interest to highlight or add sticky note to.


SA Coast -
Annotea Q
Annotating
Authoring
Authoring
Simplified
Diigo Web Collector - Capture and Annotat
chrome-extension://pnhplgjpclknigjpccbcnmicg...
Authoring and Annotation of Web Pages in CREAM.pdf
Stickynote
Cancel
Save

- Creating RElational Annotation-based Metadata for the Semantic Web
- Annotation Framewor
  - First create content, se
- **Propose**: Combine A
- Hide the Border betwe Authoring
- Extend CREAM Framework and **OntoMat-*Annotizer***, requires
  - Modes of Interaction
  - Meta Ontology

metadata

Introduction - Relational Metadata - Modes of Interaction - Design of CREAM - Meta Ontology - Outlook & Summery

Figure 12: a selected text to be annotated

3.5.4 GATE annotation tool

GATE (Kenter & Maynard, 2005) is a platform that can support all Natural Language Processing(NLP) software(Kenter & Maynard, 2005). One of the most important services provided by GATE is annotation.

With GATE annotation tool, the annotation types can be defined in annotation schema or otherwise. If the annotation types are not defined in the annotation schema there is need to define the annotation properties of the annotation type every time that the annotation is created.

By default, some real-world Entities (person, location, organization, etc.) are defined in the annotation schema.

The prevalent concepts that are used in GATE are resources. The two most important types of resources in Gate are:

Language resources: contain the documents, documents collections (corpus) for annotation.

Processing resources: algorithms and programs or applications that are used to modify or automatically create annotations over a document or a corpus.

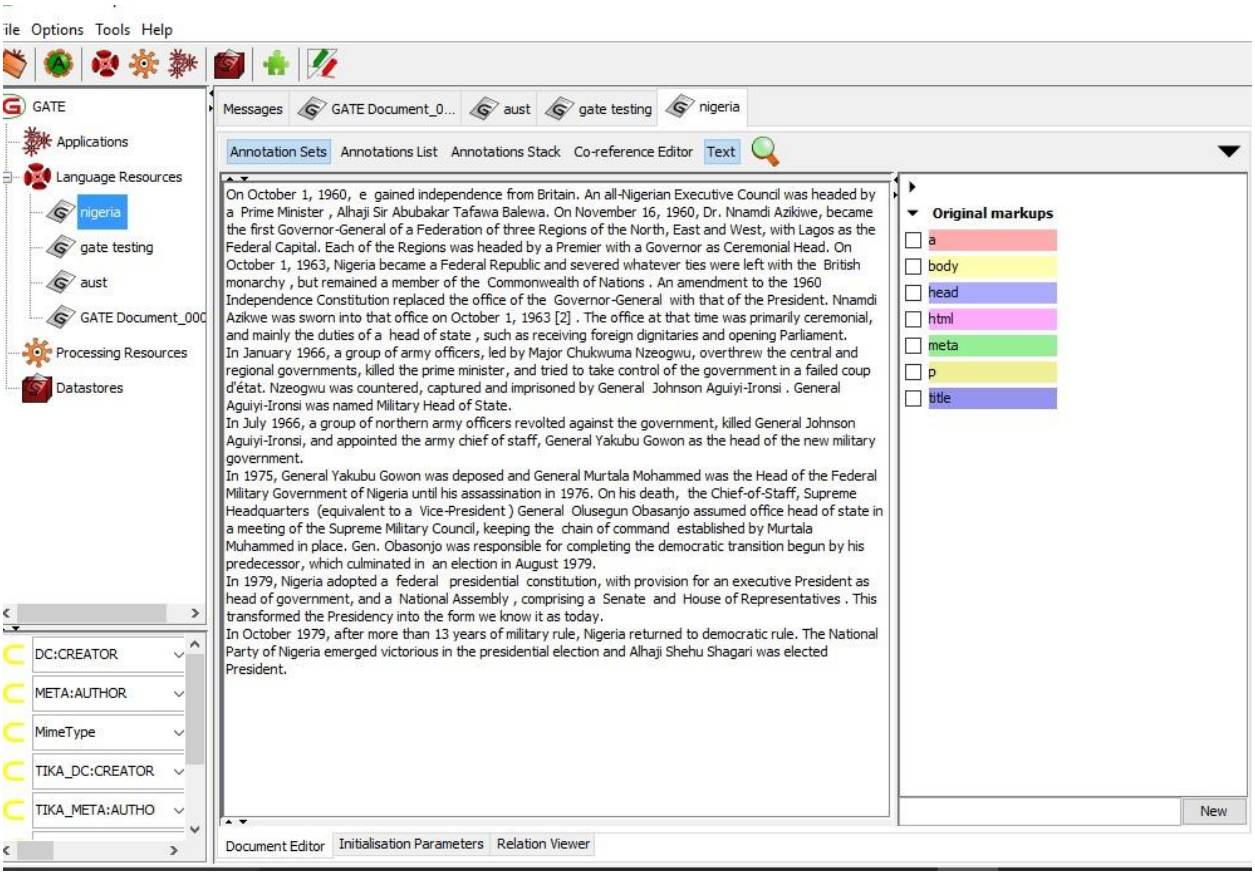
3.5.4.1 Functionality of GATE annotation tool

With Gate annotation tool, the annotation done can be manual, semi-automatic or automatic. The manual annotation is done by human whereas the automatic annotation is achieved by using gate applications like gazetteers, ANNIE, etc.

The annotations are stored in data store and they are viewed by clicking on annotation sets. Gate annotation tool supports ontologies that can be used for automatic generation of annotation schema (Kenter & Maynard, 2005).

3.5.4.1.1 Manual annotation using GATE

The document to be annotated is uploaded in the language resources and the uploaded documents should be in the following formats: HTML, XML, PDF, Word, CoNLL, CSV, JSON, plain text, etc., and the exported format are XML or HTML.



On October 1, 1960, e gained independence from Britain. An all-Nigerian Executive Council was headed by a Prime Minister , Alhaji Sir Abubakar Tafawa Balewa. On November 16, 1960, Dr. Nnamdi Azikiwe, became the first Governor-General of a Federation of three Regions of the North, East and West, with Lagos as the Federal Capital. Each of the Regions was headed by a Premier with a Governor as Ceremonial Head. On October 1, 1963, Nigeria became a Federal Republic and severed whatever ties were left with the British monarchy , but remained a member of the Commonwealth of Nations . An amendment to the 1960 Independence Constitution replaced the office of the Governor-General with that of the President. Nnamdi Azikwe was sworn into that office on October 1, 1963 [2] . The office at that time was primarily ceremonial, and mainly the duties of a head of state , such as receiving foreign dignitaries and opening Parliament.

In January 1966, a group of army officers, led by Major Chukwuma Nzeogwu, overthrew the central and regional governments, killed the prime minister, and tried to take control of the government in a failed coup d'état. Nzeogwu was countered, captured and imprisoned by General Johnson Aguiyi-Ironsi . General Aguiyi-Ironsi was named Military Head of State.

In July 1966, a group of northern army officers revolted against the government, killed General Johnson Aguiyi-Ironsi, and appointed the army chief of staff, General Yakubu Gowon as the head of the new military government.

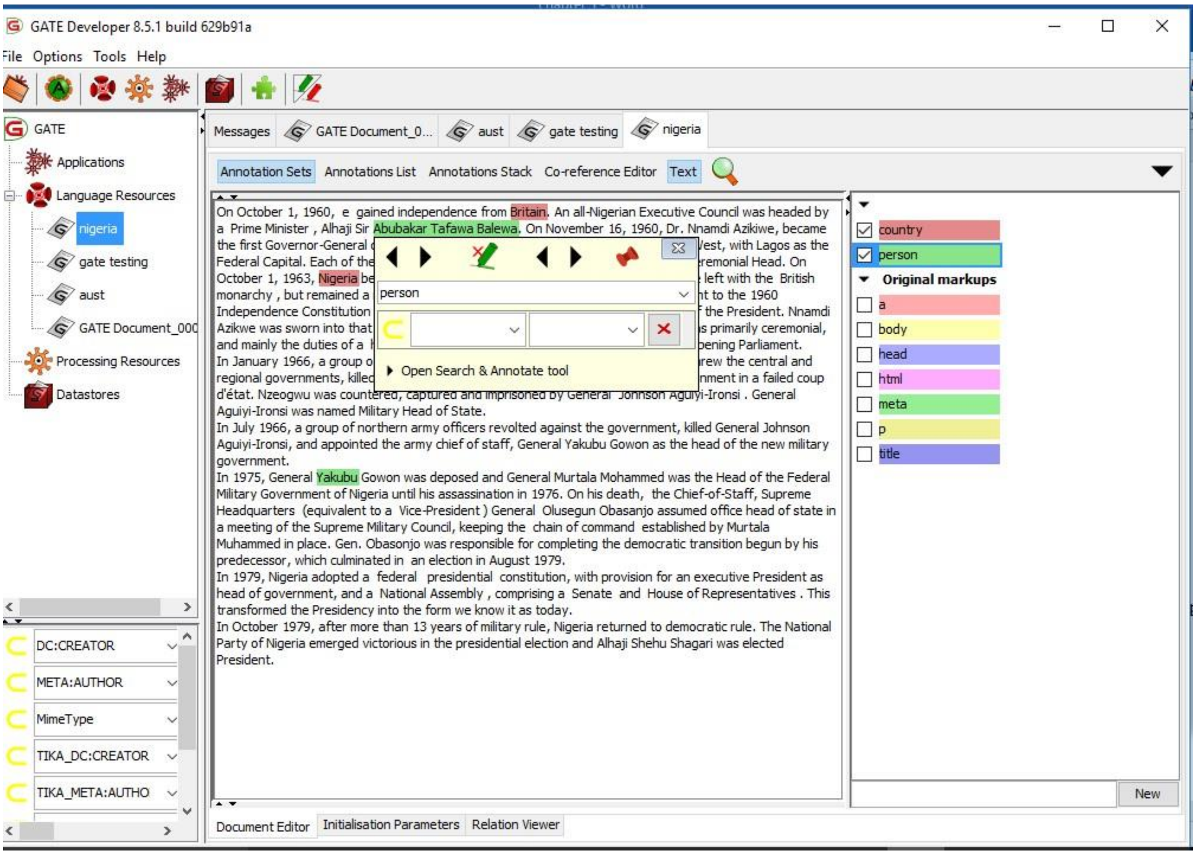
In 1975, General Yakubu Gowon was deposed and General Murtala Mohammed was the Head of the Federal Military Government of Nigeria until his assassination in 1976. On his death, the Chief-of-Staff, Supreme Headquarters (equivalent to a Vice-President) General Olusegun Obasanjo assumed office head of state in a meeting of the Supreme Military Council, keeping the chain of command established by Murtala Muhammed in place. Gen. Obasonjo was responsible for completing the democratic transition begun by his predecessor, which culminated in an election in August 1979.

In 1979, Nigeria adopted a federal presidential constitution, with provision for an executive President as head of government, and a National Assembly , comprising a Senate and House of Representatives . This transformed the Presidency into the form we know it as today.

In October 1979, after more than 13 years of military rule, Nigeria returned to democratic rule. The National Party of Nigeria emerged victorious in the presidential election and Alhaji Shehu Shagari was elected President.

Figure 13: Original GATE mark-ups displayed

For manual annotation, the user the text to be annotated and drag the mouse on the related annotation type if the annotation types are already defined in the annotation schema. If there is no defined schema, the user uses the keyboard shortcut(ctrl E) to define the annotation type and a popup window appears in which the user defines the annotation types (*CEUR workshop proceedings*, n.d.). The created annotations are automatically stored in annotation sets.



On October 1, 1960, e gained independence from Britain. An all-Nigerian Executive Council was headed by
a Prime Minister , Alhaji Sir Abubakar Tafawa Balewa. On November 16, 1960, Dr. Nnamdi Azikiwe, became
the first Governor-General … West, with Lagos as the
Federal Capital. Each of the … remonial Head. On
October 1, 1963, Nigeria be … left with the British
monarchy , but remained a … nt to the 1960
Independence Constitution … the President. Nnamdi
Azikwe was sworn into that … s primarily ceremonial,
and mainly the duties of a … pening Parliament.
In January 1966, a group o … rew the central and
regional governments, killed … nment in a failed coup
d'état. Nzeogwu was countered, captured and imprisoned by General Johnson Aguiyi-Ironsi . General
Aguiyi-Ironsi was named Military Head of State.
In July 1966, a group of northern army officers revolted against the government, killed General Johnson Aguiyi-Ironsi, and appointed the army chief of staff, General Yakubu Gowon as the head of the new military government.
In 1975, General Yakubu Gowon was deposed and General Murtala Mohammed was the Head of the Federal Military Government of Nigeria until his assassination in 1976. On his death, the Chief-of-Staff, Supreme Headquarters (equivalent to a Vice-President) General Olusegun Obasanjo assumed office head of state in a meeting of the Supreme Military Council, keeping the chain of command established by Murtala Muhammed in place. Gen. Obasonjo was responsible for completing the democratic transition begun by his predecessor, which culminated in an election in August 1979.
In 1979, Nigeria adopted a federal presidential constitution, with provision for an executive President as head of government, and a National Assembly , comprising a Senate and House of Representatives . This transformed the Presidency into the form we know it as today.
In October 1979, after more than 13 years of military rule, Nigeria returned to democratic rule. The National Party of Nigeria emerged victorious in the presidential election and Alhaji Shehu Shagari was elected President.

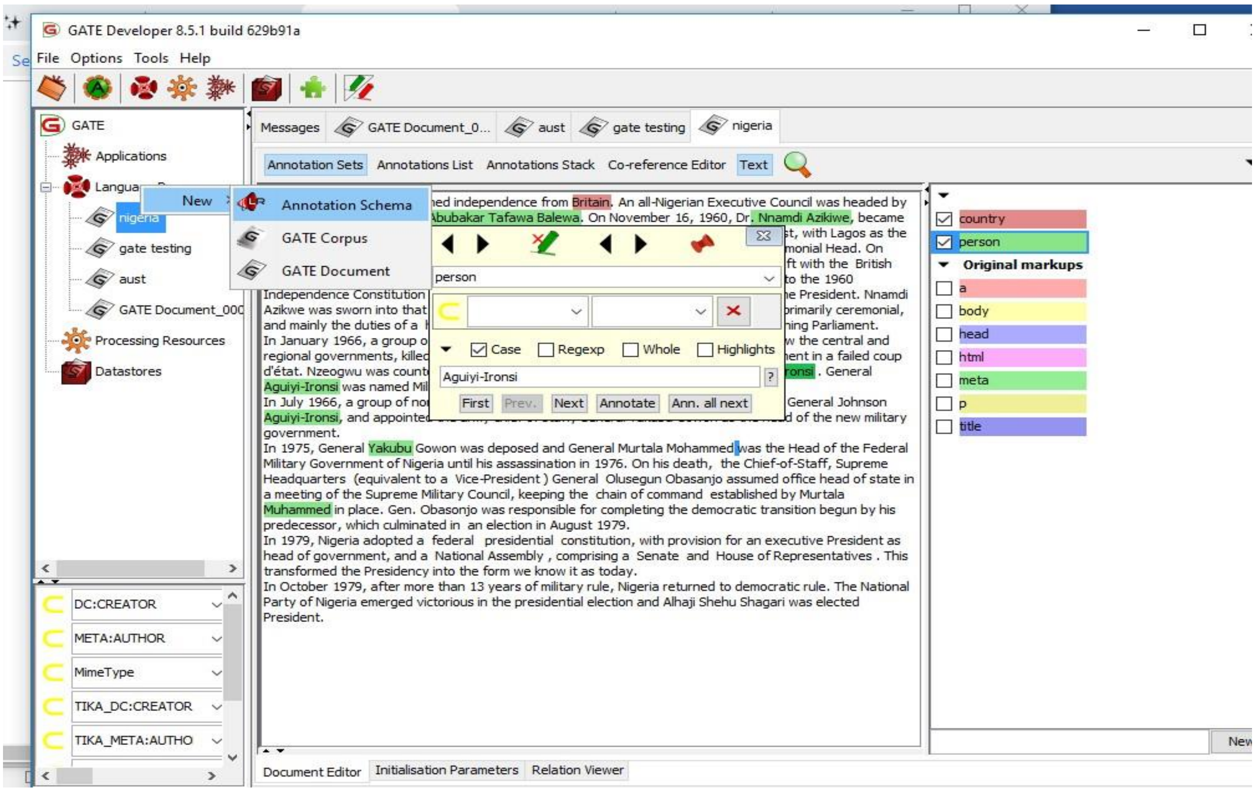
Figure 14: Deletion of annotation in GATE

3.5.4.1.2 Semi-automatic with GATE

For semi-automatic annotation, the user run a process resource over the corpus for automatic annotation, then the user can make corrections or add annotations manually.

3.5.4.1.3 Automatic annotation with GATE

For automatic annotation, choose or create a model to use to annotate, click on OpenSearch and annotate in the annotation editor window. Use the first button to choose the first expression to annotate, use annotate button if correct and NEXT button otherwise.



ned independence from Britain. An all-Nigerian Executive Council was headed by
Abubakar Tafawa Balewa. On November 16, 1960, Dr. Nnamdi Azikiwe, became
st, with Lagos as the
monial Head. On
ft with the British
to the 1960
Independence Constitution … e President. Nnamdi
Azikwe was sworn into that … primarily ceremonial,
and mainly the duties of a … hing Parliament.
In January 1966, a group o … w the central and
regional governments, killed … ent in a failed coup
d'état. Nzeogwu was count … ronsi . General
Aguiyi-Ironsi was named Mil
In July 1966, a group of no … General Johnson
Aguiyi-Ironsi, and appointed … d of the new military
government.

In 1975, General Yakubu Gowon was deposed and General Murtala Mohammed was the Head of the Federal Military Government of Nigeria until his assassination in 1976. On his death, the Chief-of-Staff, Supreme Headquarters (equivalent to a Vice-President) General Olusegun Obasanjo assumed office head of state in a meeting of the Supreme Military Council, keeping the chain of command established by Murtala Muhammed in place. Gen. Obasonjo was responsible for completing the democratic transition begun by his predecessor, which culminated in an election in August 1979.

In 1979, Nigeria adopted a federal presidential constitution, with provision for an executive President as head of government, and a National Assembly , comprising a Senate and House of Representatives . This transformed the Presidency into the form we know it as today.

In October 1979, after more than 13 years of military rule, Nigeria returned to democratic rule. The National Party of Nigeria emerged victorious in the presidential election and Alhaji Shehu Shagari was elected President.

Figure 15: Creation of a new schema

Gate also uses a created pipeline application to automatically annotate a single document and a pipeline corpus to automatically annotate a corpus.

For the automatic annotation in gate, the application that will automatically annotate the corpus is uploaded by clicking on processing resources; then click on the application, to automatically annotate the whole corpus.

3.5.5 Doccano annotation tool

Doccano implemented by Hironsan**,** is an open source annotation tool for text.

It is an annotation tool implemented for human and presents features for sequence labelling, text classification and sequence to sequence (“doccano - Document Annotation Tool,” n.d.).

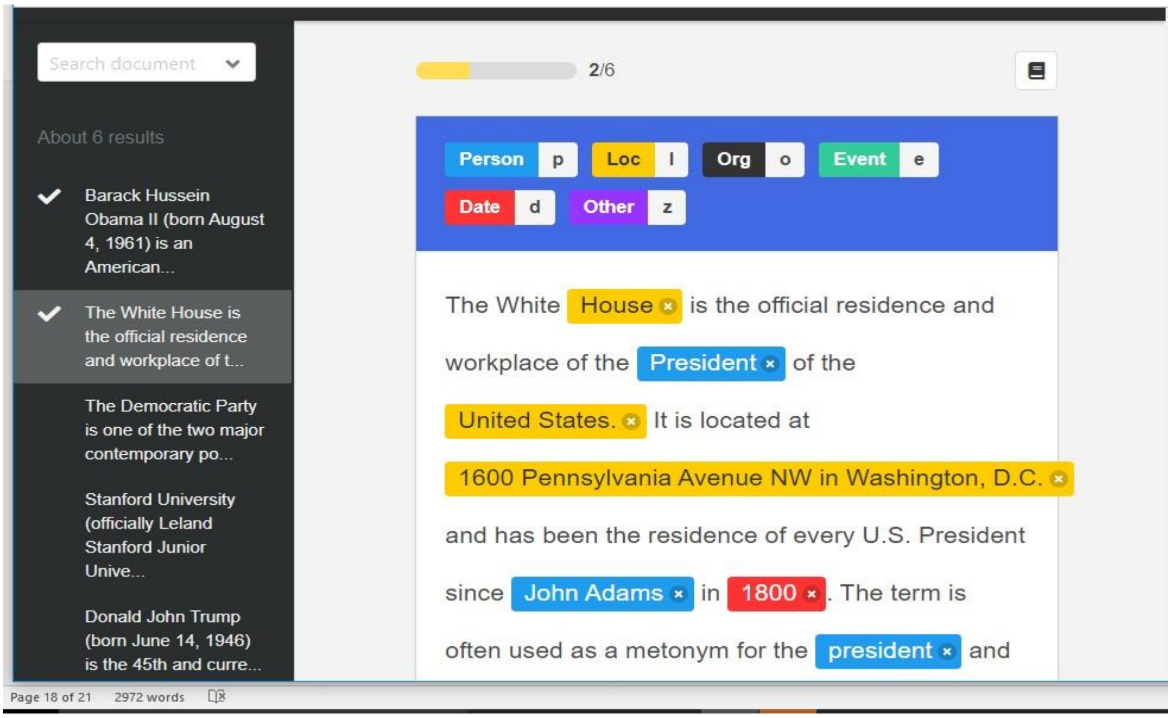
3.5.5.1 Doccano features

- Collaborative annotation
- Emoji support
- Future automatic label
- Multi-language support

3.5.5.2 Functionalities of Doccano

There are three main types of annotation that can be performed using Doccano annotation tool:

Named entity recognition: a span text is selected and annotated accordingly to the corresponding defined entities.



✓ Barack Hussein Obama II (born August 4, 1961) is an American...

✓ The White House is the official residence and workplace of t...

The Democratic Party is one of the two major contemporary po...

Stanford University (officially Leland Stanford Junior Unive...

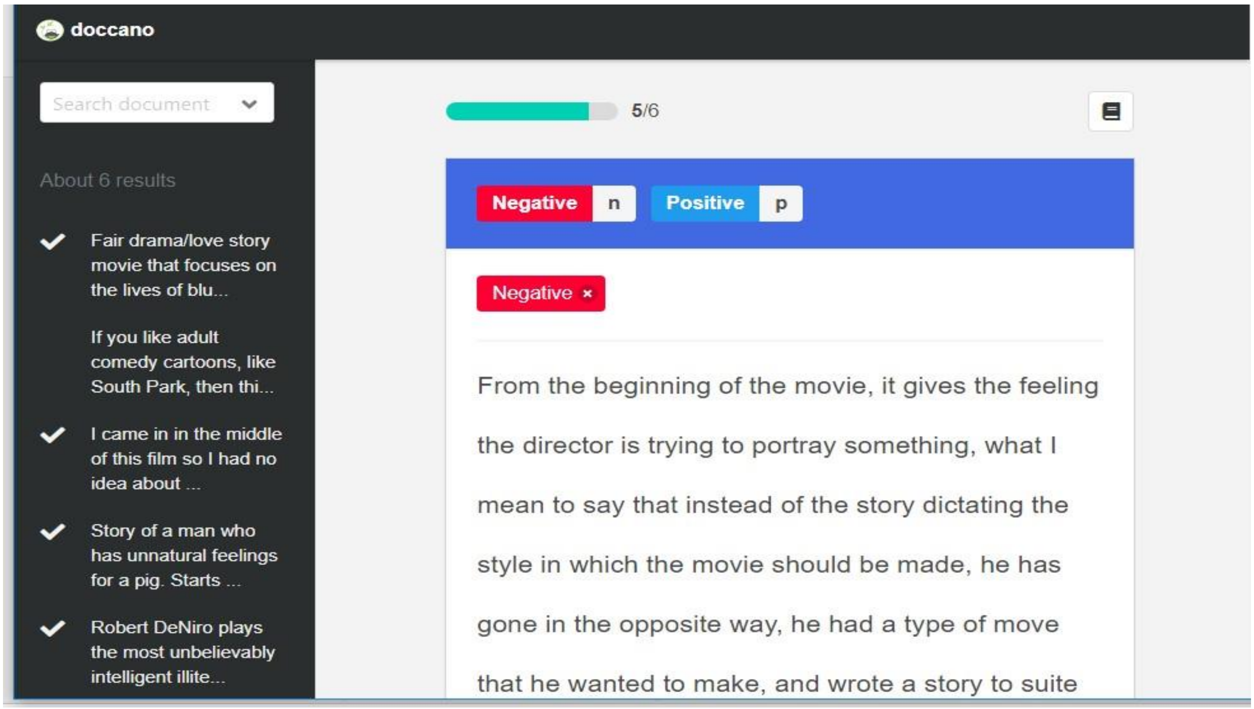
Donald John Trump (born June 14, 1946) is the 45th and curre...

The White House is the official residence and workplace of the President of the United States. It is located at 1600 Pennsylvania Avenue NW in Washington, D.C. and has been the residence of every U.S. President since John Adams in 1800. The term is often used as a metonym for the president and

Figure 16: Entity extraction in Doccano

Sentiment analysis: is a task of classifying texts and topics into different categories.

Since a text may fall into more than one category, the annotation can be multi-labels ("Open source text annotation tool for machine learning practitioner," n.d.).

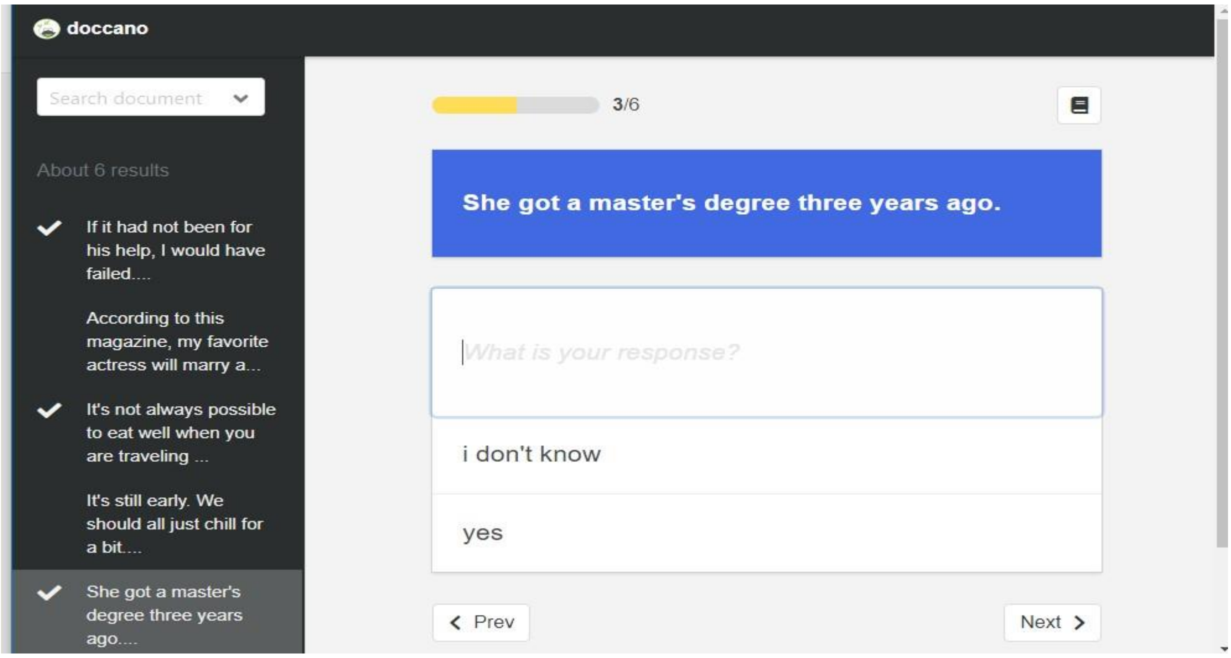


- ✓ Fair drama/love story movie that focuses on the lives of blu...
- If you like adult comedy cartoons, like South Park, then thi...
- ✓ I came in in the middle of this film so I had no idea about ...
- ✓ Story of a man who has unnatural feelings for a pig. Starts ...
- ✓ Robert DeNiro plays the most unbelievably intelligent illite...

From the beginning of the movie, it gives the feeling the director is trying to portray something, what I mean to say that instead of the story dictating the style in which the movie should be made, he has gone in the opposite way, he had a type of move that he wanted to make, and wrote a story to suite

Figure 17: Sentiment analysis with Doccano

Machine translation is one of the sequence to sequence tasks that enable multiple responses, in case more than one response can be provided.



✓ If it had not been for his help, I would have failed....

According to this magazine, my favorite actress will marry a...

✓ It's not always possible to eat well when you are traveling ...

It's still early. We should all just chill for a bit....

✓ She got a master's degree three years ago....

She got a master's degree three years ago.

What is your response?

i don't know

yes

Figure 18: Machine translation using Doccano

3.5.5.3 Functionalities of Doccano

The user creates a new project and select the project type depending on his/her needs. When the project has been created the “import data” button is displayed so that the user can upload data. Only two types of file formats can be uploaded:

- **CSV (Comma Separated Values) file:** file must contain a header with a text column or be one-column csv file.
- **JSON (JavaScript Object Notation) file**: each line contains a JSON object with a text key.

3.5.5.3.1 Manual annotation using Doccano

After selecting a JSON file from to be annotated, upload dataset. A dataset page is displayed containing all the uploaded documents in one project.

The user can start defining labels by clicking on **labels** button. After defining your labels and entities, you can start your annotation by clicking on **annotate data** button. You are now able to start annotating the uploaded document.

3.5.5.3.2 Semi-automatic annotation using Doccano

A machine learning model is uploaded to automatically annotate the uploaded document and the user can make some corrections on the wrong predictions.

3.5.5.3.3 Automatic annotation using Doccano

The uploaded document is automatically annotated by the uploaded already trained machine learning model.

3.5.6 Prodigy annotation tool

Prodigy implemented by (Ines Montani and Matthew Honnibal, August 4, 2017), is an active learning annotation tool that allows data scientists to annotate by themselves. It combines a machine learning model together with the answers from the user to provide high accuracy annotations.

With prodigy, the model is kept in the loop for it to participate in the training process and learn from the answers from the user ("Prodigy · An annotation tool for AI, Machine Learning &amp; NLP," n.d.). It uses its knowledge to figure out what to ask the user; the answers provide by the user are used to train and update the machine learning model in real time.

To configure prodigy system, you need to configure a python function to return the components as dictionaries.

The tool comes with built-in recipes to train and evaluate text classification, named entity recognition, image classification and word vector models.

The annotations are stored in an SQL database by default to minimize the system requirement.

The tool presents a web application that provides an interface that can help to annotate text, extract entities, text classification and image classification directly from the web browser. This web application can run on all operating platforms including mobile device.

The imported data can be in any text format or image. It gives the user to upload his/her own machine learning model and use it for annotations.

3.5.6.1 Functionalities of prodigy annotation tool

The types of annotations that can be done in prodigy are: entity recognition, semantic relations, entity labelling and image labelling and classification. We are used to annotation tools that present span text that contains entity to the user and ask him/her to highlight it, but prodigy presents different labels in the dropdown that user can choose from ("Prodigy · An annotation tool for AI, Machine Learning &amp; NLP," n.d.).

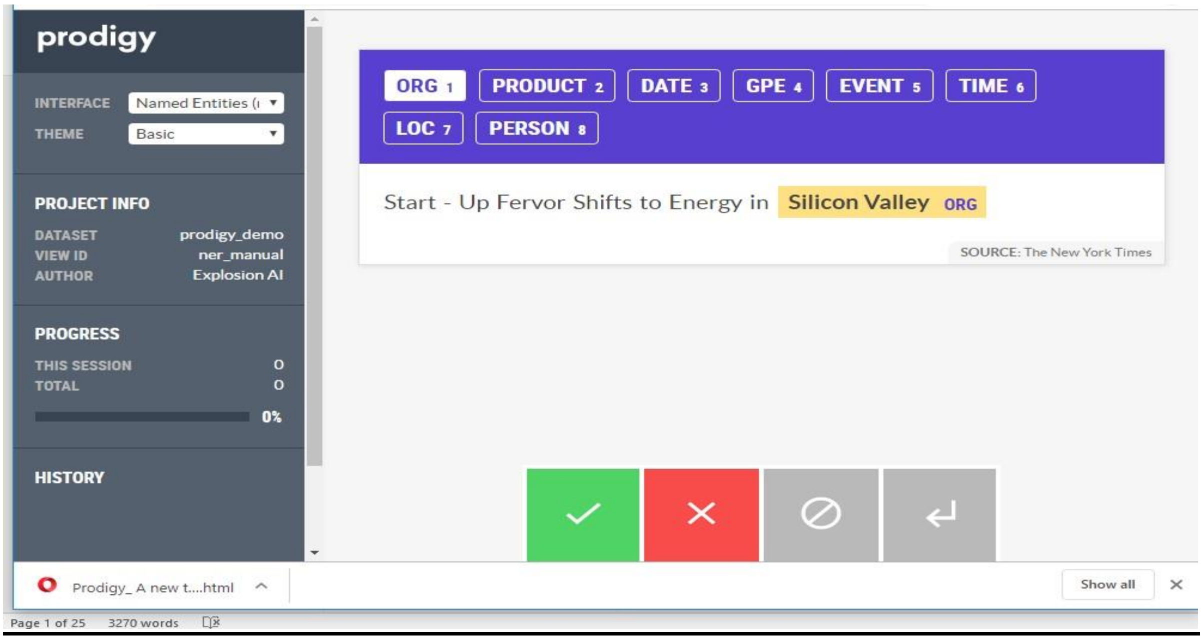
Prodigy supports and annotate any text format or image and allows the user to integrate his/her own machine learning model.

Prodigy can be run on different platforms even on mobile devices. With prodigy annotation tool, the user is focused on only one annotation task at a time which gives high accuracy annotations.

The annotation process starts by loading the datasets. You can either load your own datasets or make use of provided sample datasets. The following are the annotation types that can be done in prodigy.

3.5.6.1.1 Entity recognition with Prodigy

The python function (recipe) starts the server and use the spacy to model to detect entities, the annotator will be able to maintain the correct detected entities and make correction on the wrong predicted entities. You can either accept, reject or ignore the predictions. At the same moment the user is making corrections, the model is being updated as well.



INTERFACE Named Entities (
THEME Basic

PROJECT INFO
DATASET prodigy_demo
VIEW ID ner_manual
AUTHOR Explosion AI

PROGRESS
THIS SESSION 0
TOTAL 0
0%

HISTORY

Start - Up Fervor Shifts to Energy in Silicon Valley ORG

SOURCE: The New York Times

Figure 19: Entity recognition with Prodigy

3.5.6.1.2 Text classification

This the annotation task of classifying of text into different categories. Whether the user is doing image detection, entity detection information extraction, semantic role labeling or sentiment analysis, Prodigy provides easy, flexible and powerful annotation facilities.

Active learning feature keeps the annotations efficient even if the classes are heavily imbalanced (“Text Classification · Prodigy · An annotation tool for AI, Machine Learning &amp; NLP,” n.d.).

3.5.6.1.3 Computer vision

Labelling images, detecting objects from the image and image classifications. Label images in-house for image annotation tasks such as object detection, image segmentation and image classification.

Use Prodigy's fully scriptable back-end to build powerful active learning workflows by putting your model in the loop (“Computer Vision · Prodigy · An annotation tool for AI, Machine Learning &amp; NLP,” n.d.).

Chapter Four

Comparative analysis

4.1 Features used for comparison

The study of the above selected annotation tools led to a comparative analysis between the studied tools. Different features and properties of each of the studied tools have been compared in order to evaluate an annotation tool's performance measure.

After studying and testing the chosen tools, the figured-out features and properties that are used to compare the tools are:

1. **Availability and cost:** For the user to choose an annotation tool to use among all the available annotation tools is not an easy decision to make. Availability and cost of an annotation tool should be taken into consideration. Some of the tools are open source whereas others are not. Some tools are downloadable for free and others require payment for the license before use. Some user will like to use the tool on cloud, others will like to use the tool on-premises. For example, a user that needs to annotate webpages will always look for a webbased annotation tool, whereas a user who is seeking for high privacy for his/her annotations will always prefer to use the tool locally on his/her private infrastructure. When it comes to cost the web-based annotation tools are always cheap, but when it comes to security and privacy of data, it is always better to have your annotation tool installed on your infrastructure whether locally or on cloud.

2. **Availability for testing:** It is always better if possible, to test a tool before deciding on using it for your project. Some tools provide an online trial version or live demo and videos.

   This is very important for researchers and users when testing tools. Though some features are always missing in the trial versions, but it gives the user a view on what and how an annotation tool will serve.

3. **Operating platform and requirements:** One should consider the operating platform required by the annotation tool before choosing the annotation tool. Some of the available annotation tools are web-based annotation tool, and don't restrict on any operating platform, all that is required is that the user be connected to the internet, some of these web-based tools are peculiar to some browsers, and some work with any web browser. Unlike web-based annotation tools, the tools that are not web-based need custom installation on the user's local infrastructure or on cloud infrastructure (AMAZONE, GOOGLE, etc.).

4. **User friendliness:** An annotator or the person who does annotation is not always expected to be a programmer or a data engineer, annotation tools are used for different purposes (teaching, knowledge sharing, research, etc.). Some annotation tools present a complex interface to operate and hard to configure. For example, some tools require you to configure a server; some tools like prodigy require the user to have some python knowledge before he/she can configure it.

5. **Target data:** To my knowledge there are only three categories of target data. The target data may be in form of text, image and videos. The choice of an annotation tool should base on the target data.

6. **Machine learning support:** Machine learning models are mostly used by some annotation tools to automatically make predictions of annotations. This reduces the time that can be used to annotate and provides more annotation accuracy as the model predicts and the user makes corrections on the wrong predictions.
7. **Import format:** For text annotation tools, the user should consider the supported format of documents that can be imported by an annotation tool before choosing a tool and before uploading a document. Some annotation tools can annotate any text document, whereas some are selective.
8. **Export data format:** Exported data format differs from one tool to another.
9. **Annotation techniques:** Depending on the annotation's accuracy and quality and time constrained, someone may choose an annotation tool based on the annotation techniques that the tool provides. Some of the annotation tools provides all the three annotation techniques (manual, automatic and semiautomatic). In this, the choice is left for the user to decide which annotation technique to use for his/her annotations. Some annotation tools only provide manual annotation technique or manual and automatic techniques.
10. **Annotation types:** Depending on the annotator purpose, the annotation tool is chosen considering the annotation types that can be performed by the tool. Most of the annotation tools support annotation types like entity recognition, relations creation, entity labelling, document labelling and few of them present the normalization feature.
11. **Annotations storages:** Annotations storage differ from one annotation tool to another. For some annotation tools the annotations are saved automatically and for some others the user is in charge to save annotations.

When it comes to storage, the annotations can either be saved on local databases or server databases.

12. **Share annotations and team collaboration**: Annotations are like comments or interpretation of the document. Some annotation tools allow users in the same group to share annotations in order to share knowledge and to collaborate as a team.
13. **Searching option:** Most of the annotation tools present a searching option for easy access of documents and annotations and allow a group of users to work on the same projects.

4.2 Comparative table of annotation Tools

Tools Features	Tagtog	Annotea	Diigo	Gate	Doccano	Prodigy
Availability	+ On cloud +premises	Web-based	Browser extension	+ web APP +custom installation	+on cloud +clone it or download it	+ web App +custom installation
Operating platform and requirement	Docker version >= 18.03 + Windows + Linux + MacOS RAM 16-32GB DISK 50GB	Serves: +W3C +Annotea server + Zannot +Dart +PyNotea +Bakunin +Annotea server **Clients:** +Amaya +Annozilla +Yannotea	**Web browser:** +Chrome +Firefox +Internet explorer	Gate software installed JDK +Windows +Linux +MacOS	+Python3.6 +Django2.7 +Node.js8.0 +Google Chrome deployed to: +Azure +Heroku +AWS	**Platform:** + macOS, + OSX, + Linux, + Windows +Smart devices python version: +3.5, +3.6, +3.7 Architecture**:** x86-64 Spacy: v2.1.
User friendly interface and easy to learn	Yes	Yes	Yes	Yes	complex	Complex
Target data	Text	Text	Text	Text	Text	Text

Support machine learning	Yes	No	No	No	Yes	Yes
Import format	+PDF +TXT +HTML +source code f + markdown, + CSV	+HTML +XML	+PDF	+HTML +XML +PDF +CoNLL +CSV +JSON +Plain text	+Plain text +CSV +JSON +CoNLL	+Any text format
Export format	+HTML + XML +Txt	XML	PDF	±XML +HTML +JSON +CSV	+CSV File +JSON	+JSONL
Annotation types	+Entity recognition +Relations creation +Entity labelling +Document labelling +normalization	+Advice +Explanation +question mark +Comment +Reply +seeAlso +example +change	+Highlighting +Adding sticky notes +Tagging +Bookmarking	+Entity extraction +Text classification +Entity labelling +Document labelling	+Entity extraction +Sentiment analysis +Machine translation	+Entity recognition +Text classification +Computer vision

Annotation s storages	+local or cloud	+Remotely on the client in or on the	+remotel y on the server	Data stores	PostgreSQL database	SQLite, MyS QL, PostgreSQL
	infrastructu re +server	server in RDF database				
Team collaborati on and shared annotation s	Yes	Yes	Yes	Yes	Yes	No
Searching option	Yes	Yes	Yes	No	No	No
Annotation techniques	+Manual +semiautoma tic +automatic	+Semiautoma tic	+Semiautom ati c	+Manual Semiautoma tic +automatic	+Manual, +Semiautoma tic +automatic	+Manual, +semiautoma tic +automatic
Available for free testing	Yes	Yes	Yes	Yes	online demo	Online demo

Table 2: comparison table of annotation tools

4.3 Advantages of the annotation tools

Each tool presents its advantages. The advantages of an annotation tool are used to evaluate a tool for a user to be make his/her choice.

4.3.1 Advantages of Tagtog annotation tool

Here are some advantages of using Tagtog annotation tool:

- Tagtog minimize time and cost due to its speed and efficiency.
- Collaborative annotation.
- The user can change or delete the annotation depending on his/her permission.
- The annotation follows user guidelines which leads to high annotation accuracy.
- The user can make corrections and validations of pre-annotated data.
- Supports multiple languages (French, Swahili, Chinese, English, etc.).
- Most of annotation tools search by key, but Tagtog searching tool, searches by concept.
- Normalization
- Flexible: easily integrate with your workflow

4.3.2 Advantages of Annotea

- Annotea supports open ontologies which simplifies the extensibility of the data.

- The structure of annotated document is limited to HTML and XML-based documents. This makes the annotated document to be well structured.
- The user is able to classify the annotations at the time they are being created because the type of annotation is a metadata about annotation itself.
- The user can be able to reuse the information stored in RDF databases without modifying them.
- Local (private) and remote (shared) annotations. Annotations can be stored either locally in the user's host computer or in an annotation server.

4.3.3 Advantages of Diigo

- Diigo allows the user to bookmark webpages and read them later.
- The attached highlights and stickies can serve as a reminder.
- Trough Diigo annotation, you can share pages via social sites like twitter, Facebook, etc.
- Diigo is supported by all browsers and almost all operating platforms (computers, iPhone, Android, iPad, etc.).
- You can either keep your annotations private or share them in your groups.

4.3.4 Advantages of Gate

- The annotation types that do not occur in the text to be annotated, are not present in the annotation sets frame.
- Language resources and program resources are all stored together in a data store.
- Ontology based annotations.
- Restore application from file.
- Linking web pages to Ontologies using Information Extraction.

- Learning and evolving Ontologies via natural language analysis and lexical semantic network traversal.
- An Ontological Gazetteer for attaching instances of concepts in texts to Ontologies.

4.3.5 Advantages Doccano

- Supports team collaboration.
- Annotates any text regardless the language.
- Auto-labelling.
- Easy deployment to different web platform (Azure, Heroku, AWS).

4.3.6 Advantages of Prodigy

- Cloud-free
- Prodigy is self-contained and extensible.
- Customized web application with 13 annotation interfaces.
- Lifetime licenses
- Storage back-ends choice
- Supports any text format
- Active learning
- Achieve high accuracy annotation for the fact that the user is only focused on one task at a time

Chapter Five

Conclusion and Recommendations

5.1 Conclusion

In this research, a study of six text annotation tools has been done. The features, functionalities and properties of the tools have been discussed in details and tested.

After studying and critically testing the annotation tools, a comparative table that shows the comparison of the tools was produced. The comparison primarily focused on different parameters and properties inherent in the text annotation tools studied. The advantages of using each annotation tool were figured out.

All the studied tools have different features and properties; and they differ from one tool to another. Thus, there is no annotation tool that can be taken to be better than others. Depending on the project requirements and specifications, the user should always be able to find a tool that complies with his/her need.

The comparative table will guide users in making their choices of annotation Tools depending on what they are looking for and what is provided or presented by an annotation tool.

5.2 Recommendations

For future work, this research study can further be improved by studying as many annotation tools as possible, including image annotation tools and videos annotation tools focusing more on those that support Machine Learning.

References

Abahai, A. (2015). *Automated Document Annotation for Effective Data Sharing*. *8491*, 4–7.

AI-enabled Text Annotation Tool | PDF, Markdown, CSV, html, tweets, &amp; many more types of Documents. (n.d.). Retrieved May 19, 2019, from https://www.tagtog.net/

Bipboy, M. (2017). *Using Social Annotation Tool to Enhance Learning : A Literature Review*. 1–17.

Bodain, Y., & Robert, J. M. (2007). Developing a robust authoring annotation system for the Semantic Web. *Proceedings - The 7th IEEE International Conference on Advanced Learning Technologies, ICALT 2007*, (Icalt), 391–395. https://doi.org/10.1109/ICALT.2007.119

Brusilovsky, P. (2005). *Efficient techniques for adaptive hypermedia*. (June), 12–30. https://doi.org/10.1007/bfb0023957

Cejuela, J. M., McQuilton, P., Ponting, L., Marygold, S. J., Stefancsik, R., Millburn, G. H., & Rost, B. (2014). Tagtog: Interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. *Database*, *2014*, 1–8. https://doi.org/10.1093/database/bau033

CEUR workshop proceedings. (n.d.). Retrieved from https://gate.ac.uk/sale/tao/

Computer Vision · Prodigy · An annotation tool for AI, Machine Learning &amp; NLP. (n.d.). Retrieved May 19, 2019, from https://prodi.gy/features/computer-vision

Dasiopoulou, S., Giannakidou, E., Litos, G., Malasioti, P., & Kompatsiaris, Y. (2011). A survey of semantic image and video annotation tools. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6050*, 196–239. https://doi.org/10.1007/978-3-642-20795-2_8 doccano - Document Annotation Tool. (n.d.). Retrieved May 19,

2019, from https://doccano.herokuapp.com/

Euzenat, J., & Shvaiko, P. (2013). Ontology matching, 2nd Edition. In *Dados*. https://doi.org/10.1007/978-3-540-49612-0

Gosal, G. P. S. (2015). A Survey on Semantic Annotation of Text. *Ijarcsse*, *5*(9), 54–57.

Hanbury, A. (2008). A survey of methods for image annotation. *Journal of Visual Languages and Computing*, *19*(5), 617–627. https://doi.org/10.1016/j.jvlc.2008.01.002

Kahan, J., & Koivunen, M. (2001). Annotea : An Open RDF Infrastructure for Shared Web. *Proceedings of the 10th International Conference on World Wide Web*, 623–632.

Kenter, T., & Maynard, D. (2005). Using gate as an annotation tool. *University of Sheffield, Natural Language Processing Group*, (June). Retrieved from http://www.ia.hiof.no/softengin/ias/literature/sw/annogate.pdf

Khurana, K., & Chandak, M. B. (2013). *Study of Various Video Annotation Techniques*. *2*(1), 909–914.

Kiefer, C. (2016). Assessing the quality of unstructured data: An initial overview. *CEUR Workshop Proceedings*, *1670*, 62–73.

Kiryakov, A., Popov, B., Terziev, I., Manov, D., & Ognyanoff, D. (2004). Semantic Annotation , Indexing , and Retrieval 2 The Missing Fibres of the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, *2*(1), 49–79.

Li, J., Tang, J., Li, Y., & Luo, Q. (2009). RiMOM: A dynamic multistrategy ontology alignment framework. *IEEE Transactions on Knowledge and Data Engineering*, 36. https://doi.org/10.1109/TKDE.2008.202

Manthalu, S. (2014). *Annotating web search results*.

Nixon, L., & Troncy, R. (2014). Survey of semantic media annotation tools for the web: Towards new media applications with linked media. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-11955-7_9

Okunoye, O., Oladejo, F., & Odumuyiwa, V. (2010). *Dynamic Knowledge Capitalization through Annotation among Economic Intelligence Actors in a Collaborative Environment*.

Open source text annotation tool for machine learning practitioner. (n.d.). Retrieved May 19, 2019, from https://pythonawesome.com/open-source-text-annotationtool-for-machine-learning-practitioner/

Passin, T. B. (2004). *EXPLORER ' S GUIDE TO THE SEMANTIC WEB*.

Pernelle, & Nathalie. (2017). *Semantic enrichment of data : annotation and data linking To cite this version : M émoire S emantic enrichment of data : annotation and data linking*.

Prodigy · An annotation tool for AI, Machine Learning &amp; NLP. (n.d.). Retrieved May 19, 2019, from https://prodi.gy/

Raut, P., & Sawarkar, G. B. (2016). *A Survey on Unstructured Document Annotation Using Content and Query Value Based*. *3*(5), 154–159.

Robert, C. A. (2007). L ' ANNOTATION POUR LA RECHERCHE D ' INFORMATION DANS LE CONTEXTE D ' INTELLIGENCE ECONOMIQUE Thèse pour l ' obtention du Doctorat de l ' Université Nancy 2 L ' ' annotation pour la recherche d ' ' information dans le contexte d ' ' intelligence économiqu. *Victoria*.

Slimani, T. (2013). Semantic Annotation: The Mainstay of Semantic Web. *International Journal of Computer Applications Technology and Research*. https://doi.org/10.7753/ijcatr0206.1025

Text Classification · Prodigy · An annotation tool for AI, Machine Learning &amp; NLP. (n.d.). Retrieved May 19, 2019, from https://prodi.gy/features/text-classification

Tools. (n.d.). Retrieved May 19, 2019, from https://www.diigo.com/tools/

Tyler Manolovitz. (2005). *What is Diigo ? What Does Diigo Do ? Breakdown of Features & Tools*.

Uren, V., Hall, W., & Keynes, M. (n.d.). *Title : Semantic Annotation for Knowledge Management : Requirements and a Survey of the State of the Art*.