# DATA SENSE PLATFORM

**[1,2]MARIANA DIAS, [1,2,4]JOAO C. FERREIRA, [2]RUI MAIA, [2]PEDRO SANTOS, [1]RICARDO RIBEIRO, [1,3]ANA LUCIA MARTINS**

[1]Instituto Universitário de Lisboa (ISCTE-IUL), [1,2]InovInesc Inovação - Instituto De Novas Tecnologias, [3]Business Research Centre (BRU-IUL), [4]Information Sciences, Technologies and Architecture Research Center(ISTAR-IUL)
E-mail: jcafa@iscte-iul.pt

**Abstract**- The current manual or semi-automatic document preservation process suffers from various problems that particularly affect the handling of confidential or sensitive information, such as the identification of sensitive data in documents requiring human intervention that is costly and propense to generate error, and the identification of sensitive data in large-scale documents does not allow an approach that depends on human expertise for their identification and relationship. DataSense will be highly exportable software that will enable organizations to identify and understand the sensitive data in their possession in unstructured textual information (digital documents) in order to comply with legal, compliance and security purposes, identify and classify and relate sensitive data (Personal Data) present in large-scale non-structured information in a way that allows entities and/or organizations to understand it without calling into question security or confidentiality issues, and allowing companies that focus on their clients to better understand their profile from information collected from sensitive data or consent data or algorithms. The Data Sense project will be based on 3 key layers using the current potential of NLP technologies and the advances in machine learning (NER), Disambiguation and Co-referencing (ARE) and Automatic Learning and Human Feedback. It will also be characterized by the ability to learn from human feedback automatically, correcting and iteratively improving the AI model that supports it.

**Indexterms**- Sensitive Data, Natural Language Processing, Text Mining, Named Entities Recognition, Co-reference Resolution.

## I. INTRODUCTION

In the context of an information society where more and more documents are generated and collected from various sources and by various entities, it is only natural that this situation raises more and more security concerns. The complexity and severity of security issues in systems and even individuals depends crucially on how organizations deal with sensitive data of any kind. These are problems that have worsened over time in a fully digitized society that generates processes, and stores large-scale amounts of information that easily leads to a loss of control over the content of these documents. In the past, as an example for documents that needed to be public, the data considered to be sensitive to an entity or individual and abstracted by manual procedures, duly documented and structured using fixed rules in a process called "sanitization", were manually identified. More recently, tools have been created that help the identification process with a particular focus on structured information such as emails, addresses, phone numbers or credit cards, while leaving all sensitive data of a textual and unstructured nature as is the case of names, medical information, criminal records, religion to the care of human expertise to identify and treat them. All this manual and semi-automatic process suffer from several problems, namely:

1. Identifying sensitive data in one document (or several) requires tasks that are manual, error prone, and therefore very costly;
2. Their identification in large-scale documents (e.g. thousands of documents) does not allow an approach that depends on human expertise in their identification and relationship in most cases;
3. It is not possible to automatically establish a relationship between several unstructured documents and the sensitive data present, so it is not possible to verify the content of the documents with regard to the presence of sensitive data.

Since the "identification" of sensitive data makes up an important part of the whole process (even if one uses only human expertise), this is only part of an even bigger problem. A large number of documents with textual and unstructured information may contain references to data that are inherently linked: The name and address of an individual can be found in a document but your medical or criminal information may be scattered in several other documents. The referencing of sensitive data in unstructured documentary information is essential to associate individuals and organizations with dispersed information. This structuring of information, resolution of references and classification is crucial for the protection of the confidential information present in a documental database. Incorrect management of this type of data can put public or private organizations in very complex situations even in illegal situations. In order to combat such situations, it is necessary for organizations to have the means to detect them and to carry out, in parallel, integrated management of all sensitive data in accordance with existing standards and legislation. For this, it is essential that organizations and entities have the ability to perceive "where" they are, what "type" they are and "how" they relate the data they have. Only with a strong

understanding of sensitive data, namely their identification, classification and the relationships they have (regardless of their format) will it allow organizations to deal with a problem that has become too complex, expensive and under more tight. This understanding, once obtained, allows organizations to perceive, create and systematize preventive security policies, educate users in their manipulation, set tight controls for sensitive data, and implement rules in accordance with current legislation. There are mandatory responses that will have to be given by entities and organizations to a number of existing issues, such as the right to forgetfulness, the request for access to personal data stored by users, temporary authorizations to store and process personal and sensitive data, as well as, the automatic processing of sensitive information.

In our work, we create a platform that allows acting in the area of the discovery of data considered Sensitive (Sensitive Data Discovery). DataSense has two fundamental objectives:

1. Allow the identification, classification and relationship of sensitive data present in unstructured information on a large scale in order to allow entities and organizations to obtain an understanding of their sensitive data;
2. Allow organizations to respond immediately to the content and network (direct and indirect relationships) of the sensitive data they store and process (e.g., right to forget).

In order to respond to the aforementioned objectives, DataSense is based on four concepts essential to overcome the state of the art of application and proposes a hybrid architecture that will take the risk of applying the area of Natural Language Processing and Automatic Learning (Machine Learning) in the critical area of sensitive data protection. The concepts, described in detail in the next chapter are: Sensitive Data (Personal Data), Natural Language Processing, Humanly readable multi-format unstructured information analysis and training supported in human feedback. These basic concepts of the proposed solution are supported by three layers of Artificial Intelligence: Identification of Named Entities, Machine Learning models for resolution of Coreference and Entity Linking, Human Feedback and incremental learning of the models.

## II. RELATED CONCEPTS AND WORK

In the banalization of the commercial discourse on Artificial Intelligence solutions, there was a considerable growth of business investment in the most diverse sub-areas of this topic, which does not escape the Natural Language Processing (NLP) and where the DataSense is located. NLP is used in numerous business applications ranging from personal assistants in smartphones to real-time translation systems and social-emotional analysis. More recently, NLP has begun to be expanded to incorporate more mature models with better levels of efficiency, and the result is more intelligent and capable applications. In the commercial area, the use of natural language processing at an advanced level for identification, classification, relationship and automatic learning based on human feedback is not an area that has developed sufficiently, as happened for example in other areas such as the processing and identification of speech that is currently used in many areas.

There are several examples of applications that work in the area of eDiscovery (Electronic Discovery). Cicayda (https://cicayda.com/), for example, looks for documentary information, legal information data to catalogue and perform a risk analysis using non-detailed natural language techniques. Another solution in the market is called Onna (https://onna.com/) that allows a search in different repositories but uses standards techniques only to detect unstructured information found and catalogued. Both solutions andessentially generic systems of Electronic Discovery that classify only the metadata of the documents and not their contentand in some cases with very simple approaches to natural language processing such as regular expression processing..Another problem associated with these systems is that the technological approach does not allow for a co-reference resolution of discovered entities nor the ability to improve based on human feedback, which makes them unable to learn over time. Also, these systems do not allow its use in other languages like Portuguese. In this context and knowing that Portuguese is a language with more than 200 million speakers in the world, its important considerate for these type of systems.

However, there are much more advanced possibilities that can be applied and that DataSense integrates in its approach. Some of the most advanced concepts in this technical-scientific area and some of the open challenges are described below.

Progress in the area of Artificial Intelligence, particularly in the area of natural language technologies has been notable, with visible effects on the quantity and quality of products, systems and applications based on natural interaction. Firstly, because there are areas where the sensitivity of information is decisive and any error can have serious consequences. For example, it is not possible to apply massively and easily natural language processing systems in the legal area of the courts or in the medical field. Resolve this type of problems implies necessarily the ability to extract information, classified information and identify documents in large databases and relational documentary information. There is an insufficient number of the corpus or annotated data sets to train and validate this type of systems in the European Portuguese language and in the specific area of sensitive information. Resources in European Portuguese are generally much more limited than those in languages such as English, and

therefore there are not many production systems based on natural language processing in PT-EU. The context and challenges of the applications supported by Artificial Intelligence, namely in the area of Natural Language Processing (NLP) are evidenced in the solutions of Extraction and Information Retrieval and Named Entity Recognition. These are aimed at obtaining the semantic structure - the objects, their relations and actions from data in written natural language, which in the most complex cases, may not be structured.

In addition to the challenges inherent in the complexity of large documentary systems with various data sources in various formats, under typically poor quality, morphological variability is added. The different ways of writing a sentence and the ambiguity of the natural language itself (different meanings of a word, expression or phrase) characterize a high level of complexity in the development of a solution based on Natural Language Processing. This challenge for European Portuguese is even greater when compared to more exploited languages such as English or French, or even Portuguese-Brazilian [1] [2].

In the context of the DataSense project, it is important to mention that it is very relevant to identify and define the fundamental ontology for the project. This should be distinguished from ontologies used in other domains and languages [3] by the integration of semantic knowledge in the area of sensitive data in the area of information structuring, namely extraction and retrieval.

Natural Language is a common component of all Artificial Intelligence applications based on natural language understanding.Most NLP-based projects for specific domains use rule-based modules, such as regular expressions and syntax rules. However, this approach entails two problems: 1) the system only recognizes a limited set of rules; 2) Extension or improvement requires manual labour of someone who knows the domain and the formalism of rule-making. Other techniques, based on Machine Learning, can learn to classify or even generate new rules but assume the existence of a known documentary corpus.

One of the technical-scientific challenges that the system will have to deal with is an inter-document reference. While intra-document reference allows the delimitation of the scope of co-referencing of entities, references between documents raise important difficulties with regard to the disambiguation of entities that may not be fully and uniquely identified. In addition to the previously mentioned technical-scientific challenges - which are just a few of the main ones - users often write based on their own writing style, fluctuations that require the need to update the models and rules used. In order for DataSense to respond efficiently, it will be necessary for all rules, intra and inter-reference, new expressions and new rules, to be automatically or semi-automatically reintroduced into the knowledge of the model. These updates allow the system to follow the dynamics of written and language evolution, with the introduction of terms, alteration of others, or changes in the rules of speech and writing that are part of our day-to-day life.

In the identification of named entities (or Named Entity Recognition) elements of the text (ex: names) and according to types of individual are identified. For example, the well-known shared task of CoNLL 2003 [4] groups the names into three classes (organizations, places and people). There are even more detailed, sometimes hierarchical categorizations in specific domains of application such as biomedical and legal texts [5][6]. Initially, the NER approaches were based on rules [7] [8] from the ones designed specifically for the language and application domain under study. Subsequently, statistical approaches and neural networks [9] emerged, which can be trained for different types of data and domains.

Coreference Resolution is a well-known area within Natural Language Processing. It has had successive developments being the target of recent application of approaches based on latent structures [10] or reinforcement learning [11] for example. Based on Recurrent Neural Networks (RNN) [9] and Long Short-Term Memory (LSTM) neural networks, and taking into account the specific linguistic knowledge of a given language and domain [12], they are considered as improving the results this complex area.

## III. PROPOSAL

The DataSense solution defines and integrates four fundamental concepts that are integrated into the field of extraction and retrieval of sensitive data present in large unstructured databases. The concepts and the hybrid approach are detailed below.

Concept 1 - Sensitive Data (Personal Data). Despite the convenience of using the acronym PII (Personal Identification Information), in the European context there is something that is not a direct synonym, but something called PD (Personal Data). Personal Data is supported by three different Directives: 95/46 / EC (Data Protection Directive that was replaced in May 2018 by the GDPR Directive), 2002/58 / EC (E-Privacy Directive which was also replaced in May 2018 by E-Privacy Regulation) and 2006/24 / EC - Article 5 (Data Retention Directive). The Data Protection Directive such as the GDPR regulates the processing of personal data in the European Union. The General Data Protection Regulation (GDPR) directive, which applies a set of rules to return control of sensitive data to citizens and also establish clear and objective rules on deadlines, mechanisms and penalties for non-implementation. The E-Privacy directive, which was replaced at the same time as the Data Protection Directive by E-Privacy Regulation, defines rules on confidentiality of information,

treatment of spam, handling of cookies, etc. Adding Directive 2006/24 / EC which regulates data retention in particular in the telecommunications industry.

For the purposes of this work, will be defined as a structure that represents the sensitive data that can identify, contact, or locate an individual. This crucial work will be carried out based on the best practices and accumulated know-how in the area of Sensitive Data. The basic structure of information can be described already, and generically, through the following information contexts: 1) Identification personal information (name, email, social security number, credit card number, voter card, etc.); 2) Information about kinship and family relationships; 3) Medical or biometric information; 4) Criminal information; 5) Information about religion; 6) Information about entities, organizations or companies directly associated (e.g. employment) with the individual; 7) Gender information.

Figure 1, shows an illustration of the sensitive data, person A and a set of sensitive data were considered.

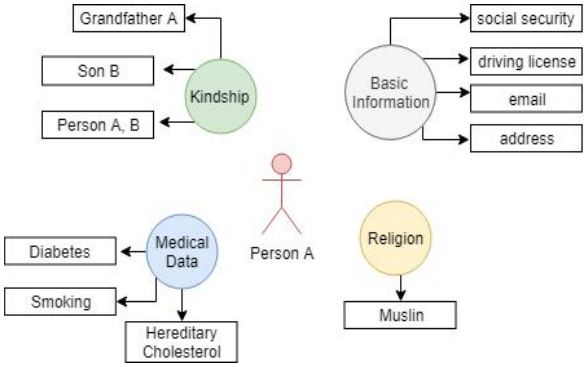


**Figure 1: Sensitive Data Illustration concept**

Concept 2 - Natural Language Processing. Natural Language Processing (NLP) is a subarea of Artificial Intelligence (AI) that studies the ability and limitations of a machine to understand the natural spoken or written language of humans. The purpose of NLP is to provide computers with the ability to understand texts that humans easily understand and interpret but often do not follow the formal, syntactic and semantic characteristics of grammar and definitions considered to be formally correct in that language. "Understanding" a text means recognizing the context, performing syntactic, semantic, lexical and morphological analysis, creating abstracts, extracting information, interpreting the senses, analyzing feelings, and even learning concepts with processed texts. Thus the use of natural language processing is mandatory in the solution and totally innovative in this area of application - namely for European Portuguese - given that it will allow, through trained NLP models, to perform the identification, classification and extraction of relations between sensitive data.

Concept 3 - Multi-format and unstructured information analysis. Once the solution is able to identify, classify and create relationships between sensitive data, as previously mentioned, it is necessary to apply a drill-down mechanism of unstructured textual information. For the realization of this mechanism will be used various techniques of analysis in documents and applied the models of NLP developed in Concept 2 with the objective of extracting the sensitive data found in the various documents analyzed. Figure 2 illustrates a possible high-level relationship between several documents analyzed.The relationship criterion is based on the identification and classification of the sensitive data found in each document. Relationships can be circular and multilevel.

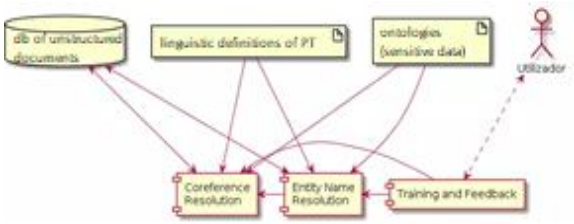


**Figure 2: Document Relationship Analysis**

Concept 4 – Training supported by human feedback.

In this concept, it will be possible to ask the user through a specific web interface, the ability to correct predictions and results obtained by the models in order to influence and improve the capabilities of NLP models. This capability gives the solution a unique way of learning from errors and inaccuracies committed by the NLP model in relation to the identification, classification and relationship between sensitive data. Typically, these training processes involve three phases:

1. The trained model uses the prior formal knowledge of the language and data that has been previously classified to process a set of documents.
2. Periodically, and according to the system's trust levels about the process, the system sends a user the documents whose data identification or extraction of information was not performed with the expected degree of confidence. Also, periodically (and independently of the results of the Sensitive Data Discovery process) the system submits one or more documents to a user for validation of the result of the identification and classification.
3. User changes to the process allow you to inform the system of the correct option. The new phases of training of the model are based on feedback the so system error is progressively being minimized, resulting in an overall improvement of the system and allowing better alignment with the user's expectations and understanding.

## IV. TECHNICAL-SCIENTIFIC APPROACH TO THE SOLUTION

The DataSense solution intends to aggregate previously defined concepts (1 to 4) into a simple, scalable system that is capable of fulfilling all the objectives it proposes. The tool will have a capacity

to process large amounts of data and will achieve a level of accuracy close to the human being. For this, we describe here the essential steps in the design and implementation of the system based on Natural Language Processing for Information Extraction and Retrieval.

In the design of NLP models, the creation of a well-defined corpus is the genesis of development. The corpus is the basic input of the trained models and is the main source of information about the implementation of rules, expressions or other formal definitions in the system. The latter is often used in the areas of Machine Learning and Information Systems and allows us to interpret the structured documentary perspective that is intended for the project. With regard to the data needed to train the models (statistical or linguistic, for example) used in the project, DataSense has the advantage of pre-existing a large set of unstructured documents and others already structured.

One of the most complex tasks in the NLP-based Information Extraction and Recovery areas is the mapping of unstructured information dispersed in one or several documents in a relational information structure. This important technical-scientific task emerges as the basis of the project and will be approached with a stacked method framework based on the following layers:

a) NER- based Regular Expressions and Machine Learning techniques: a set of approaches based on Regular Expressions and Machine Learning models that allow identifying Named Entities (NE) [13]. This layer will allow us to construct a form of information visualization structured according to the Sensitive Data. It should be noted that the construction of regular expressions will be carried out through an innovative way.

Likewise, it is important to note that the definition of sensitive data in the context of the project is the definition of Personal Identification Information (PII) - that is, information that can identify, contact or locate an individual in the following contexts:

•Basic identification information (name, email, social security number, credit card number, etc.)
•Information about kinship and family relationships
•Medical or biometric information
•Criminal information
•Information about membership, education or religion
•Other (to identify during project execution)

In the context of the project will also evaluate some metrics of the methods and models (precision, F1-score, recall) that allow evaluating the performance in the identification and classification of sensitive data. The decision of which metrics to use will check for those with higher accuracy, fewer false positives, less ambiguity and metonymy, and future learning ability as a foundation in layer c).

b)Co-referencing and Automatic Relations Extraction: A second layer of the process will be ensured by the knowledge of the language [14] and by the models trained for Entity Linking and Co-referencing Resolution. The idea of extracting information is to link two entities (entities whose data are sensitive) present in one or more unstructured documents.

Thus, the idea is to extract semantic relationships between entities so that these are the final result of the information being sought. An example is the ability to process a document (or several) and associate a person's name with their address and also the city where they live, as well as the degree of kinship associated with another person.

This syntactic processing of documents allows the generation of classification of each term or expression. Then, using a word engine embeddings [15] [16] (based on neural networks-LSTM) will extract the characteristics of the sentences and paragraphs that will be used in a classifier of the Support Vector Machines or Conditional Random Fields[17][18]. These classifiers serve to disambiguate and assign a score to each association in the previous layer. In this domain, during the execution of the project, several neural network topologies will be considered in order to evaluate their use as classifiers. For this, we will analyze some of the most updated works in these areas applicable to the linguistic domain [19]. These networks will allow co-referencing in a hybrid way, in conjunction with heuristic methods. All methods have inherently different definitions, but the result will be assessed in the TOEFL test for the ability to construct less ambiguous and more precise intra-documentary relationships between data considered to be sensitive.

c) Update by Explicit Human Feedback: This layer will allow the users to be able to verify the assignments and confirm or change the values automatically placed by the system in the information structure, that is, in the document database. The information introduced will allow the adaptation of the models used in the first layer of NER and in the second layer of Co-referencing Resolution. Both of these layers can, in certain situations, suffer from problems of ambiguity or false positives inherent in the linguistic domain itself. For this reason, it is only natural that they have an accuracy not as high as you would expect. In order to solve this problem, it is intended, whenever necessary, to submit the NER (Named Entity Recognition) and CRR (Coreference Resolution) engines to human validation in situations where their accuracy is lower than a certain threshold previously defined as satisfactory. This mechanism will allow the models to learn automatically as a result of user feedback (with Machine Learning techniques) in order to improve theireffectiveness. This third technological layer of the project - in the area of documentary management of sensitive information - provides the system of innovative supervised and semi-supervised learning capability for Portuguese-European. The result of this hybrid three-layered approach will be a system capable of

identifying, classifying relationships and processing unstructured textual information on a large scale. It will also be characterized by the ability to learn from human feedback automatically, correcting and iteratively improving the artificial intelligence model that supports the system itself. Vocabulary validation / correction should be done based on what is known by the system and not on external dictionaries. There are many editing measures, such as the minimum editing distance [20][21] that will allow you to check for spelling mistakes or alternative ways of writing. These approaches also allow us to monitor some language dynamics and understand a set of less clear situations so that, through explicit feedback, the system can be informed of the correct or updated representation for a term.

## CONCLUSION

We describe our work towards a severe problem of sensitive information. As a way to give answers that allow to understand and define controls for the sensitive data and to solve the automatic processing of sensitive information. This system will be implemented using Natural Language Processing techniques and Machine Learning techniques, of supervised and unsupervised learning. DataSense must bea functional prototype capable of correctly executing a set of tasks that must obey a set of objectives. Objectives can be subdivided into three components of the system: Recognition and Classification of Named Entities, Co-Reference Resolution and Training supported by human feedback. In addition to discovering and classifying sensitive data, DataSense must be able to identify the entities present in the documents and databases, finding relationships between them and create a network linking the discovered entities to the sensitive data encountered.

## REFERENCES


[1] E. Fonseca, R. Vieira, and A. A. Valin, “Coreference Resolution in Portuguese: Detecting Person, Location and Organization,” Learn. Nonlinear Models, vol. 12, pp. 86–97, Jan. 2014.

[2] J. G. C. de Souza, P. N. Gonçalves, and R. Vieira, “Learning Coreference Resolution for Portuguese Texts,” in Computational Processing of the Portuguese Language, 2008, pp. 153–162.

[3] “OntoNotes Release 4.0 - Linguistic Data Consortium.” [Online]. Available: https://catalog.ldc.upenn.edu/ldc2011t03. [Accessed: 16-Feb-2018].

[4] E. F. Tjong Kim Sang and F. De Meulder, “Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition,” in Proceedings of the Seventh Conference on Natural Language

[5] J. Kazama, T. Makino, Y. Ohta, and J. Tsujii, “Tuning Support Vector Machines for Biomedical Named Entity Recognition,” in Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain - Volume 3, Stroudsburg, PA, USA, 2002, pp. 1–8.

[6] M. Surdeanu, R. Nallapati, and C. Manning, “Legal claim identification: Information extraction with hierarchically labeled data,” in Workshop Programme, 2010, p. 22.

[7] R. Collobert and J. Weston, “A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning,” in Proceedings of the 25th International Conference on Machine Learning, New York, NY, USA, 2008, pp. 160–167.

[8] S. J. Wiseman, A. M. Rush, S. M. Shieber, and J. Weston, “Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution,” 2015.

[9] H. Palangi et al., “Deep Sentence Embedding Using Long Short-term Memory Networks: Analysis and Application to Information Retrieval,” IEEEACM Trans Audio Speech Lang Proc, vol. 24, no. 4, pp. 694–707, Apr. 2016.

[10] S. Martschat and M. Strube, “Latent structures for coreference resolution,” Trans. Assoc. Comput. Linguist., vol. 3, no. 1, pp. 405–418, 2015.

[11] K. Clark and C. D. Manning, “Deep reinforcement learning for mention-ranking coreference models,” ArXiv Prepr. ArXiv160908667, 2016.

[12] B. Dhingra, Z. Yang, W. W. Cohen, and R. Salakhutdinov, “Linguistic Knowledge as Memory for Recurrent Neural Networks,” ArXiv170302620 Cs, Mar. 2017.

[13] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification,” Lingvisticæ Investig., vol. 30, no. 1, pp. 3–26, Jan. 2007.

[14] S. J. Wiseman, A. M. Rush, S. M. Shieber, and J. Weston, “Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution,” 2015.

[15] H. Palangi et al., “Deep Sentence Embedding Using Long Short-term Memory Networks: Analysis and Application to Information Retrieval,” IEEEACM Trans Audio Speech Lang Proc, vol. 24, no. 4, pp. 694–707, Apr. 2016.

[16] D. Ganguly, D. Roy, M. Mitra, and G. J. F. Jones, “Word Embedding Based Generalized Language Model for Information Retrieval,” in Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 2015, pp. 795–798.

[17] T. Joachims, “Text categorization with Support Vector Machines: Learning with many relevant features,” in Machine Learning: ECML-98, 1998, pp. 137–142.

[18] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon, “Adapting Ranking SVM to Document Retrieval,” in Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 2006, pp. 186–193.

[19] B. Dhingra, Z. Yang, W. W. Cohen, and R. Salakhutdinov, “Linguistic Knowledge as Memory for Recurrent Neural Networks,” ArXiv170302620 Cs, Mar. 2017.

[20] E. S. Ristad and P. N. Yianilos, “Learning string-edit distance,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 5, pp. 522–532, May 1998.

[21] G. Sidorov, H. Gómez-Adorno, I. Markov, D. Pinto, and N. Loya, “Computing text similarity using Tree Edit Distance,” in 2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC), 2015, pp. 1–4.


★★★