Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

everis
an NTT DATA Company

Analizando las directivas de la EU – Tecnologías del Lenguaje y Big Data

25 de Abril

DESARROLLO Y POTENCIAL DE LAS TECNOLOGÍAS DEL LENGUAJE

attitude makes the difference

conceptos

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

everis
an NTT DATA Company

big data | analytics | cognitive

la interpretación de everis:

Center of Excellence CADC
everis
ITAINNOVA
INSTITUTO TECNOLÓGICO DE ARAGÓN

La Comisión Europea

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

everis
an NTT DATA Company

Hay 32.966 funcionarios generando documentos.

24 idiomas oficiales.
Traducciones.
Jerga legal.

la complejidad…

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

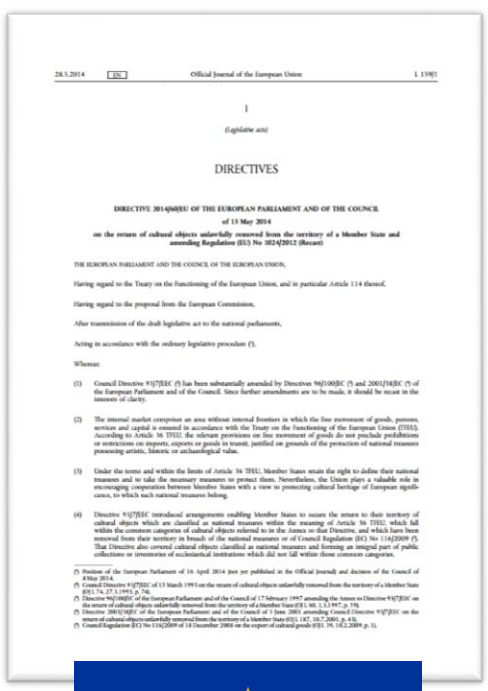
everis
an NTT DATA Company

El proceso legislativo es largo, complejo.

La directivas nacen en Europa, se traducen a todos los idiomas oficiales.

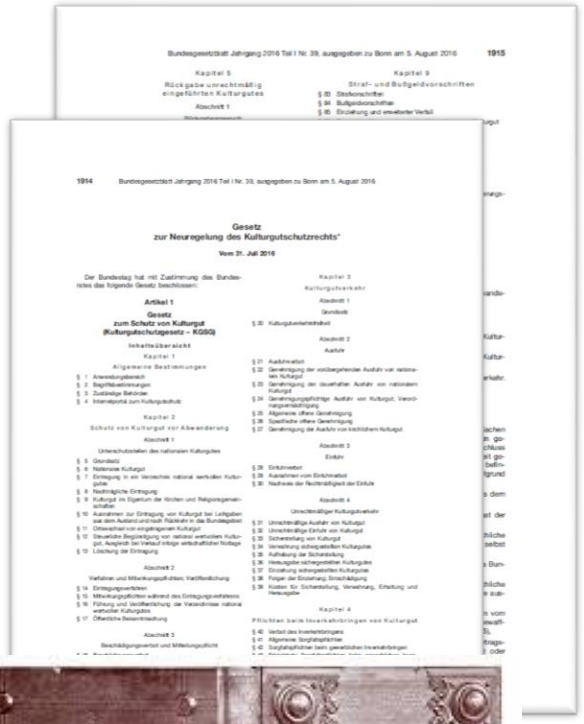
Se trasladan al marco legal de cada estado miembro.

Cada estado informa (opcionalmente) a la Comisión de cada transposición.



DIRECTIVES

DIRECTIVE 2014/60/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL
of 15 May 2014
on the return of cultural objects unlawfully removed from the territory of a Member State and amending Regulation (EU) No 1024/2012 (Recast)



Gesetz
zur Neuregelung des Kulturgutschutzrechts*

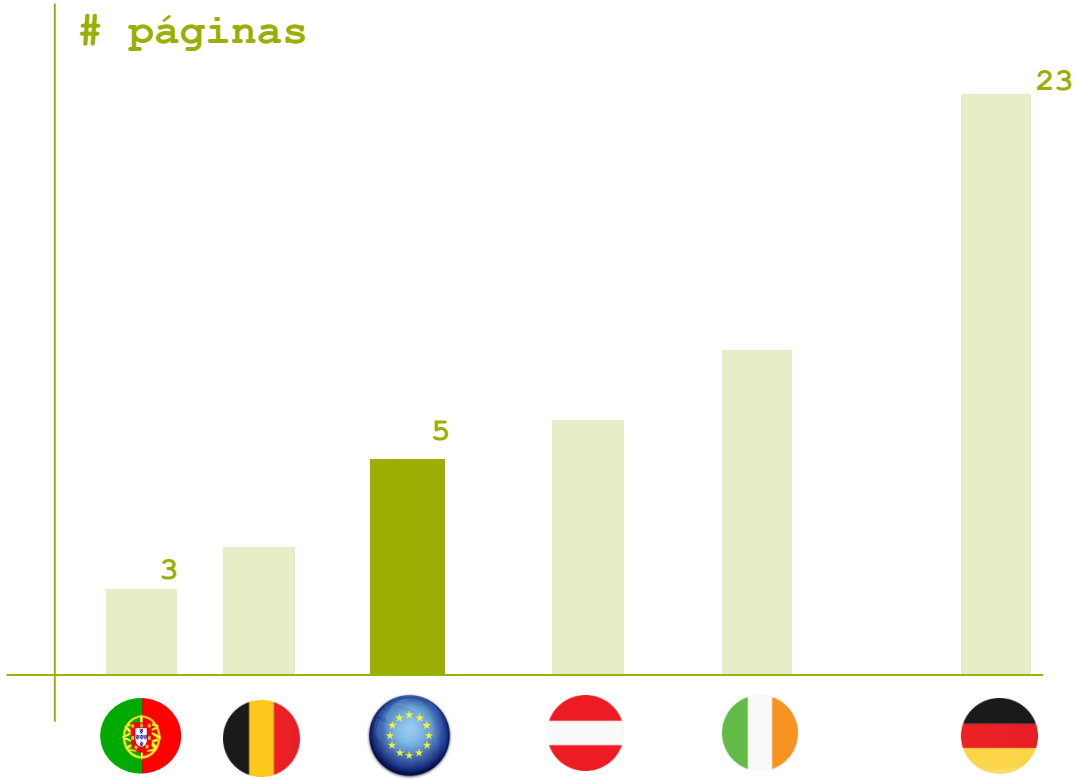
y la idiosincrasia…

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

everis
an NTT DATA Company

La directiva, sus transposiciones y el contenido.

Quizás es una medida del tamaño de las palabras, el amor por la literatura o la capacidad de síntesis en Europa.



el reto

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

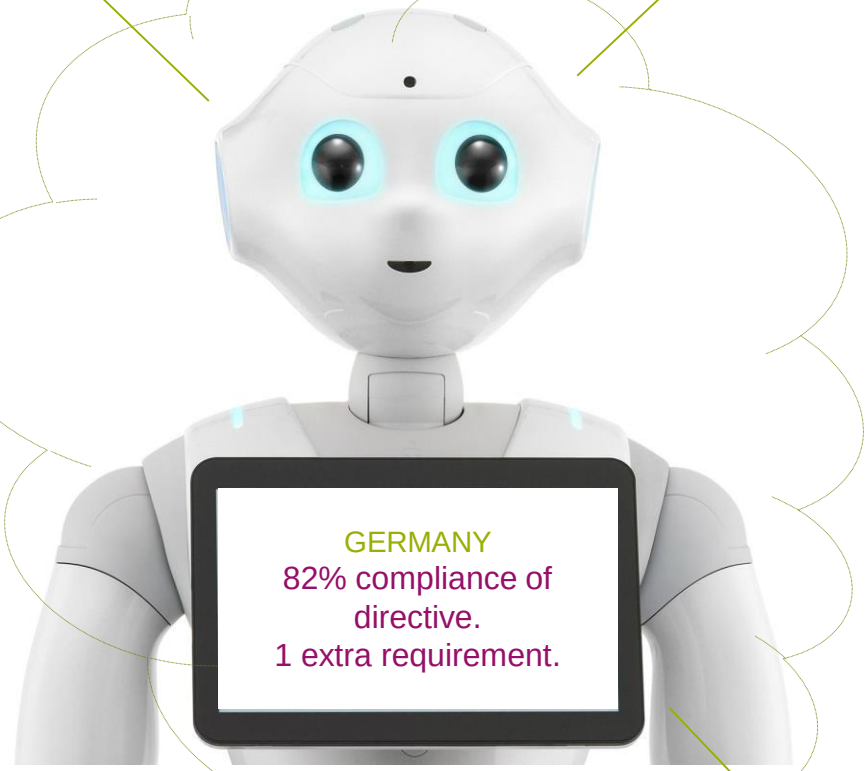
everis
an NTT DATA Company

QUE HABLE TODOS LOS IDIOMAS DE LA COMISIÓN

QUE INTERPRETE LAS TRANSPOSICIONES Y LAS COMPARE CON LA DIRECTIVA

“NECESITAMOS UNA MÁQUINA

QUE HAGA MAGIA CON TODOS ESTOS PAPELES.”



QUE MIDA LA CONFORMIDAD DE LA TRANSPOSICIÓN CUANTITATIVAMENTE

QUE DETECTE INCONFORMIDADES O GOLD-PLATING

la primera: analítica clásica

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

everis
an NTT DATA Company

Natural Language Processing (NLP).

Implementamos un modelo con diccionarios, stop words, lematización … para matrices de comparación “artículo vs párrafo”.

	Art1	Art2	Art3	Art4	Art5	Art6	Art7	Art8	Art9	Art10
Paragraph1	14%	8%	8%	10%	12%	8%	8%	10%	12%	10%
Paragraph2	10%	8%	10%	10%	8%	8%	8%	10%	10%	18%
Paragraph3	8%	13%	7%	24%	13%	4%	8%	3%	6%	16%
Paragraph4	10%	8%	9%	8%	13%	10%	8%	13%	9%	12%
Paragraph5	10%	10%	10%	10%	10%	12%	10%	10%	10%	10%
Paragraph6	10%	10%	10%	10%	8%	15%	10%	7%	8%	13%
Paragraph7	12%	9%	9%	9%	9%	12%	9%	12%	9%	9%
Paragraph8	31%	9%	9%	6%	4%	6%	5%	9%	7%	11%
Paragraph9	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
Paragraph10	10%	10%	10%	8%	8%	10%	8%	12%	8%	12%
Paragraph11	20%	5%	8%	6%	9%	24%	8%	7%	7%	6%
Paragraph12	13%	9%	9%	11%	9%	11%	13%	9%	9%	11%
Paragraph13	14%	9%	9%	9%	12%	9%	9%	9%	9%	9%
Paragraph14	9%	9%	9%	9%	9%	9%	12%	9%	12%	12%
Paragraph15	10%	10%	10%	10%	10%	10%	10%	12%	10%	10%

Figure 5 - Example of compliance matrix

la segunda: big data

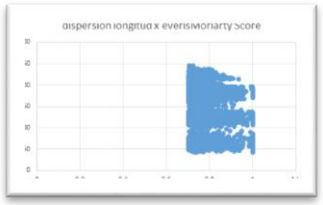
Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

everis
an NTT DATA Company

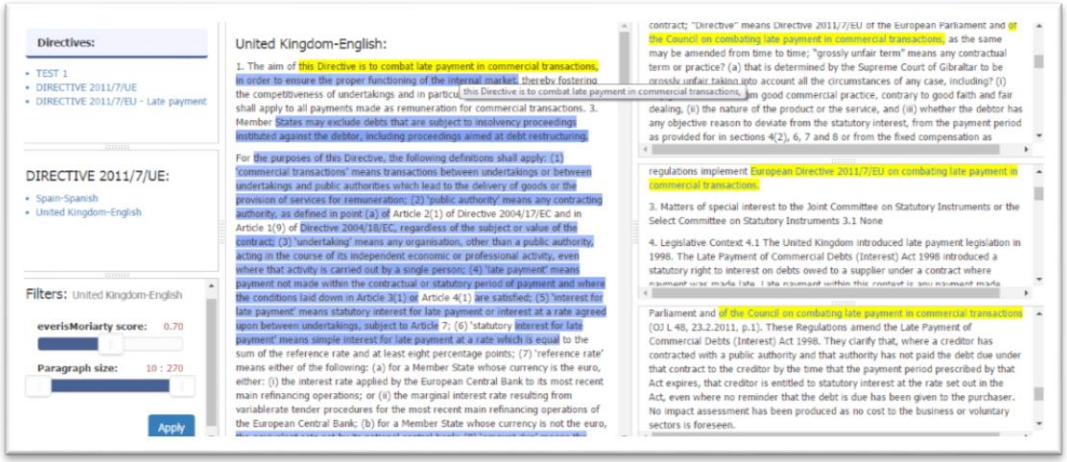
Analítica y Big Data.

everis MORIARTY
Cognitive Applications Development Platform · powered by ITAINNOVA

Spark



Implementamos un proceso con técnicas Big Data para procesar comparaciones de n-gramas de longitud variable de 1 a 250.



cadena_orig	everisMoriarty_score	longitug	cadena_origen	cadena_final
			that the period for payment fixed in the contract does not exceed the time limits provided for in paragraph 3, unless otherwise	period for payment fixed in a relevant contract shall not exceed the time limits provided for in section 6 unless otherwise expressly agreed
6301	0,769230769	176	expressly agreed in the contract and provided it	contract and provided such
			means to increase awareness of the remedies for late payment among undertakings. 4. Member States may encourage the	increase awareness of the remedies for late payment among undertakings. Establishment of codes. 17. The Minister may encourage the
6944	0,768867925	204	establishment of prompt payment codes which set out clearly defined payment time limits	prompt payment codes which set out clearly defined payment time
			30 calendar days from the date of receipt of the goods or services, unless otherwise expressly agreed in the contract and any tender	exceed 30 calendar days from the date of receipt of the goods or services, unless otherwise expressly agreed in the relevant contract and
6257	0,768786127	173	documents and provided it is not grossly	not grossly unfair
			national central bank; (8) â€˜amount dueâ€™ means the principal sum which should have been paid within the contractual or	substitute the following definitionsâ€˜ â€œâ€œamount dueâ€ means the principal sum which should have been paid within the contractu
2724	0,768707483	139	statutory period	period
3862	0,768595041	120	date of receipt of the goods or services, unless otherwise expressly agreed in the contract and any tender documents and	from the date of receipt of the goods or services, unless otherwise expressly agreed in the relevant contract and provided
			fixed in the contract does not exceed the time limits provided for in paragraph 3, unless	fixed in a relevant contract shall not exceed the time limits provided for in section 6 unless otherv
3906	0,768595041	119	otherwise expressly agreed in	agreed
3932	0,768595041	113	objectively justified in the light of the particular nature or features of the contract, and that it in any event	agreed is objectively justified in the light of the particular nature or features of the relevant contract in question, and
3996	0,768595041	121	paid by the agreed date, interest and compensation provided for in this Directive shall be calculated solely on the basis	date, shall attract the interest and compensation provided for in this Act, and shall be calculated solely on the basis
			date of the receipt of the invoice or the equivalent request for payment is uncertain, 30 calendar days after the date of receipt of the	receipt of the invoice or an equivalent request for payment is uncertain, 30 calendar days immediately following the date of receipt of the g
5448	0,768361582	172	goods or services; (iii) where the	supplied by the creditor
			date of receipt of the invoice or the equivalent request for payment is uncertain, 30 calendar days after the date of the receipt of the	receipt of the invoice or an equivalent request for payment is uncertain, 30 calendar days immediately following the date of receipt of the g
5868	0,768361582	172	goods or services; (iii) where the	supplied by the creditor
3267	0,76744186	120	days, unless otherwise expressly agreed in the contract and provided it is not grossly unfair to the creditor within the	services, unless otherwise expressly agreed in the relevant contract and provided that it is not grossly unfair to the creditor. (2)
			not exceed 30 calendar days from the date of receipt of the goods or services, unless otherwise expressly agreed in the contract and	of that procedure shall not exceed 30 calendar days from the date of receipt of the goods or services, unless otherwise expressly agreed

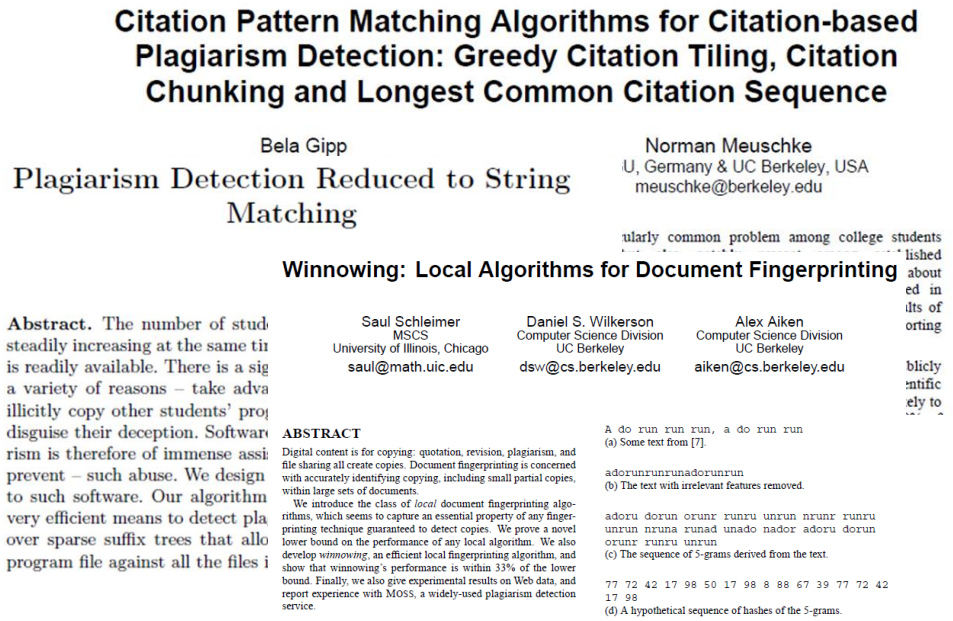
la tercera: cientifica

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

everis
an NTT DATA Company

Analítica, Big Data y Ciencia.

Apliquemos nuevos avances en investigación analítica e introduzcamos las mejoras que nos lleven a una solución óptima.



Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence

Bela Gipp

Plagiarism Detection Reduced to String Matching

Norman Meuschke
U, Germany & UC Berkeley, USA
meuschke@berkeley.edu

ularly common problem among college students
lished
about
ed in
ilts of
orting

Winnowing: Local Algorithms for Document Fingerprinting

Saul Schleimer
MSCS
University of Illinois, Chicago
saul@math.uic.edu

Daniel S. Wilkerson
Computer Science Division
UC Berkeley
dsw@cs.berkeley.edu

Alex Aiken
Computer Science Division
UC Berkeley
aiken@cs.berkeley.edu

Abstract. The number of stud
steadily increasing at the same ti
is readily available. There is a sig
a variety of reasons – take adva
illicitly copy other students' pro
disguise their deception. Softwar
rism is therefore of immense assi
prevent – such abuse. We design
to such software. Our algorithm
very efficient means to detect pla
over sparse suffix trees that allo
program file against all the files i

blicly
ntific
ely to

ABSTRACT

Digital content is for copying: quotation, revision, plagiarism, and file sharing all create copies. Document fingerprinting is concerned with accurately identifying copying, including small partial copies, within large sets of documents.

We introduce the class of *local* document fingerprinting algorithms, which seems to capture an essential property of any fingerprinting technique guaranteed to detect copies. We prove a novel lower bound on the performance of any local algorithm. We also develop *winnowing*, an efficient local fingerprinting algorithm, and show that winnowing's performance is within 33% of the lower bound. Finally, we also give experimental results on Web data, and report experience with MOSS, a widely-used plagiarism detection service.

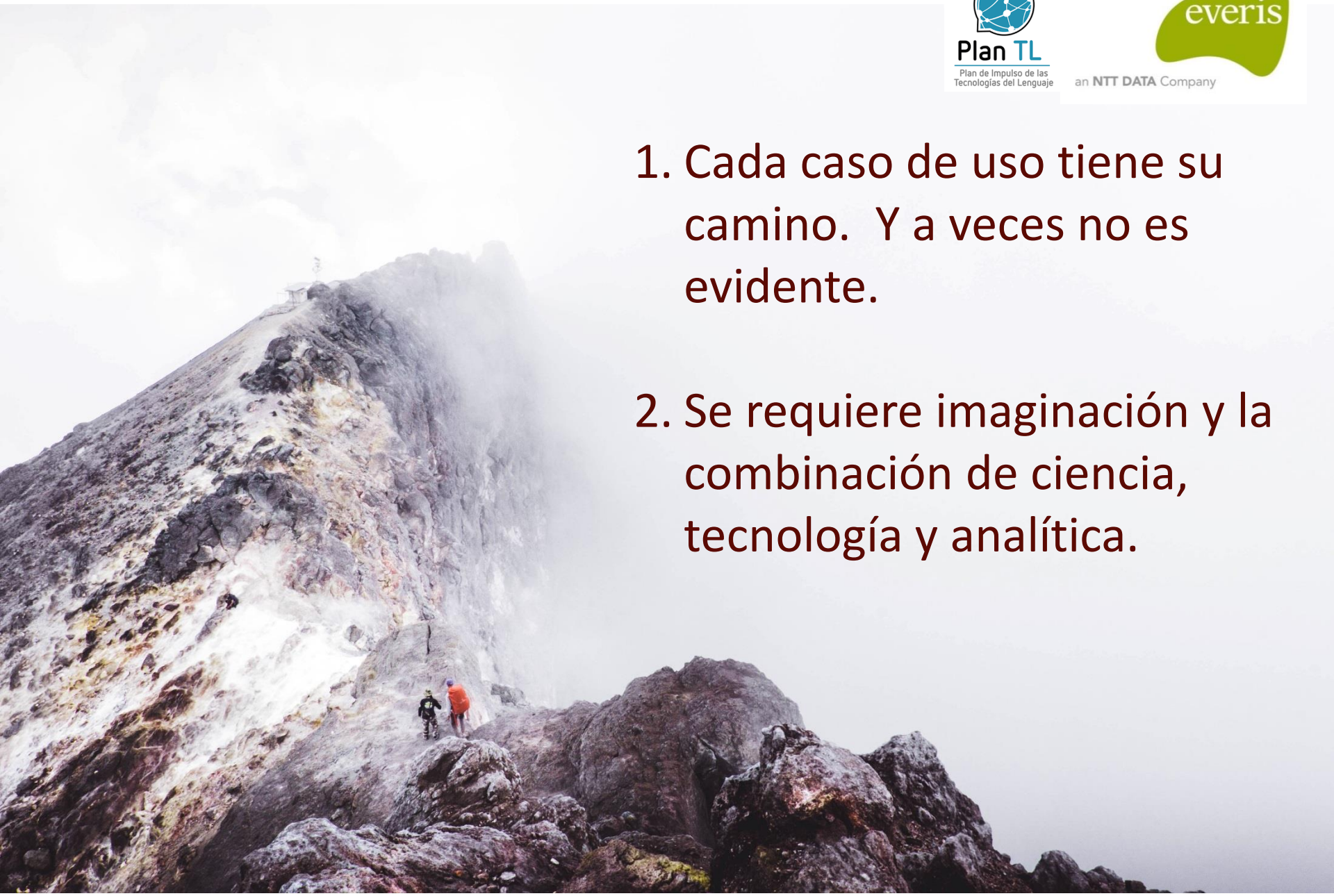
```
A do run run run, a do run run
```
(a) Some text from [7].

```
adorunrunrunadorunrun
```
(b) The text with irrelevant features removed.

```
adoru dorun orunr runru unrun nrunr runru
unrun nruna runad unado nador adoru dorun
orunr runru unrun
```
(c) The sequence of 5-grams derived from the text.

```
77 72 42 17 98 50 17 98 8 88 67 39 77 72 42
17 98
```
(d) A hypothetical sequence of hashes of the 5-grams.

- Scoring global.
- Comparación real–time para el usuario: funcionalidad de subrayado.
- Filtrado de ruido textual.
- Optimización del rendimiento.
- Optimización del balance de los filtros eM Score - longitud
- Optimización adaptada a cada lenguaje.



1. Cada caso de uso tiene su camino. Y a veces no es evidente.

2. Se requiere imaginación y la combinación de ciencia, tecnología y analítica.

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

everis
an NTT DATA Company

GRACIAS

everis MORIARTY
Cognitive Applications Development Platform · powered by ITAINNOVA

Center of Excellence
CADC

everis
an NTT DATA Company

ITAINNOVA
INSTITUTO TECNOLÓGICO DE ARAGÓN