UNIVERSITÉ D'AVIGNON
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Universidad Privada Antenor Orrego
UPAO
...una gran Universidad!

Generación de Resúmen automático de textos

Workshop UPAO 2008

Juan-Manuel Torres

juan-manuel.torres@univ-avignon.fr

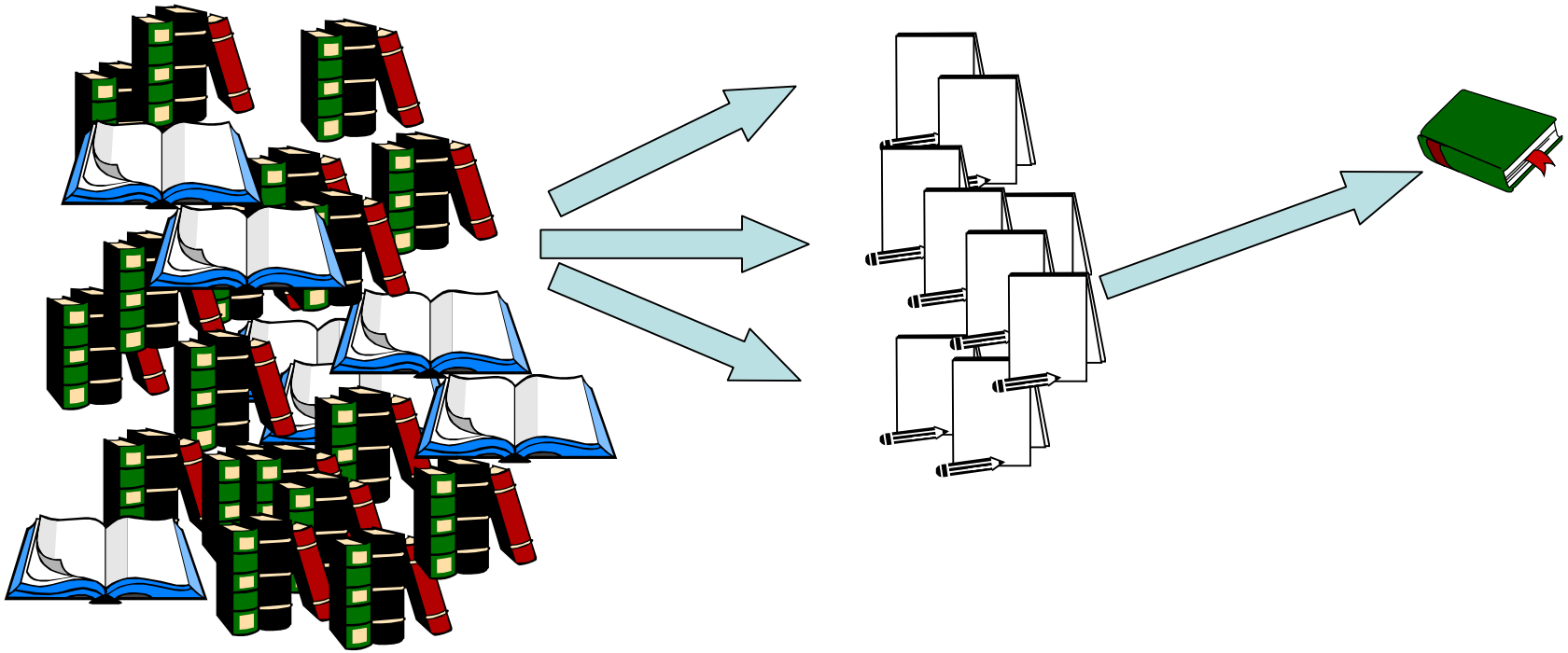
LABORATOIRE INFORMATIQUE D'AVIGNON
Université d'Avignon et des Pays de Vaucluse

Plan

- **Introduccion**
  - Definitiones y resumen humano
  - Estructura de los resumenes
- **Técnicas**
  - Linguisticas
  - Extraccion vs abstraccion
- **Resumen por extraccion**
  - Procesamiento de Lenguaje Natural
- **Sistema Cortex**
  - Algoritmo
  - Pruebas
  - Nuevas tendencias

Un excitante desafío...

Baje 1000 documentos del web, envielos al sistema de resumen y elija los mejores leyendo **los resumenes ordenados por pertinencia**...



Les premières images de Phoenix sur Mars.

La sonde américaine Phoenix a atterri sans encombre dimanche sur Mars et a commencé à livrer ses premières images. Phoenix va tenter d'y sonder le sous-sol. C'est la première sonde à atterrir de manière active (en utilisant des rétrofusées) depuis les deux rover Viking, en 1976. La sonde Phoenix s'est posée avec succès sur la planète Mars, cette nuit à 23h38 (heure française). Des problèmes pendant la descente avaient fait échouer la première tentative d'atterrissage d'un tel vaisseau sur Mars en 1999.

Mots clés : mars[10] phoenix[10] sonder[4] atterrir[3] image[3] vie[3] premier[2] encombrer[2] poser[2]
Extrait généré en 2.895 secondes à partir de 184 phrases, 13 articles (moyenne de 14.153 phr/art)

Score	Sentence
0.359	La sonde américaine Phoenix a atterri sans encombre dimanche sur Mars et a commencé à livrer ses premières images.
0.276	Phoenix va tenter d'y sonder le sous-sol.
0.235	C'est la première sonde à atterrir de manière active (en utilisant des rétrofusées) depuis les deux rover Viking, en 1976.
0.206	La sonde Phoenix s'est posée avec succès sur la planète Mars, cette nuit à 23h38 (heure française).
0.176	Des problèmes pendant la descente avaient fait échouer la première tentative d'atterrissage d'un tel vaisseau sur Mars en 1999.
0.171	Les premières images envoyées par Phoenix montrent que l'appareil a déplié ses panneaux solaires, comme prévu, après que la poussière soit retombée.
0.169	Traces de vie?
0.168	Sur les onze précédentes tentatives de poser des sondes spatiales sur Mars, seules cinq ont réussi.
0.166	Les scientifiques qui contrôlaient le voyage de la sonde spatiale au siège du laboratoire, à Pasadena, près de Los Angeles, se sont félicités de la survie de Phoenix Lander qui a renvoyé des signaux depuis Mars ; et donc ses premières photos.
0.162	La Nasa a utilisé pour réaliser Phoenix des composants conçus pour la sonde Mars Surveyor dont le programme avait été annulé en 2001 suite à l'échec de la mission Mars Polar Lander.
0.161	Lancée le 4 août 2007 sur une fusée Delta 2 de Floride (sud est), Phoenix est la première sonde à s'être posée dans la région arctique de la planète rouge pour y creuser le permafrost.
0.160	« Phoenix est aussi dotée d'instruments qui, en analysant la composition du permafrost martien, sont capables de détecter des molécules de carbone et d'hydrogène, des éléments nécessaires à la vie. La sonde pourra ainsi découvrir d'autres composants chimiques et déterminer ainsi si une forme de vie primitive a été ou est encore possible sur Mars », a déclaré Peter Smith.
	Interrogé sur le processus d'atterrissage de Phoenix Mars Lander, un autre responsable Michael A. Gross l'a qualifié de "six minutes

Titulos – Noticias – Flujo RSS

Resume automatique - Recherche Google - Mozilla Firefox

Fichier Edition Affichage Aller à Marque-pages Outils ?

http://www.google.fr/search?hs=jvS&hl=fr&client=firefox-a&rls=org.mozilla%3Afr%3Aofficial&q=Resume+automatique&btnG=Rechercher&meta=

TALN | Nautitia Gestion des... | Web Term Documen... | SENSUS demo | System Quirk | UNED Group in Natu... | Departamento de In... | Program 1 | SELFHTML: CGI/Perl... | UNIVERSIA-MIT-OP...

Web Images Groupes Annuaire Actualités plus »

Resume automatique | Rechercher | Recherche avancée | Préférences

Rechercher dans : Web Pages francophones Pages : France

Web Résultats **1** - **10** sur un total d'environ **1 250 000** pour **Resume automatique**. (**0,25** secondes)

Pertinence Mining - Traitement de l'information textuelle
... ne nécessite aucune inscription et ne sert que de DEMO de sensibilisation au couplage de deux technologies : **Résumé automatique** + Moteur de recherche ...
www.pertinence.net/ - 15k - En cache - Pages similaires

RALI: **Résumé automatique**
Les activités du RALI dans le domaine du **résumé automatique** s'articulent autour de trois projets principaux:. SumUM pour les articles scientifiques ...
rali.iro.umontreal.ca/Resume/index.fr.html - 7k - En cache - Pages similaires

Résumé automatique de textes
Le **résumé automatique** de textes : solutions et perspectives, J. Minel, - Narrative Summarization, I. Mani, - **Résumé** de textes juridiques par identification ...
www.atala.org/article.php3?id_article=24 - 18k - En cache - Pages similaires

Le **résumé automatique** des textes scientifiques et techni
Le **résumé** de texte **automatique** s'inscrit dans le cadre de l'ingénieri
To produce the RAFI system ('**Résumé Automatique** à Fragments Inc
www.atala.org/article.php3?id_article=208 - 21k - En cache - Pages sir
[Autres résultats, domaine www.atala.org]

Appel a participation : Filtrage et **resume automatique** de l ...
Évaluation, rectification et pertinence du **résumé automatique** de texte pour une utilisation en réseau, Abderrafih LEHMAM. ...
www.lri.fr/Services/Archives-majordomo/interactif/msg00030.html - 16k - En cache - Pages similaires

NetEconomie :: Pertinence.net veut démocratiser le **résumé automatique**
Pertinence.net veut démocratiser le **résumé automatique**. vedette La Jeune pousse Pertinence.net propose un service en ligne de **résumé automatique** des ...
www.neteconomie.com/perl/navig.pl/neteconomie/infos/article/20011004152944 - 25k - 14 sep 2005 - En cache - Pages similaires

sujet stage **résumé automatique**
Résumé automatique par filtrage sélectif Contexte La tâche de **résumé** est avant tout une tâche de compréhension de texte :. dans une première étape il s'agit ...
www.limsi.fr/Individu/mhp/stageDEA/NH-05.html - 3k - 13 sep 2005 - En cache - Pages similaires

Outils Froids - **Resume Automatique**
Outils de **résumé automatique**. De La Cartographie Pour Pertinence Mining · Fourre Tout Numero Quatre · Pertinence **Resume** Les Resultats De Google ...
www.outilsfroids.net/texts/ResumeAutomatique.shtml - 19k - En cache - Pages similaires

GFII - Séminaire "Le **résumé automatique** : état de l'art technique ...
MINEL Jean-Luc, Filtrage sémantique : du **résumé automatique** à la fouille de ...
Pertinence Summarizer : un outil de **résumé automatique** au service de la ...

Terminé

Démarrer | 3 Explorateur ... | 3 OpenOffice.o... | 9 Firefox | ~/Projets/Versio... | PERL - Microsoft ... | Courrier entrant ... | Microsoft PowerP... | 09:37 a.m.

Busqueda de informacion

Neural Processing Letters **16:** 201–210, 2002.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

The Minimum Number of Errors in the *N*-Parity and its Solution with an Incremental Neural Network

J. MANUEL TORRES-MORENO[1★], JULIO C. AGUILAR[2], and MIRTA B. GORDON[3]

[1]*École Polytechnique de Montréal, Département de Génie informatique, CP 6079 Succ. Centre-ville, H3C3A7 Montréal (Québec) Canada. e-mail: juan-manuel.torres@polymtl.ca*

[2]*Laboratorio Nacional de Informática Avanzada (LANIA), Rébsamen 80-91090 Xalapa, México*

[3]*Laboratoire Leibniz – IMAG (CNRS), 46, Avenue Félix Viallet, 38031 Grenoble Cedex, France*

Abstract. The N-dimensional parity problem is frequently a difficult classification task for Neural Networks. We found an expression for the minimum number of errors v_f as function of N for this problem, performed by a perceptron. We verified this quantity experimentally for $N = 1, \ldots, 15$ using an optimal train perceptron. With a constructive approach we solved the full N-dimensional parity problem using a minimal feedforward neural network with a single hidden layer of $h = N$ units.

Key words. classification tasks, minimerror, monoplan, parity problem, perceptrons, supervised learning

1. Introduction

The Neural Networks Community has studied the N-dimensional parity problem for a long time. In their celebrated book, Minsky and Papert [1] elegantly demonstrated that perceptrons are unable to solve non linearly separable problems, such as the parity of 2 inputs (or their equivalent one, the or-exclusive problem, XOR). The capacity of a simple perceptron is limited, since it is unable to solve problems that are linearly separable [6, 7]. The problem still becomes difficult in small dimensions: $N \leqslant 5$ and it is increased exponentially in function of the number of available patterns. This problem has been attacked by several methods such as Gradient Backpropagation (BP) and its variations. These methods have difficulties even in small dimensions due to the problem of the local minima in those in which the minimization of the cost function may fall. An alternative approach is to use Incremental Neural Networks methods that add units while learning errors exist, following a suitable heuristic, as we show it in Section 4.

The N-dimensional parity problem can be formulated as a supervised learning problem with a learning set of P patterns with N binary inputs $\xi_i = \pm 1$, $i = 1, \ldots, N$, and a binary output $\tau = \pm 1$:

$$\mathcal{C} = \{(\xi^\mu, \tau^\mu), \mu = 1, \ldots, P\} \tag{1}$$

Articulos cientificos, Reportes…

rity is difficult
nction is used,
ed for the gra-
is problem has
it it is a highly
ow to discrimi-
or the negative

one, $\tau = -1$. The problem is considered exhaustive learning, because all the $P = 2^N$ different patterns examples must be learned. The N-parity's problem becomes quickly complicated due to the neighboring states in input's space (its Hamming distances are $d_H = 1$) with opposite outputs. A solution with binary poids perceptrons has been should in [22].

The input's space for the N-dimensional parity and the separating hyperplanes $\boldsymbol{w}$ are represented for values of $N = 2, 3, 4$, in Figures 1(a), 1(b) and 2.

2. Finding the Minimum Number of Errors

Following a constructive approach, an incremental Neural Network adds hidden units one to one, until it is capable of eliminating learning errors. In N-dimensional parity, the problem is quite difficult for the first unit because it should find the

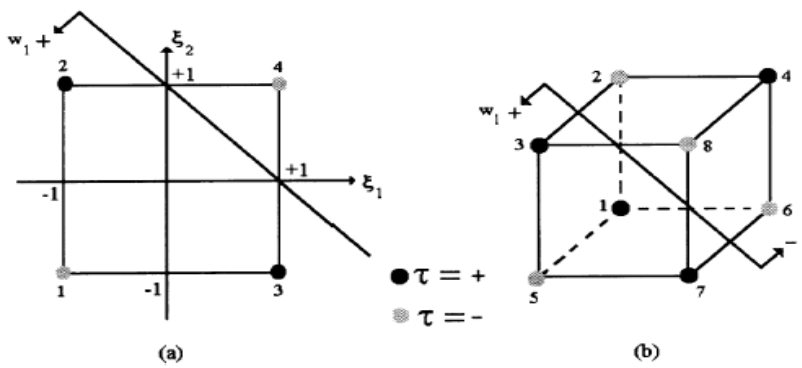


Figure 1. N-dimensional parity: (a) $N = 2$, (b) $N = 3$.

Pertinence Mining.com
Traitement de l'information textuelle

Produits et services en ligne pour le traitement de l'information

PERTINENCE MINING
82, avenue Jean Jaurès
94400 Vitry-sur-Seine, FRANCE
Tél/Fax: +33 (0)1 43 91 45 90
e-mail: contact@pertinence.net

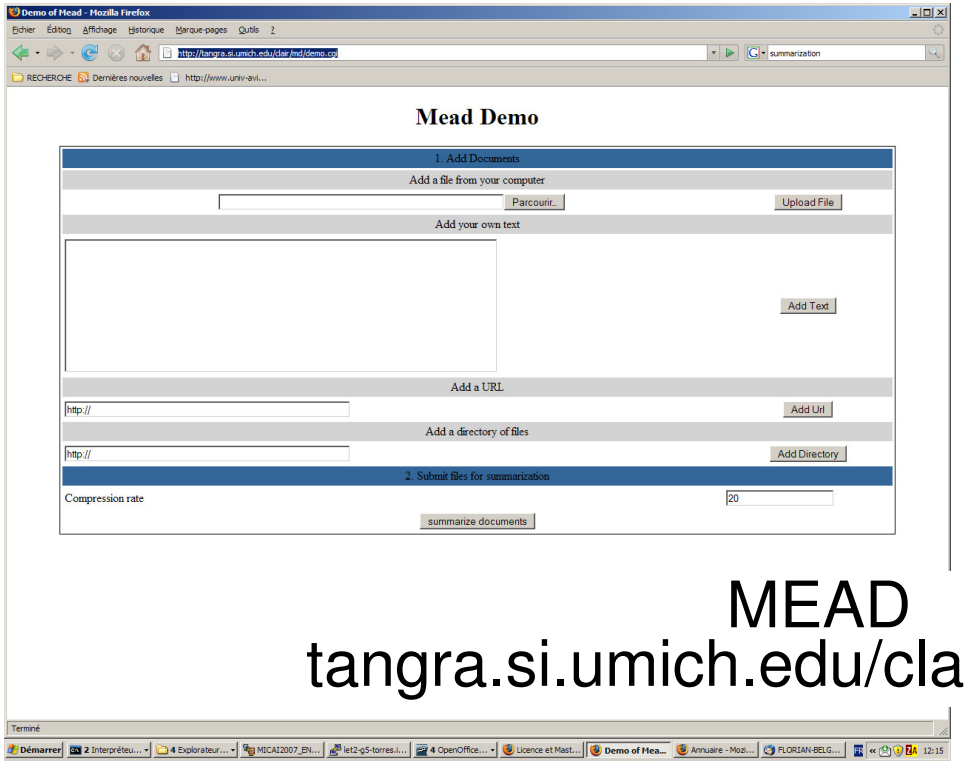
PERTINENCE MINING
SARL
Siret : 442 064 028 00016
Site web : http://www.pertinence.net ou http://www.pertinence-mining.com

www.pertinence.com

FICHE DESCRIPTIVE : PERTINENCE SUMMARIZER

OFFRE :
- Version monoposte,
- Version serveur,
- API Web Service,
- API Java

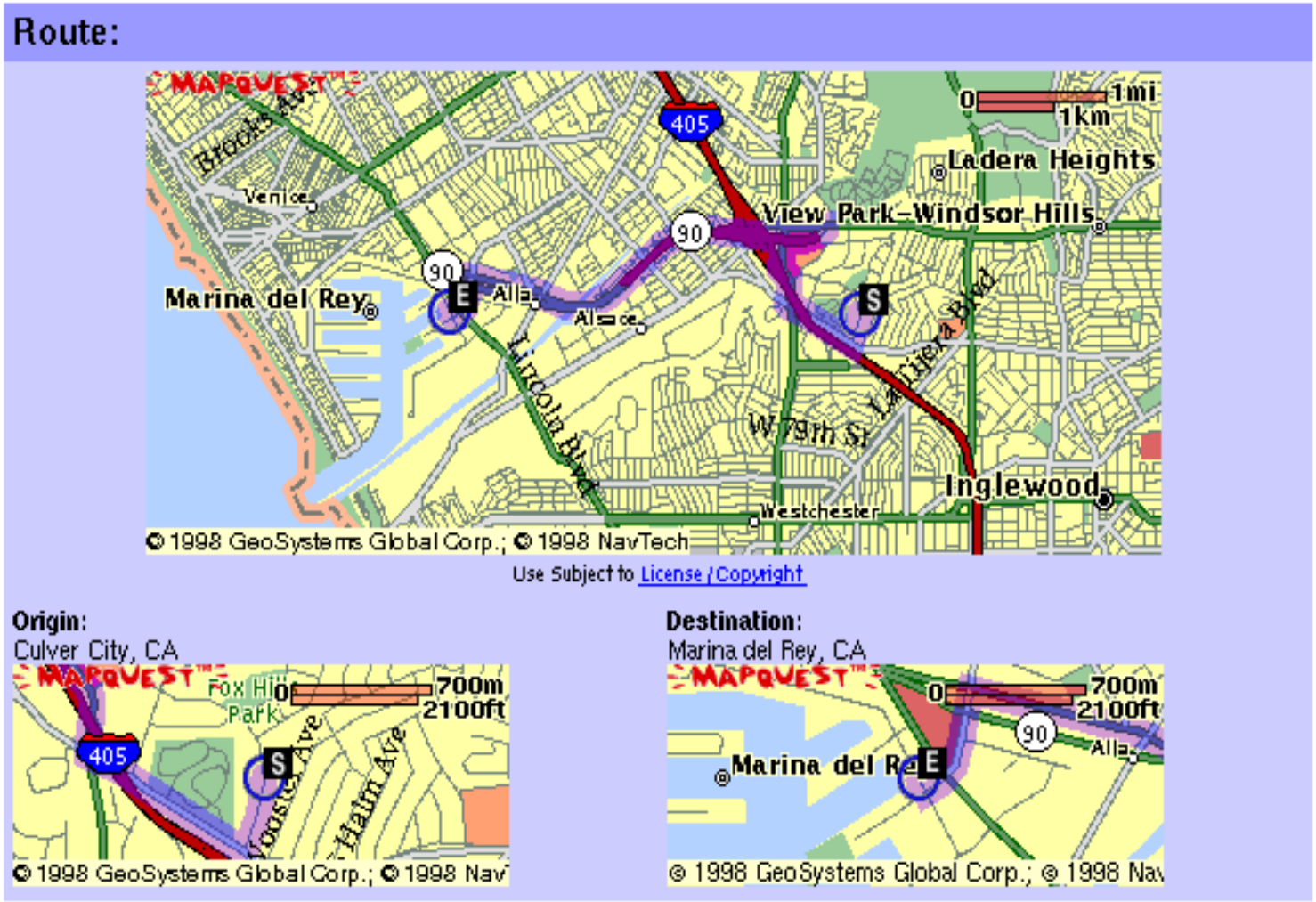
www.copernic.com



MEAD
tangra.si.umich.edu/clair/md/demo.cgi

LIA YACHS Resumenes de **Quimica Organica**
http://daniel.iut.univ-metz.fr/yachs

Mapas graficos — orientacion



GUIAS Cine / télé… — decisiones

2:30am **VC2 – 76**

The Jackal

Movie: Bruce Willis excels as "The Jackal," a cunning assassin who uses many disguises in this 1997 thriller. Richard Gere and Sidney Poitier costar as players from different sides of the law who unite to stop him.

3:00am **KCOP – 13**

The Untouchables

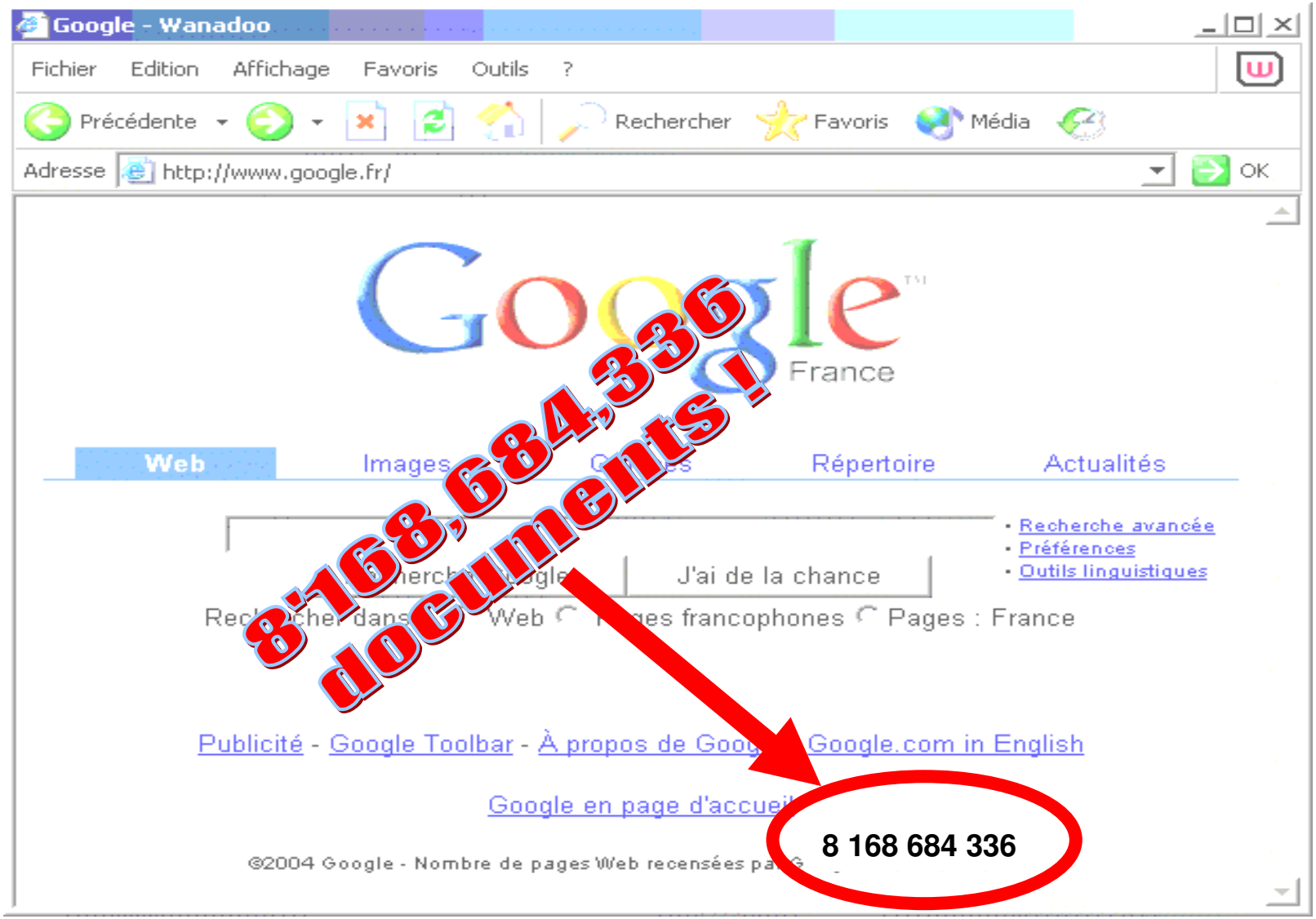
Movie: Eliot Ness (Kevin Costner) and ''The Untouchables'' take on Robert De Niro's flamboyant Al Capone in the pulse–pounding 1987 adaptation of the popular TV series.\Sean Connery won an Oscar as the Irish beat cop who shows Ness ''the Chicago way.''\ Brian De Palma directed the feature;\David Mamet wrote the script.\And yes, film majors, the scene at Union Station was lifted directly from the

3:05am **STARZ – 25**

Grosse Pointe Blank

Movie: A razor–sharp script and a fine turn by John Cusack as a troubled hit man mark 1997's "Grosse Pointe Blank," a dark comedy in which the assassin encounters his old flame (Minnie Driver of "Good Will Hunting") at a high–school reunion. Cusack's sister Joan ("In and Out") is hilarious as the killer's devoted assistant, and Alan Arkin makes the most of his small role as Cusack's terrified the

El GRAN problema…



¿Qué es un resúmen ?

« Transformación reducida de un texto fuente hacia un resumen por compresión del contenido por medio de una selección y/o generalización de lo que es importante en el texto fuente »

Karen Sparck Jones

- ***Representación abreviada y precisa del contenido de un documento***

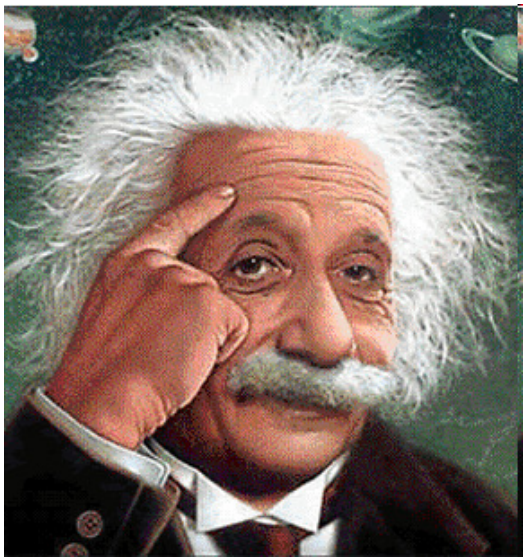
Preguntas

- ¿ Qué **tipo** de resúmen buscan las personas ?
- ¿ Cuál debe ser el **grado de sofisticación** de los sistemas de resúmen ?
  - ¿ Los métodos estadísticos son suficientes ?
  - ¿ Necesidad de métodos simbólicos profundos para "comprender" ?
- ¿ Cómo **medir la calidad** de un resúmen ?

Características del resúmen

- **Indicativo** / **Informativo**
  - Indicativo : indica la naturaleza del texto
  - Informativo : Intenta substituir al texto
- **Extracto** o **Resumen**
  - **Extracto** : frases extraidos del texto
  - **Resúmen** : reformula, comprime el texto
- **Tamaño**
  - Tasa de compresión = |Resumen| / |Texto|

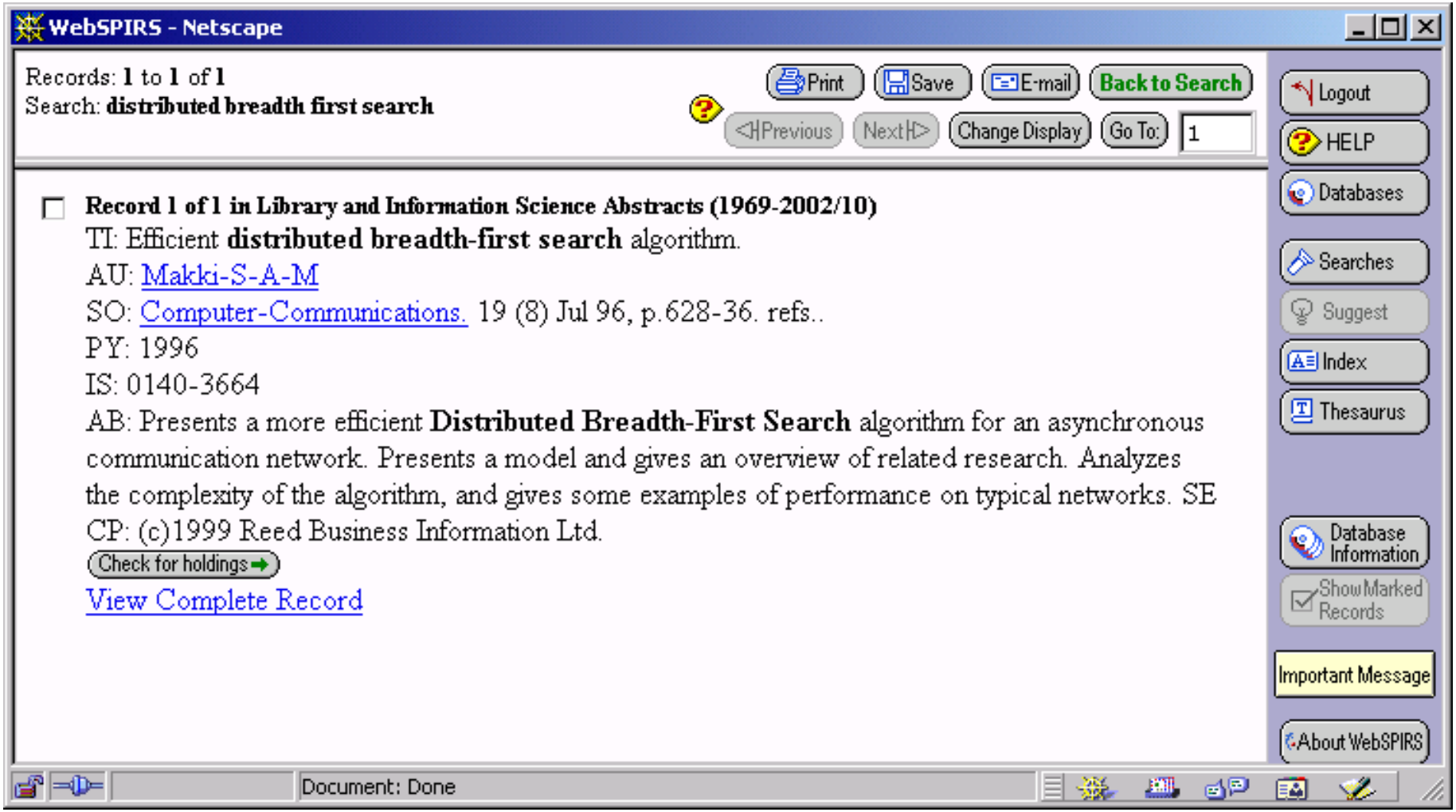
¿ Cómo hacen los humanos para producir resumenes… ?



¿ Quién hace resumenes ?

- **Autor**
- **Experto**
  - Actualizado en el dominio de los documentos a analizar
- **Profesional en resumenes**
  - Analiza los documentos y la redacción de los resumenes
  - Experto en **esta** disciplina
  - Normas de producción de resumenes

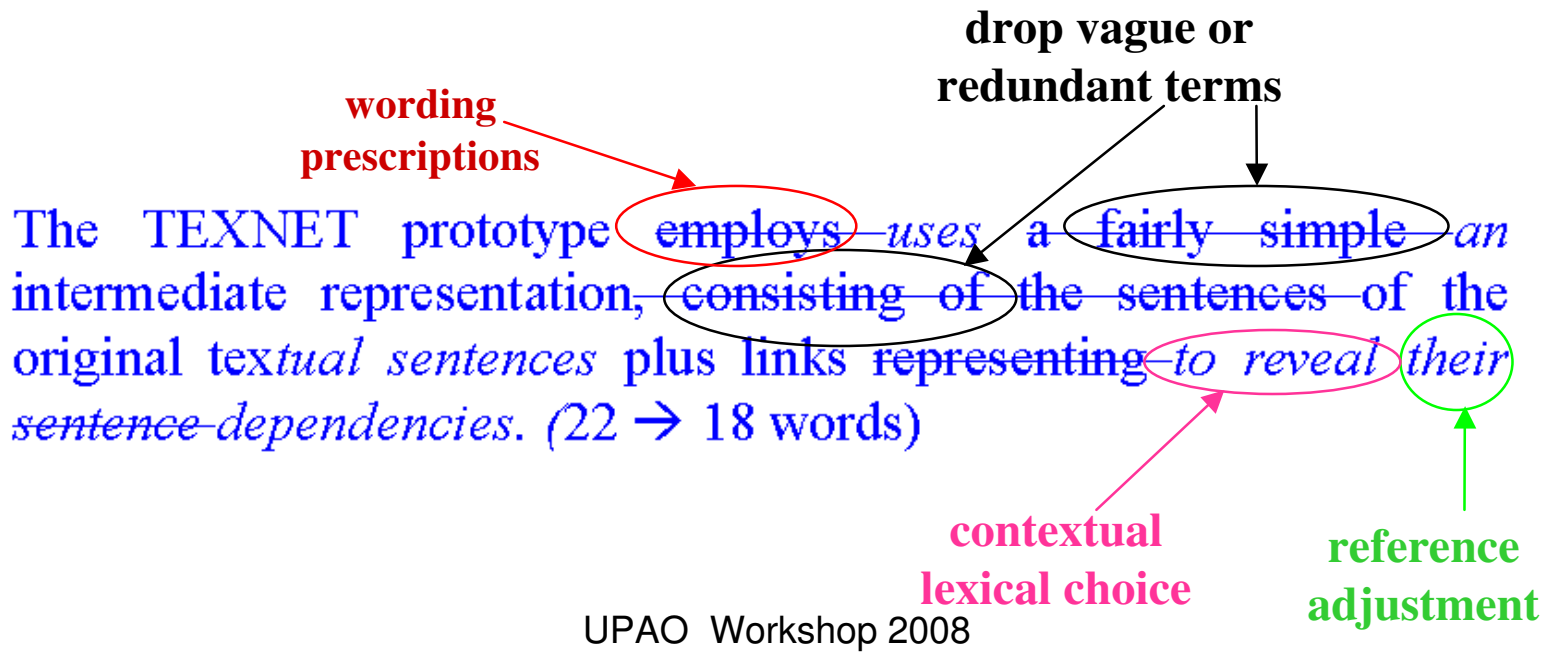
Resúmen profesional



Cremmins'82-'96: operaciones de edición

Operaciones de edición utilizadas

- **Revisión local : contenido en una frase**
- **Revisión global : contenido a través de las frases**



Tiempo de generación del resúmen

Cremmins recomienda entre 12-20 min para resumir un artículo científico típico…

¡Un tiempo mucho menor del necesario para realmente *comprenderlo* !

Endres-Niggemeyer et al. (1995, 1998)

- Los resumidores no leen ***jamás el documento completo***
- **Estructura** del documento
  - Formato
  - Disposición
  - Esquemas
- Nivel de **discurso** : de qué habla el documento

Resumidores profesionales

- Producen buenos resumenes (***savoir-faire***)
- **Indicaciones** de agencias
- **Normas** internacionales
- **Selección** de la información
  - Títulos/Encabezados/Leyendas
  - Introducción/Conclusión
  - Objetivo, metodología, resultados
  - …
- **Redacción** del resumen
  - Re-utilización de frases
  - Transformacion y aplicación de patrones de redacción

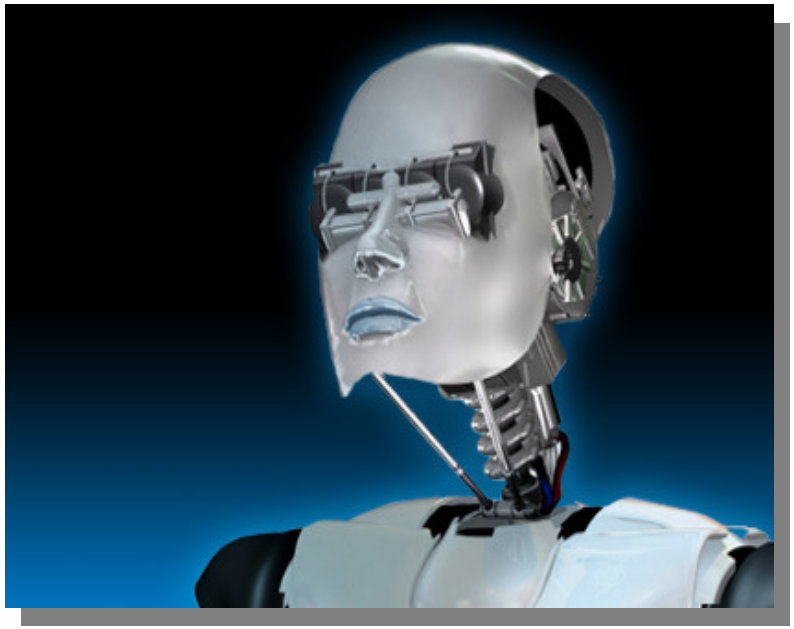
Transformaciones

- Re-expresión de conceptos
- Eliminación
  - Estructuras
  - Frases
  - Texto entre paréntesis
- Expansión de abreviaturas
- Abreviaturas
- Pegar, Separar
- **Ninguna**

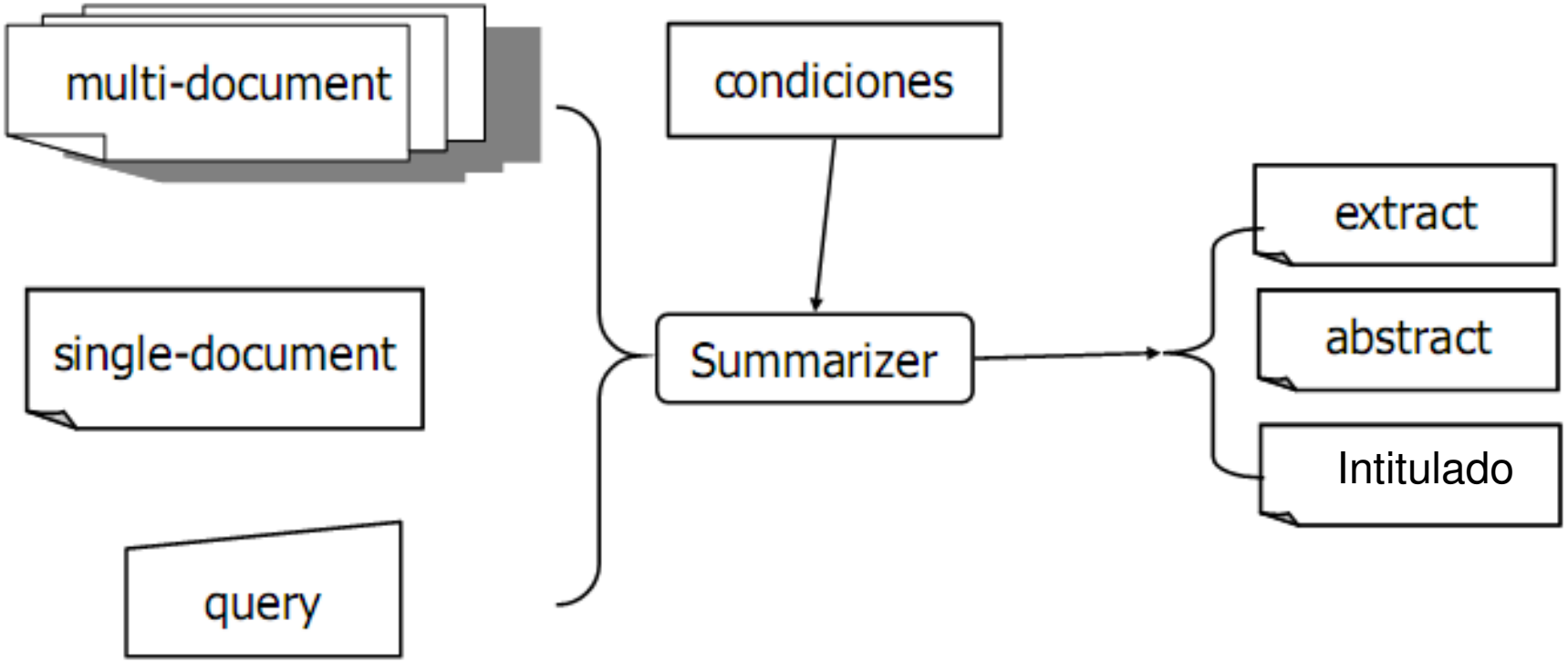
Observaciones

- **70% de las frases son copiadas literalmente** (Lin & Hovy 2003)
- ¿ La gente **sabe escribir** correctamente y **sintetizar** ideas ?
- Los resúmenes del autor ¿ son **pertinentes** ?
- Los resúmenes profesionales ¿ son **realmente** pertinentes ?

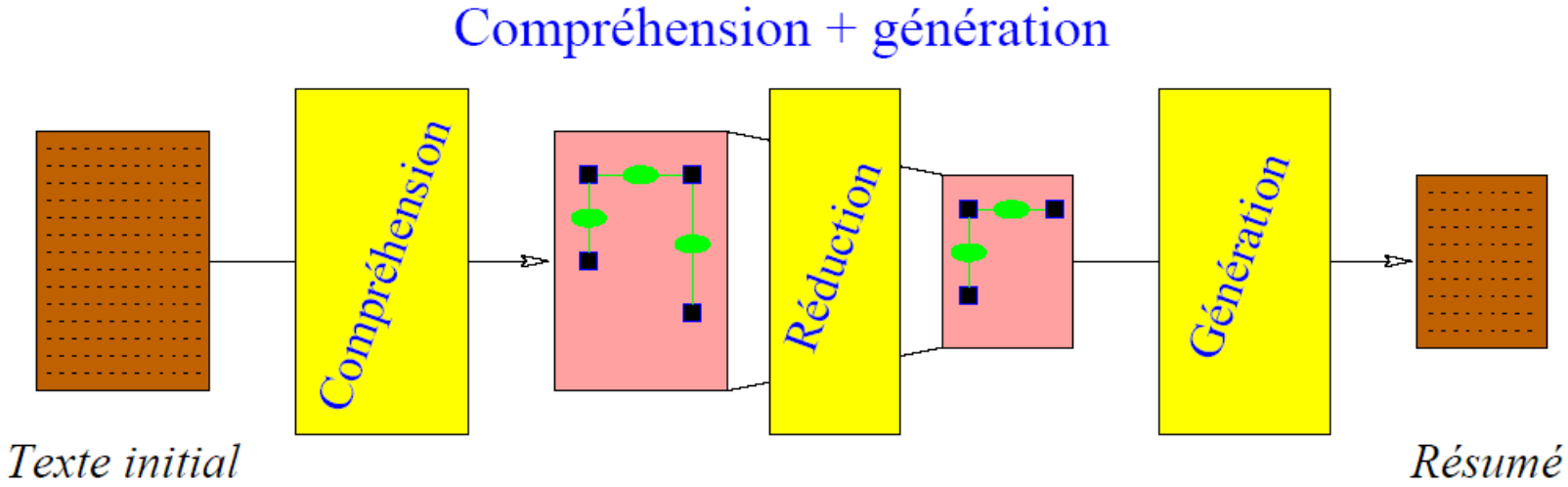
¿ Cómo hacen los sistemas automáticos para producir resumenes… ?



Esquema basico



El método del buen alumno



- Necesita conocimientos del dominio (tematica)
- Necesita conocimientos de la lengua en la cual se genera el resumen…

Resúmen por comprension : FRUMP

a small eartquake shook several Southern Illinois counties Monday night, the National Earthquake Information Service in Golden, Colo., reported. Spokesman Don Finley said the quake measured 3.2 on the Richter scale, "probably not enough to do any damage or cause any injuries." The quake occurred iabout 7:48 p.m. CST and was centered about 30 miles east of Mount Vernon, Finlay said. It was felt in Richland, Clay, Jasper, Effington, and Marion Counties.

There was an earthquake in Illinois with a 3.2 richter scale. (RESUME)

FRUMP

- Estructura de base : *sketchy-scripts* adaptación de scripts de Shank & Abelson (1977)
- Scripts contienen información importante de un evento
- 50 sketchy-scripts producidos manualmente Interpretación de un texto : lectura rapida (skimming)

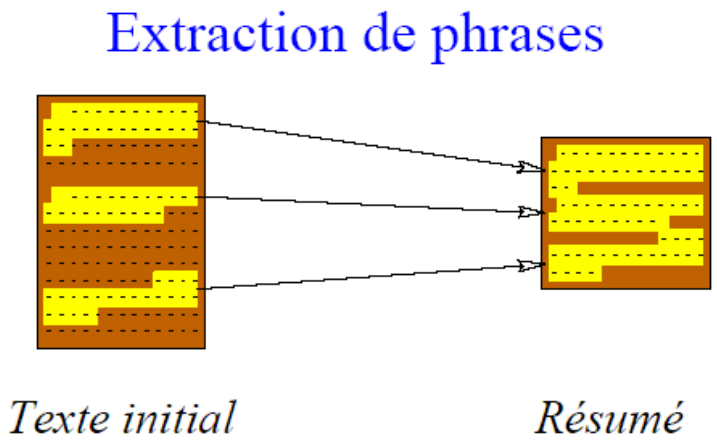
FRUMP : los contras …

- 50 scripts son muy **pocos** para interpretar el mundo
- Conocimientos ***codificados*** **manualmente**
- ¿ Cómo aprender scripts de nuevos temas ?

Ciudad del Vaticano. La noticia de la muerte del Papa *sacude* el mundo. Murió el martes pasado de forma misteriosa… (TEXTO)

Sismo en el Vaticano : un muerto. (RESUMEN)

El método del alumno flojo



Identificar frases importantes:

- Por su posicion
- Por su frecuencia de palabras
- Por sus conectores retoricos
- Por su similitud con otras frases
- …

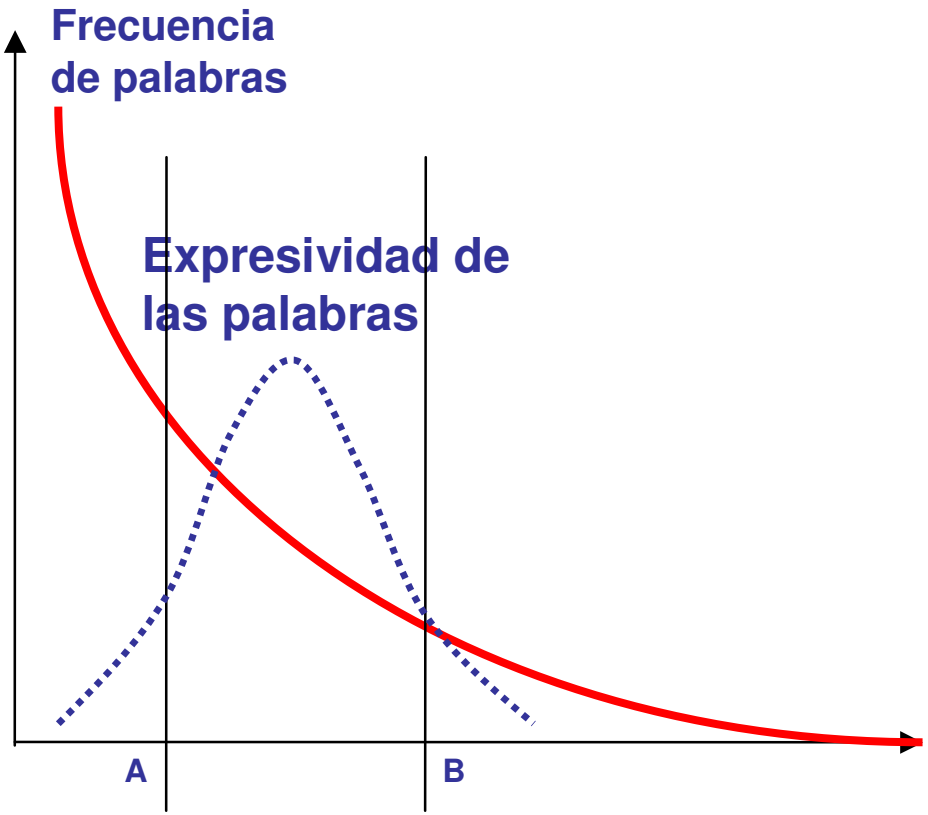
Resúmen por extracción de frases

+ Simple a programar
+ Robusto
+ Rápido

- ¿ Cómo descubrir qué tipo de información lingüistica/semántica es relevante para el contenido ?
- ¿ Coherencia ?
- ¿ Cohesión ?

Enfoque de la RI (Luhn'58-59)

- Extracto = frases significativas
- Una frase significativa contiene palabras significativas (palabras-claves)
- Palabras significativas entre A y B



(Luhn, 59)

Tabla de frecuencias

theater	3.0
pretenderlo	
codazzi	2.0
preteto	1.0
tierreno	1.0
azliun	1.0
cviles	1.0
contemporaneas	1.0
prosirio	**9.0**
política	1.0
civiltá	1.0
linfoma	3.0
...	
el	**52447**

$$w_{i,j} = \mathrm{tf}_{ij} \cdot log_2 \frac{N}{n}$$

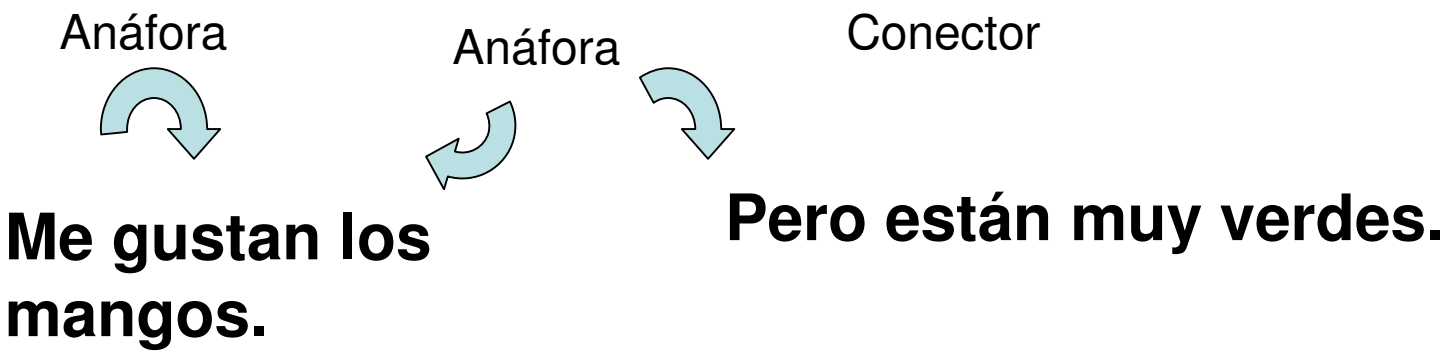
No. DE FRASES = 60644

idf (el) = log(60644/52447) = 0.063

idf (prosirio) = log(60644/9) = 3.82

Problemas : cohesión

- Extractos de frase por frase pueden ser **incoherentes** y difíciles a leer



  - Solución: adicionar frases necesarias para producir un pasaje adecuado
- ¿ existen partes sobre-representadas ?
- ¿ hay partes que se olvidaron ?

Problemas : cohesión

La cadena de supermercados Metro anunció que los negocios van muy bien. El mes pasado las ventas aumentaron 10% y se estudia la creacion de nuevos puestos. Por otro lado, la compania de supermercados Carrefour anunció perdidas de 30% al ultimo trimestre. El directorio se reunio hoy para decidir la posibilidad de reducir su personal.(CORPUS)

La cadena de supermercados Metro anunció que los negocios van muy bien. El directorio se reunio hoy para decidir la posibilidad de reducir su personal.

(EXTRACT)

Dificultad de la evaluación

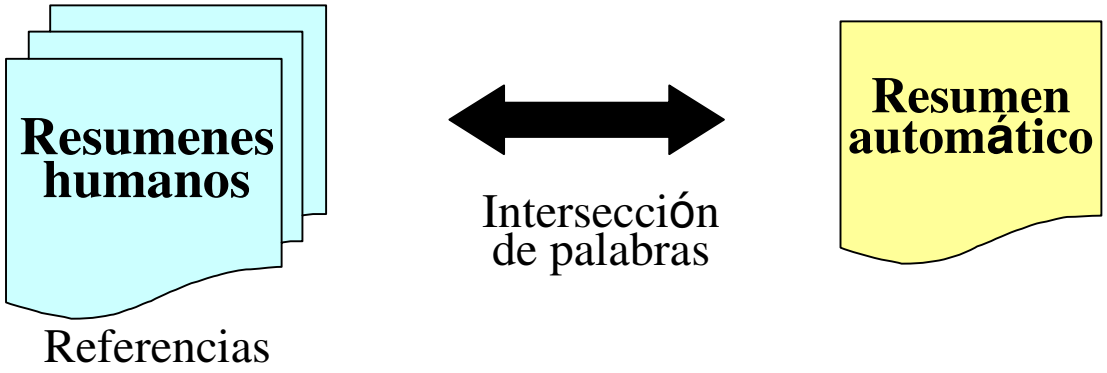
- NO EXISTE UN ÚNICO RESUMEN válido para un texto
- Lenguaje natural producido por una máquina
- Personas juzgando incrementa el coste
- Resumir conlleva compresión (reducción de tamaño). Es necesario evaluar resumenes de distintos tamaños
- Legibilidad (puede no tener relación con la calidad del resumen)
- En relación con las expectativas de la tarea

Evaluacion

- Intrínsecos
  - Calidad (por personas - No siempre acuerdo)
    - Legibilidad, comprensión, acrónimos, anáforas, integridad de la estructura, gramaticalidad, estilo impersonal, ...
  - Informatividad
    - Información preserva respecto al texto original (varias compresiones)
    - Información contiene respeto a un **resumen ideal**

ROUGE

(Lin, ROUGE DUC-NIST, 2004)



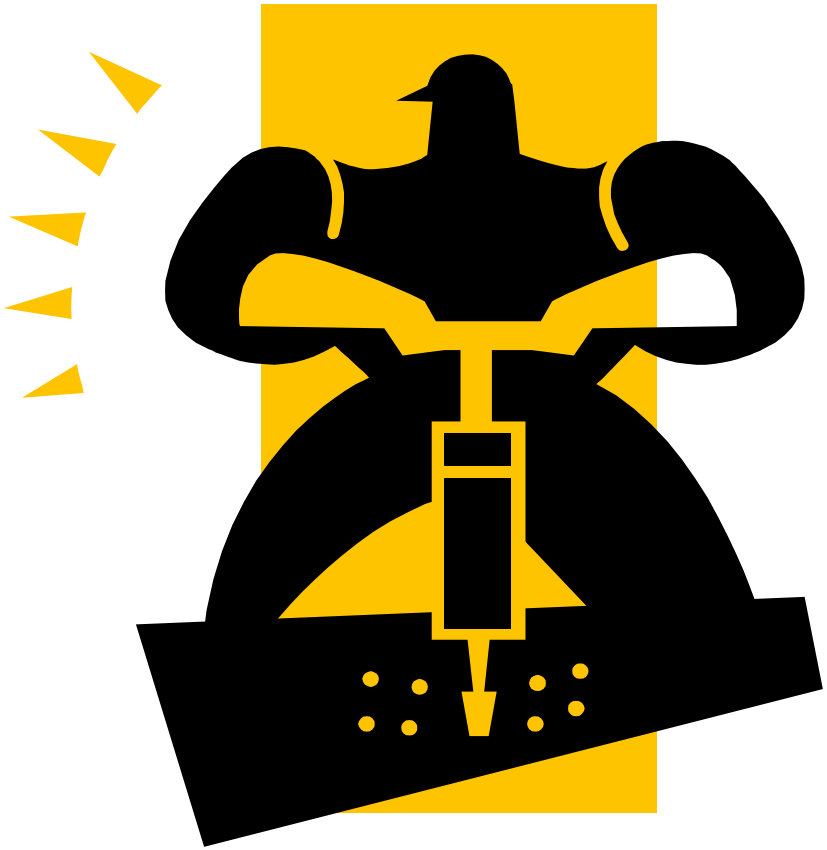
ROUGE-2 : Interseccion de bigramas
ROUGE-SU4 : Interseccion de bigramas separados por un hueco de hasta 4 palabras

Universidad Privada Antenor Orrego

ROUGE-2 : Universidad Privada, Privada Antenor, Antenor Orrego

ROUGE-SU4 : Universidad Privada, Universidad Antenor, Universidad Orrego, Privada Antenor, Privada Orrego, Antenor Orrego

Y ahora, un sistema real...



CORTEX es **O**tro **R**esumidor de **TEX**tos

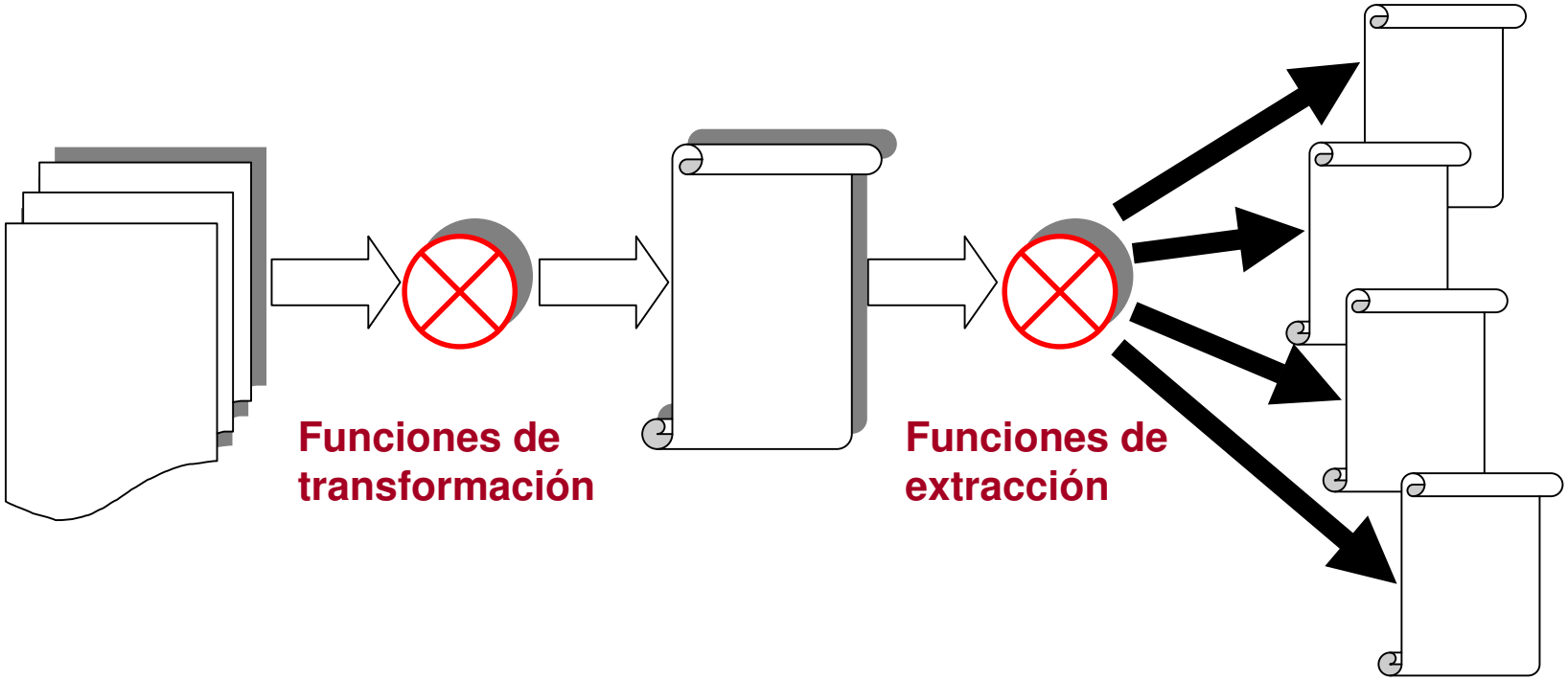
Juan-Manuel Torres-Moreno
juan-manuel.torres@univ-avignon.fr
LIA/Université d'Avignon

Sistema CORTEX

Tres niveles : (Torres, Meunier, Velazquez, 2002-2004)

1. **Módulos de transformación** : Segmentacion, Filtrado, lematización, sinónimos, anáforas, ...
2. **Módulos de extracción** : Métricas, algoritmo de decisión, resumen personalizado …
3. **Módulos de generación de resumenes**: Identificación de conceptos, generación por templates, …

Una cadena de funciones



LATAO : **Lecture et Analyse de Textes Assistés par Ordinateur** (Lectura y Analisis de Textos Asistido por Computadora)

Unidades de información

« La plumme de ma tante »

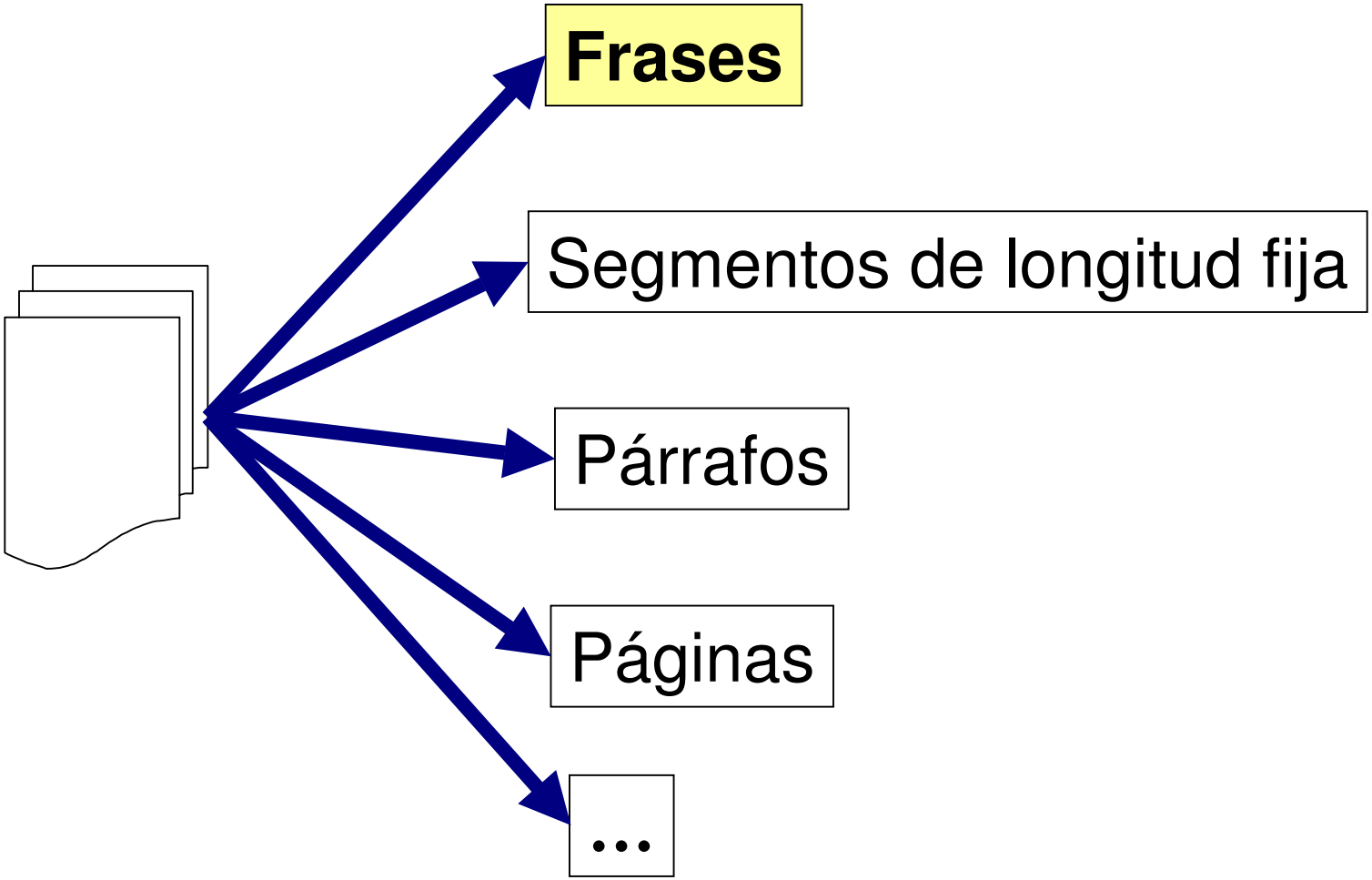
3-gramas

Palabras

```
{la, plumme, de, ma, tante}
```

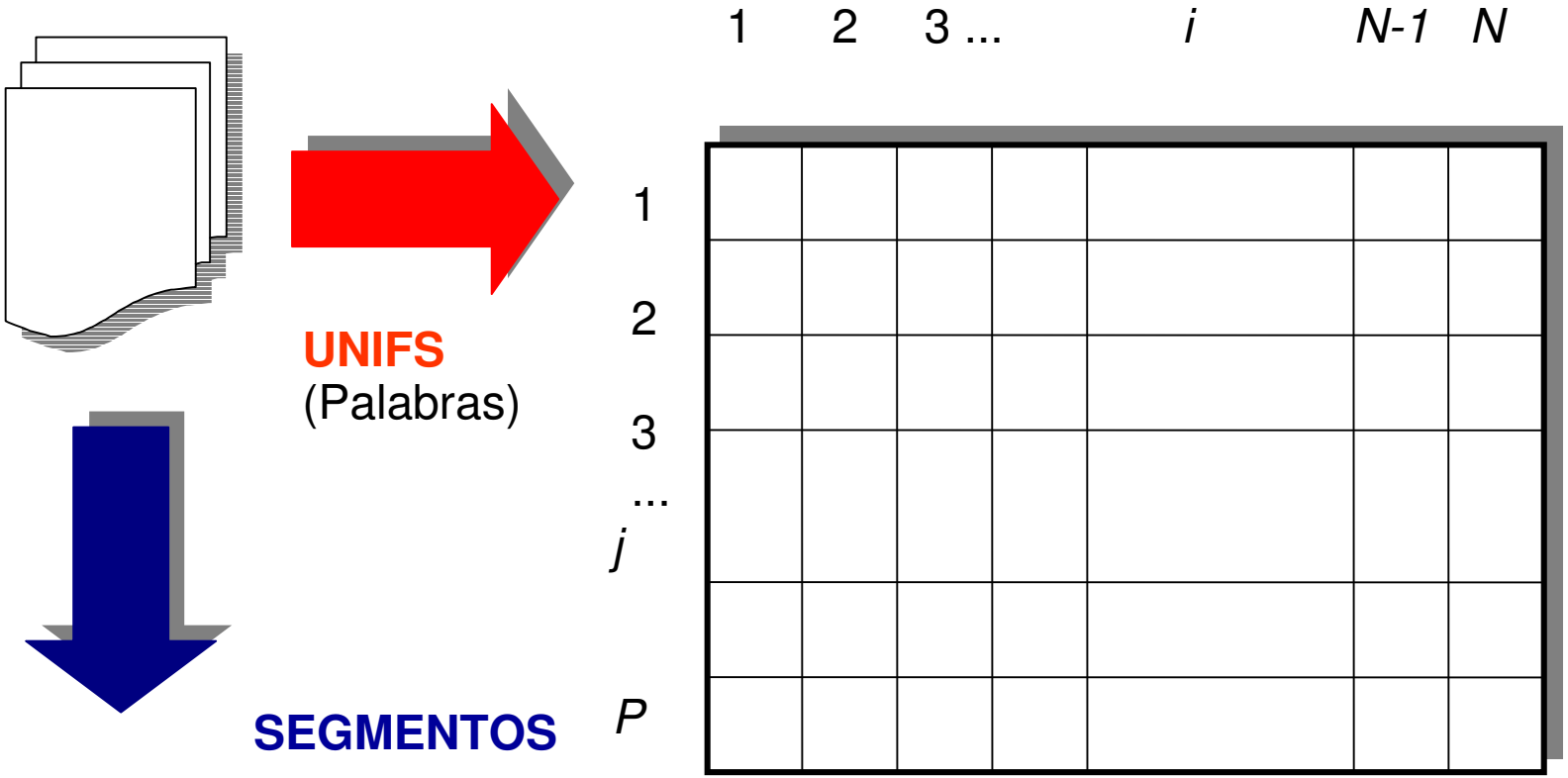
```
{la_, a_p, _pl, plu, lum, umm, mme, me_, e_d,
 _de ,de_, e_m, ma_, a_t, _ta, tan, ant, nte}
```

Unidades de segmentación



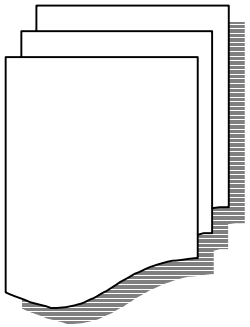
La vectorización de un texto (1)

La vectorización de un texto (2)



MATRIZ Término-Segmento

La vectorización de un texto (3)



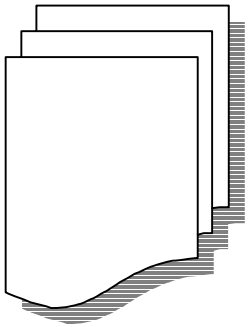
UNIFS
(Palabras)

SEGMENTOS

	1	2	3 ...		*i*	*N-1*	*N*
1	1	1	1	0		0	0
2	0	1	0	0		0	1
3	0	0	0	1		1	0
...							
j							
	1	1	0	1		0	0
P	0	0	0	0		1	0

MATRIZ Término-Segmento

La vectorización de un texto (4)



UNIFS
(Palabras)

SEGMENTOS

	1	2	3		*i*	*N-1*	*N*
1	tf	tf	tf	0		0	0
2	0	tf	0	0		0	tf
3	0	0	0	tf		tf	0
....							
j							
	tf	tf	0	tf		0	0
P	0	0	0	0		tf	0

MATRZ Término-Segmento

Matrices término-segmento

Frecuencia γ

$$\gamma = \begin{bmatrix} \gamma_1^1 & \gamma_2^1 & \gamma_3^1 & \cdots & \gamma_i^1 & \cdots & \gamma_{N_L}^1 \\ \gamma_1^2 & \gamma_2^2 & \gamma_3^2 & \cdots & \gamma_i^2 & \cdots & \gamma_{N_L}^2 \\ \gamma_1^3 & \gamma_2^3 & \gamma_3^3 & \cdots & \gamma_i^3 & \cdots & \gamma_{N_L}^3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \gamma_1^\mu & \gamma_2^\mu & \gamma_3^\mu & \cdots & \gamma_i^\mu & \cdots & \gamma_{N_L}^\mu \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \gamma_1^{N_P} & \gamma_2^{N_P} & \gamma_3^{N_P} & \cdots & \gamma_i^{N_P} & \cdots & \gamma_{N_L}^{N_P} \end{bmatrix}, \quad \gamma_i^\mu \in \{0, 1, 2, \ldots\} \quad (2)$$

Binaria ξ

$$\xi_i^\mu = \left\{ \begin{array}{ll} 1 & \text{si } \gamma_i^\mu \neq 0 \\ 0 & \text{si } \gamma_i^\mu = 0 \end{array} \right\}$$

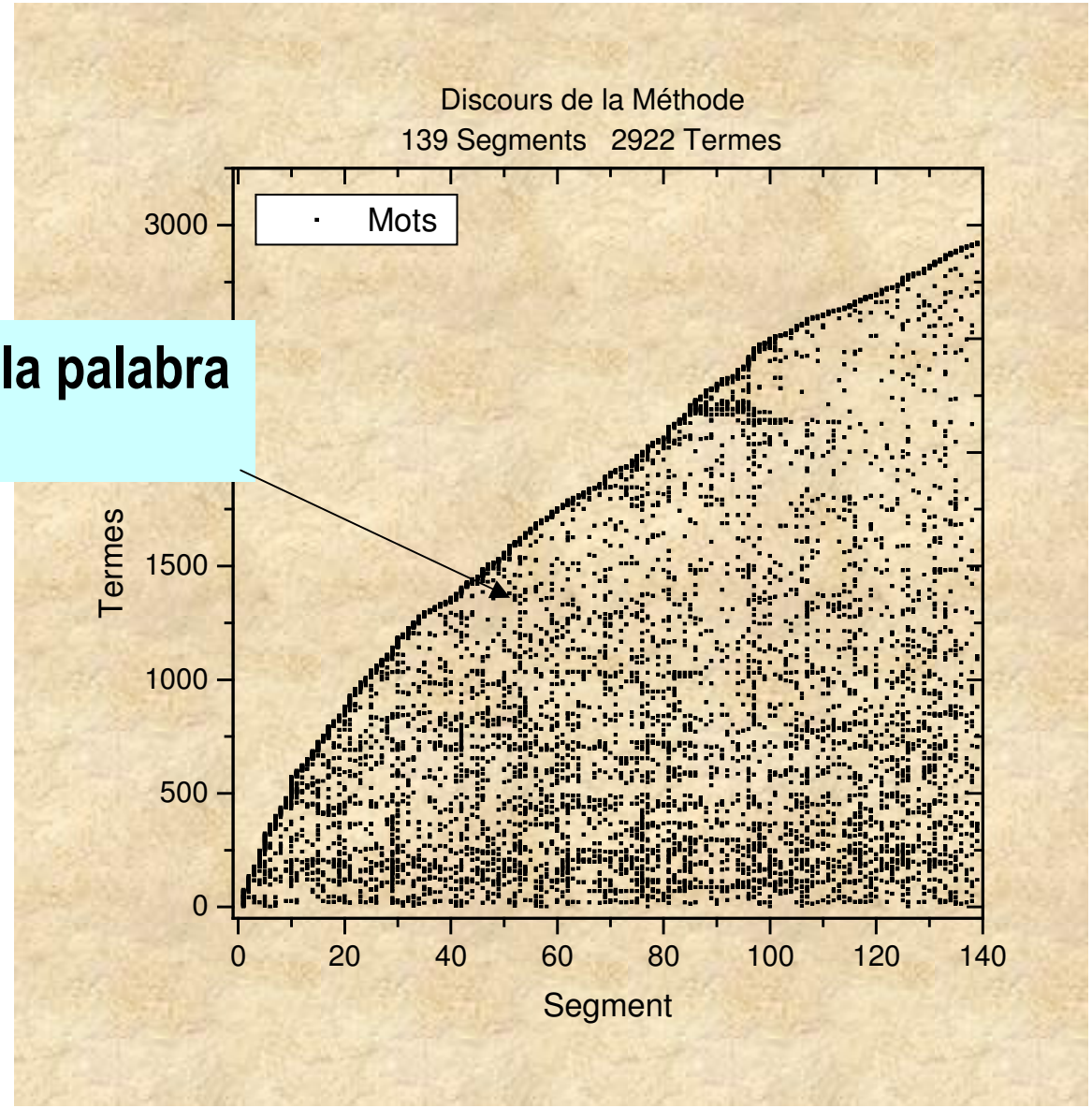
« Huella digital » de un texto

- ¿ Cómo utilizar la información que contienen las matrices del texto?

- ¿ Se pueden visualizar las características propias de un texto ?

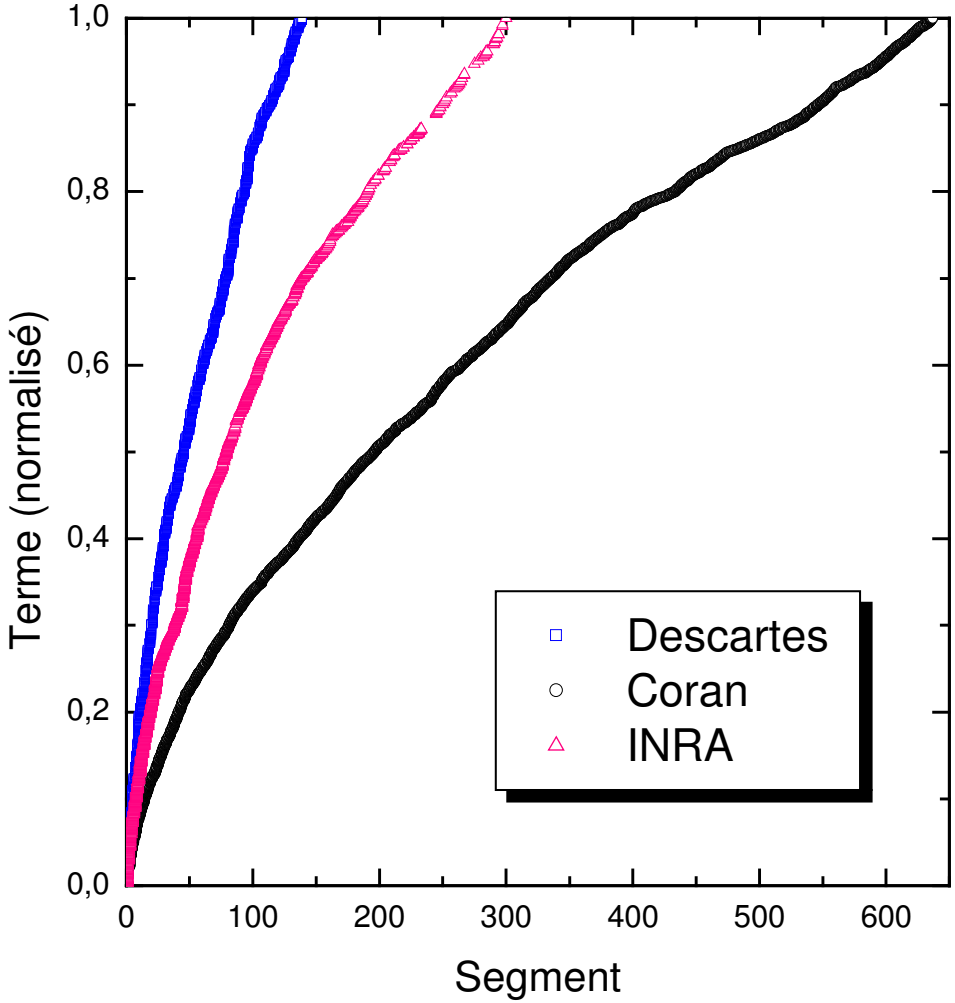
Matriz Termino-Segmento

presencia de la palabra *i* en la frase *j*

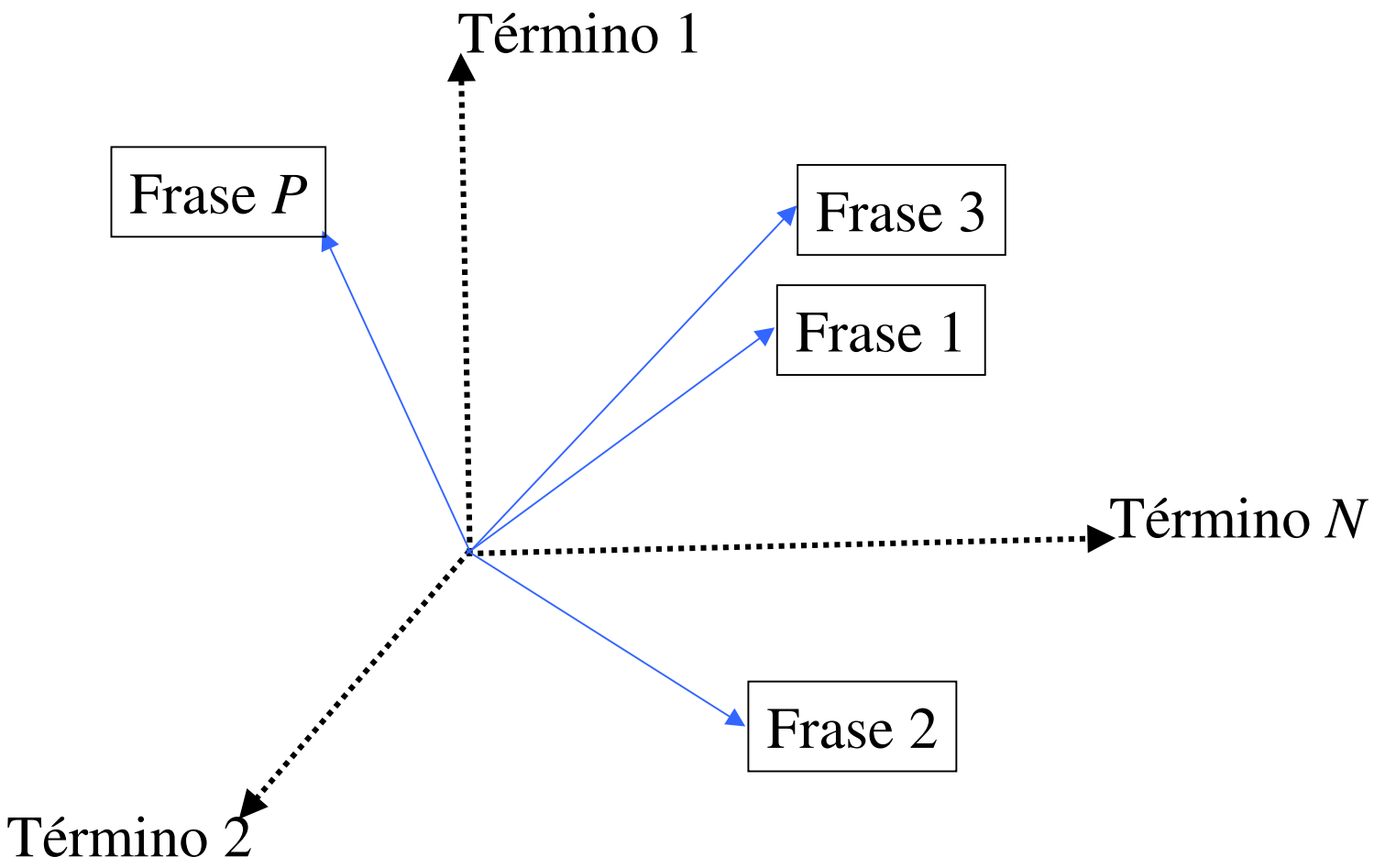


¿ Identificación de estilos ?

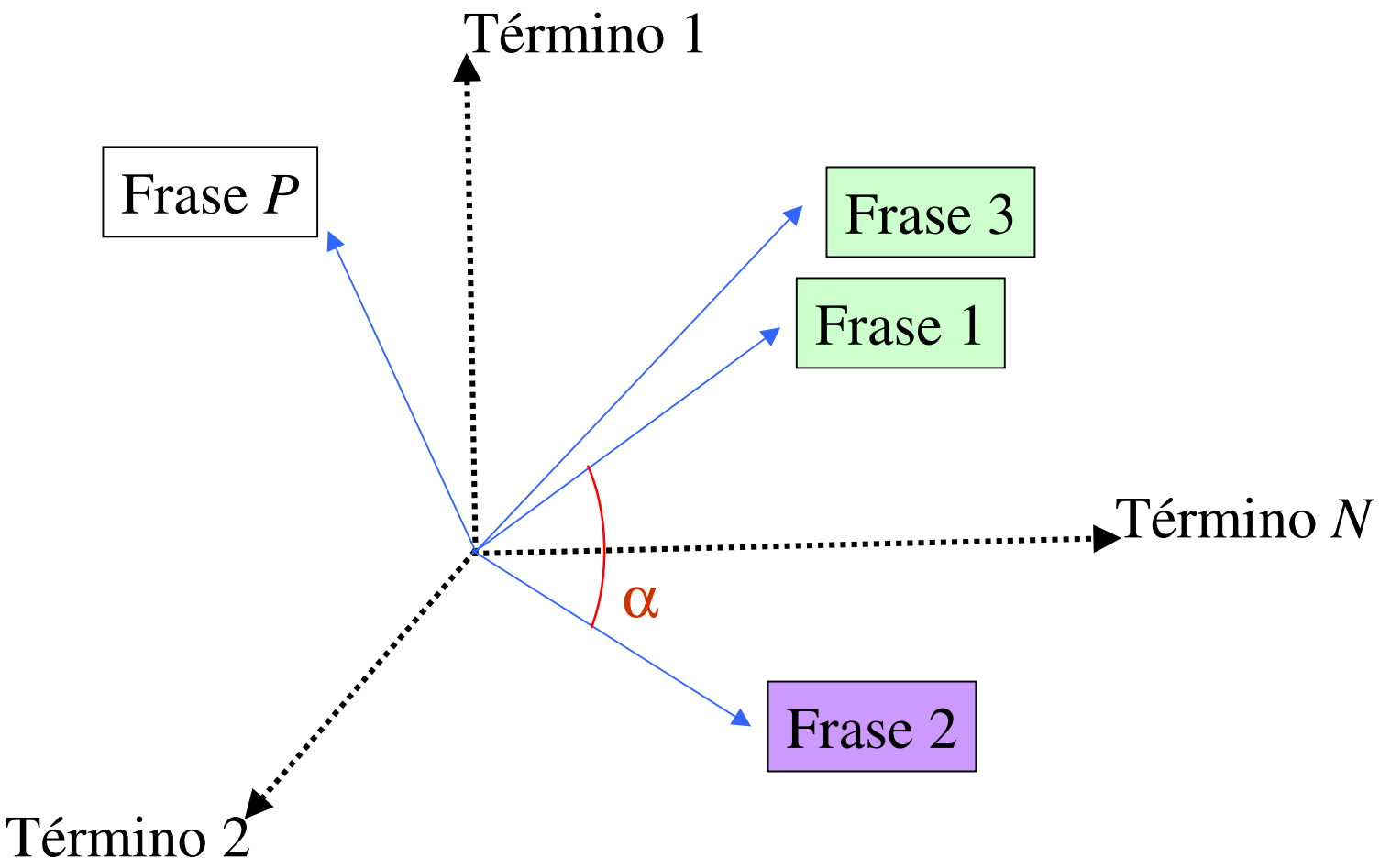
Riqueza
del léxico



Modelo vectorial



Modelo vectorial



Ejemplo

- Texto « Puces»
  - Invasión de “puces” (pulgas) y de “poux” (piojos) : Puces Bio

- Pequeño comunicado en Internet

Invasion de puces et de poux - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.admin.ch/cp/f/1997Sep10.064053.8237@idz.bfi.admin.ch.html Go

Google Search Web Search Site PageRank Page Info Up Highlight

CONFOEDERATIO HELVETICA
Les autorités fédérales de la Confédération suisse

Page d'accueil
Courrier
Recherche

Archives des communiqués
septembre 1997

deutsch
italiano

Resumen autor

Invasion de puces et de poux

Communiqué aux agences de presse cu 9 septembre 1997

Avenches: Invasion de puces et de poux à l'école de recrues d'infanterie d'exploration et de transmission 213

Les cantonnements de la compagnie IV de l'école de recrues d'infanterie d'exploration et de transmission 213, stationnée à Avenches, sont envahis par les puces et les poux. Des piqûres de puces ont été relevées sur plus d'un tiers des militaires. On a aussi retrouvé des cadavres de poux sur 3 militaires. Des mesures d'urgence ont été prises en conséquence.

Des piqûres de puces ont été diagnostiquées sur plus d'un tiers des 155 hommes de la compagnie IV de l'école de recrues d'infanterie d'exploration et de transmission 213. Des cadavres de poux, mais aucun oeuf, ont également été décelés sur 3 militaires. Ces insectes sont transmis par contact personnel. La cause de cette invasion n'est pas claire; ces insectes semblent toutefois avoir essaimé à partir du local de garde.

Le médecin de troupe a donné immédiatement les soins nécessaires aux militaires concernés et il a ordonné les mesures d'hygiène qui s'imposaient. Des produits spéciaux ont été remis pour les soins corporels. Tout le matériel personnel de la compagnie a été emballé hermé-tiquement et apporté à l'arsenal cantonal de Fribourg. La troupe sera déplacée dans un complexe industriel. Une section d'hygiène de l'école de recrues d'hôpital 268, stationnée à Moudon, va désinfecter tous ces cantonnements. On estime qu'avec ces mesures sanitaires appropriées la troupe pourra réintégrer ses cantonnements vendredi au plus tard.

Renseignements complémentaires: Colonel EMG Peter Meier, commandant de l'école de recrues d'infanterie d'exploration et de transmission 213 Téléphone 026-350 82 00

Corpus del texto

Done Internet

Start Voyages-sn... Invasion ... 3½ Floppy (... poster arco2001_p... Microsoft P... EN 20:27

Invasion de puces et de poux - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.admin.ch/cp/f/1997Sep10.064053.8237@idz.bfi.admin.ch.html

CONFOEDERATIO HELVETICA
Les autorités fédérales de la Confédération suisse

Page d'accueil
Courrier
Recherche

Archives des communiqués
septembre 1997

deutsch
italiano

Resumen autor

Resumen Cortex

Invasion de puces et de poux

Communiqué aux agences de presse cu 9 septembre 1997

Avenches: Invasion de puces et de poux à l'école de recrues d'infanterie d'exploration et de transmission 213

Les cantonnements de la compagnie IV de l'école de recrues d'infanterie d'exploration et de transmission 213, stationnée à Avenches, sont envahis par les puces et les poux. Des piqûres de puces ont été relevées sur plus d'un tiers des militaires. On a aussi retrouvé des cadavres de poux sur 3 militaires. Des mesures d'urgence ont été prises en conséquence.

Des piqûres de puces ont été diagnostiquées sur plus d'un tiers des 155 hommes de la compagnie IV de l'école de recrues d'infanterie d'exploration et de transmission 213. Des cadavres de poux, mais aucun oeuf, ont également été décelés sur 3 militaires. Ces insectes sont transmis par contact personnel. La cause de cette invasion n'est pas claire; ces insectes semblent toutefois avoir essaimé à partir du local de garde.

Le médecin de troupe a donné immédiatement les soins nécessaires aux militaires concernés et il a ordonné les mesures d'hygiène qui s'imposaient. Des produits spéciaux ont été remis pour les soins corporels. Tout le matériel personnel de la compagnie a été emballé hermé-tiquement et apporté à l'arsenal cantonal de Fribourg. La troupe sera déplacée dans un complexe industriel. Une section d'hygiène de l'école de recrues d'hôpital 268, stationnée à Moudon, va désinfecter tous ces cantonnements. On estime qu'avec ces mesures sanitaires appropriées la troupe pourra réintégrer ses cantonnements vendredi au plus tard.

Renseignements complémentaires: Colonel EMG Peter Meier, commandant de l'école de recrues d'infanterie d'exploration et de transmission 213 Téléphone 026-350 82 00

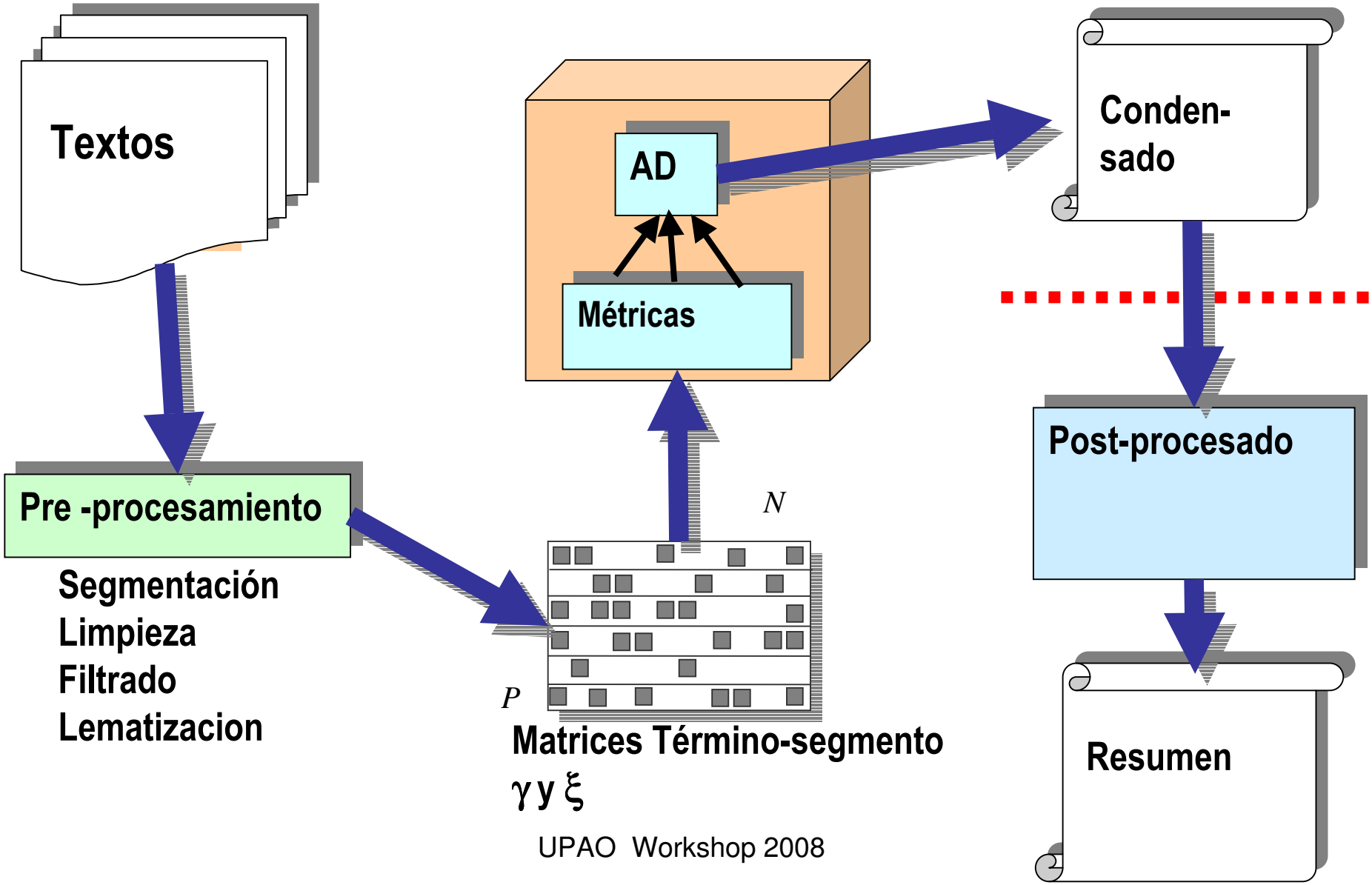
El regimiento de la compania IV de la escuela de infanteria y de transmision 213 de Avenches, fue invadido por pulgas y piojos. Una seccion de higiene de la escuela del hospital 268 de Moudon, va a désinfectar los cuarteles

Done

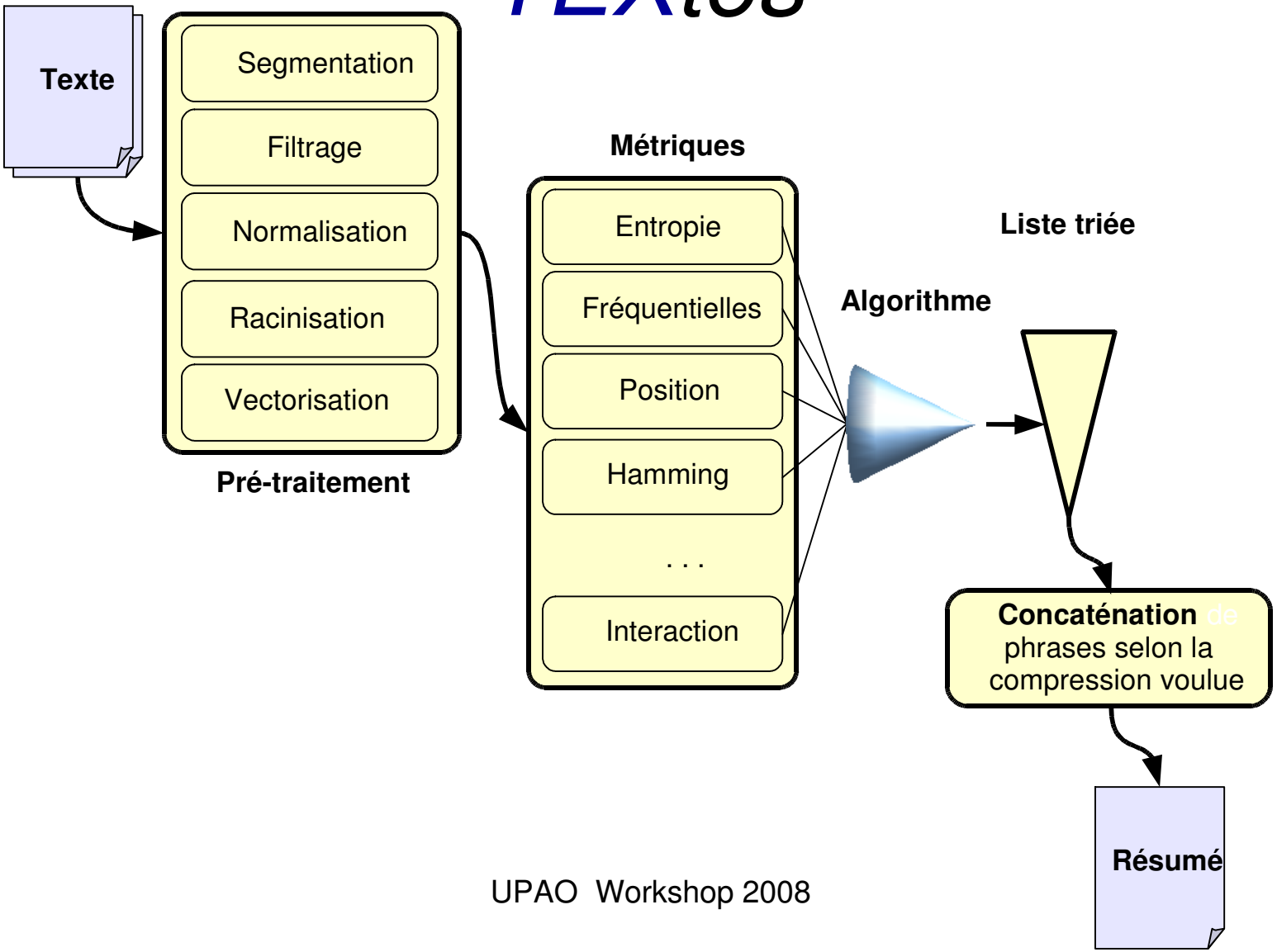
Internet

Start | Voyages-sn... | Invasion ... | 3½ Floppy (... | poster | arco2001_p... | Microsoft P... | EN 20:27

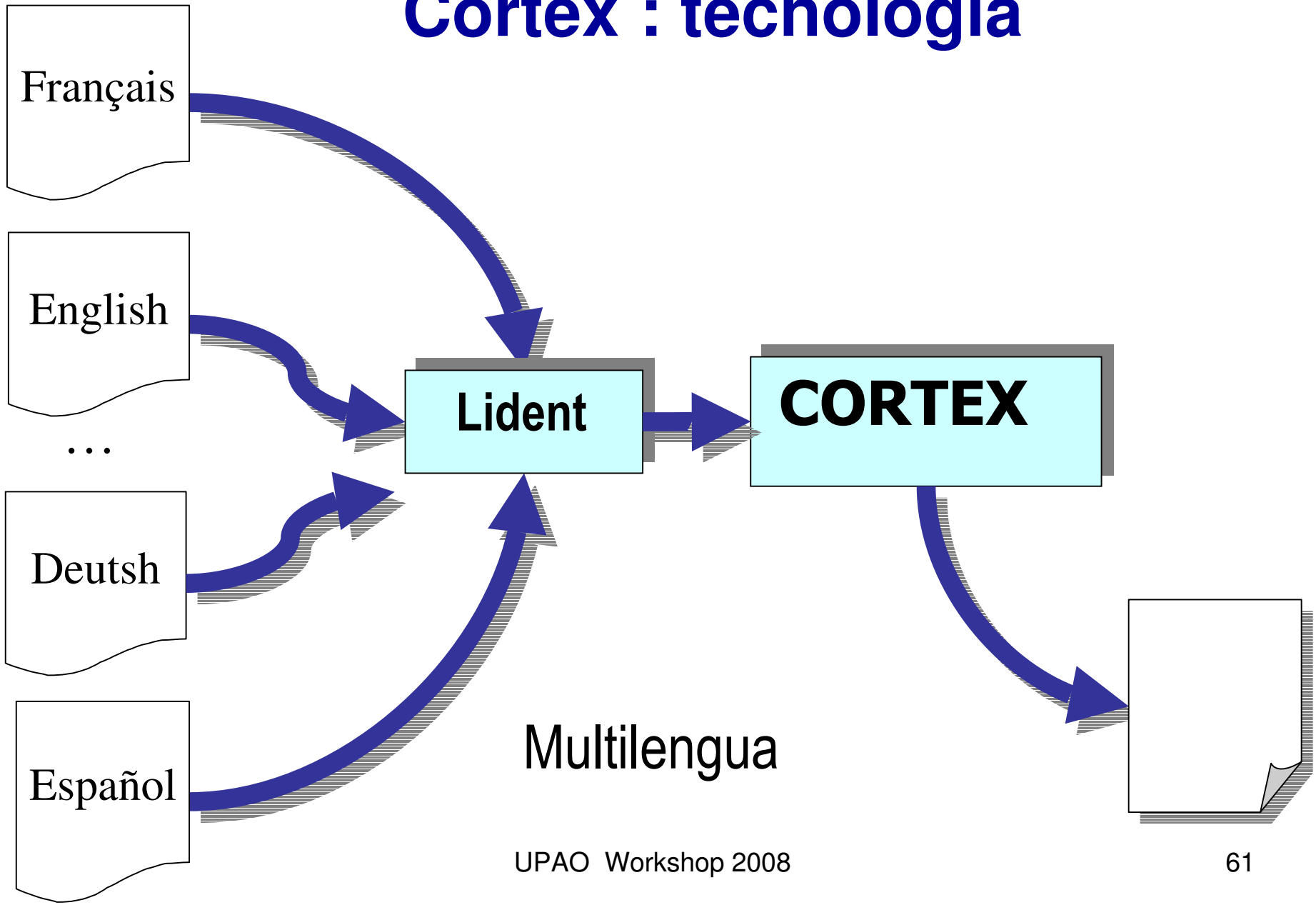
CORTEX: una maquina de resumenes



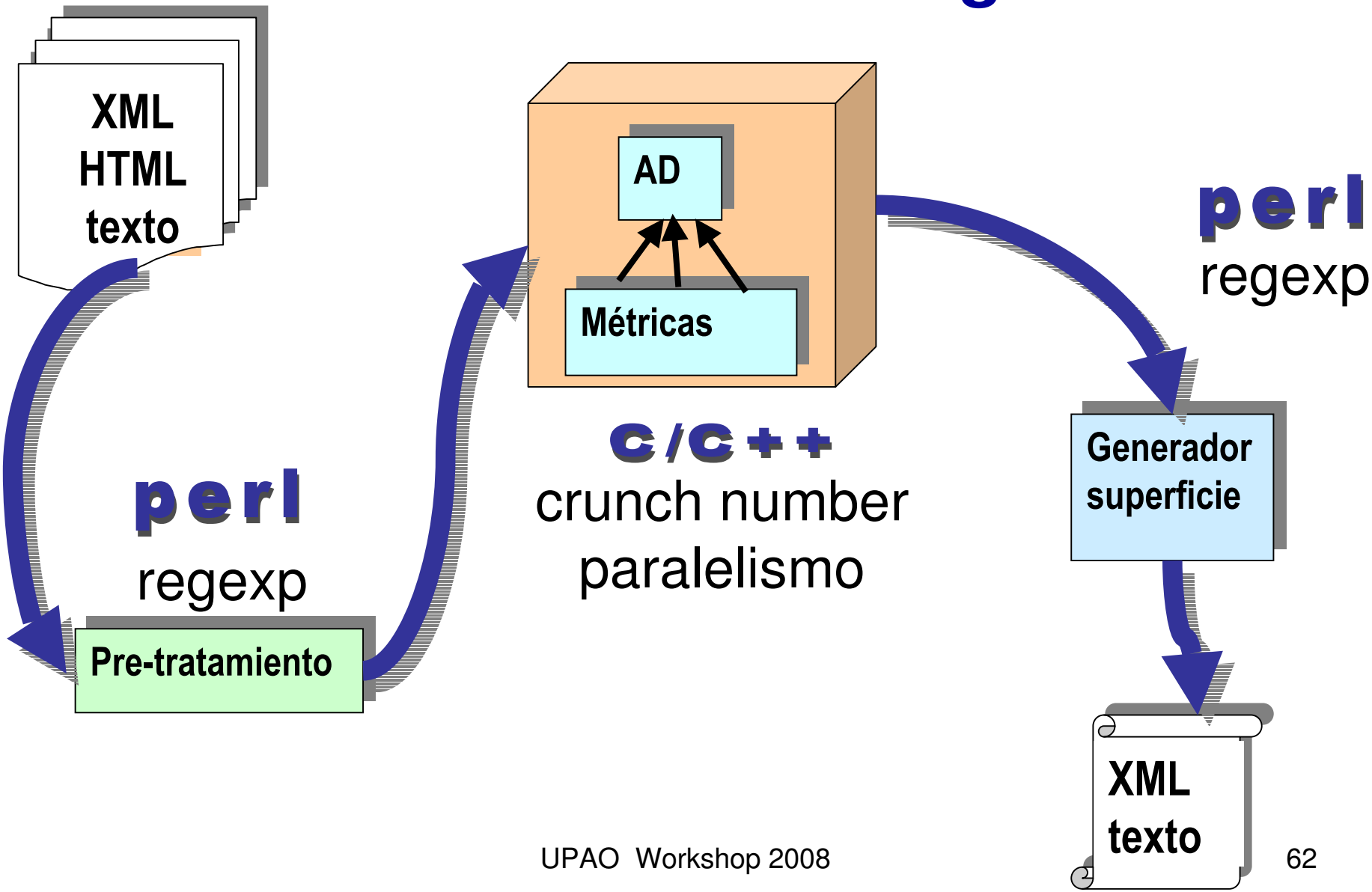
Cortex : *Otro Resumidor de TEXtos*



Cortex : tecnologia



Cortex : tecnología



Pre-procesamiento

- Segmentación
- Filtrado
- Lematización
- Generación de matrices

Medio simple de acercarse a la semántica

Segmentación

- Por frases
  - Marcadores . : ? !
- Por párrafos
  - n frases
  - Identificación del párrafo

Filtrado

- Texto entre paréntesis (bla blabla bla, bla)
- Números / cifras
- Palabras funcionales
  - Artículos, conjunciones, pronombres...
- Verbos funcionales
  - SER, ESTAR TENER, DEBER,…
- Expresiones
  - Es decir, entonces, a lo mejor, ...

Filtrar palabras funcionales

Palabras muertas del vocabulario

- ¿Cuáles ?
  - Palabras muy frecuentes (artículos, coordinación, pronombres …) : **de, la, l', le, à, les, et, des, d', en, un , du, une, …**

- ¿ Las palabras-funcionales son realmente inútiles ?
  - Utiles : clasificar textos
  - Utiles : reconocimiento de la palabra
  - Inútiles : indexación de documentos
  - **Inútiles : resumen por extracción**

Lematizacion

- **Cantar, cantemos, canten, cantaban, cantantes, cantó, cantarian, cantador, cancion, cancionero,…**

  $\Rightarrow$ **CANTAR**

- **Reducir el espacio vectorial**

- **Diccionarios**
  - ~ 200 000 formas en francés
  - ~ 500 000 formas en español

Vectorizacion

Las casas azules de mis tias.
Una de mis tias se llama Lulu.
Me gusta mucho su casa.
El azul es mi color !
Yo tengo unos zapatos azules nuevos.

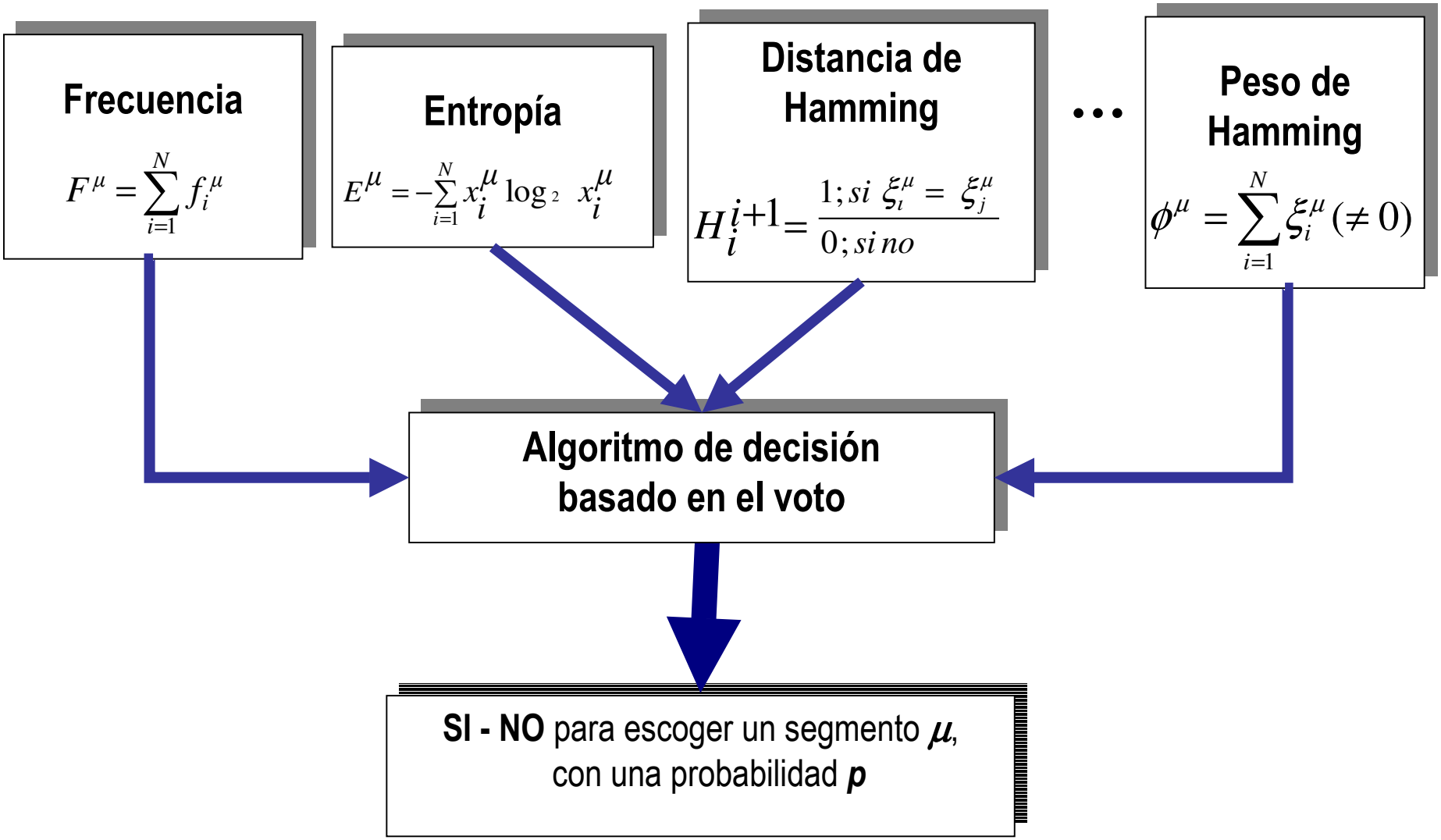
casa	azul	tia	llamar	lulu	gustar	nuevo	zapato	color
TF	TF	TF	0	0	0	0	0	0
0	0	TF	TF	TF	0	0	0	0
TF	0	0	0	0	TF	0	0	0
0	TF	0	0	0	0	0	0	TF
0	TF	0	0	0	0	TF	TF	0

- **Modelo vectorial (bag-of-words)**
- **Palabras filtradas, normalizadas y lematizadas**

Métricas

- **Frecuenciales**
  - Frecuencia
  - Probabilidades
  - TF*IDF
- **Informacionales**
  - Entropía
  - Interacción
- **Hamming**
  - Peso, longitud, distancia
- **Títulos**
  - Títulos, sub-títulos…
- **Posición**
- **Otras**

CORTEX : estructura



Ejemplo

Palabras

Frases

	a	b	c	d
	i=1	*i=2*	*i=3*	*i=4*
μ=1	1	1	0	0
μ=2	1	10	1	0
μ=3	1	0	1	2

Matriz frecuencias

	a	b	c	d
	i=1	*i=2*	*i=3*	*i=4*
μ=1	1	1	0	0
μ=2	1	1	1	0
μ=3	1	0	1	1

Matriz binaria

Matrices huecas

BIBLE.TXT : 99.27% está vacío (CORAN tambien!)

- **801920** palabras en total / 50019 frases
- 8195 palabras diferentes / 49993 frases
- **5777** palabras (*f* >2) / 45250 frases
- Tiempos en G5 Mac OSX
  - **Segmentación en 8 segundos**
  - **Resumen en 5 minutos**

Matrices huecas

N columnas

0	0	0	1	0
3	0	0	0	0
0	0	0	0	0
7	0	0	0	2

k elementos $\neq 0$

P líneas

P líneas

línea	columna	γ
1	4	1
2	1	3
4	1	7
4	5	2

Frecuencia de las palabras

- Palabras **pertinentes** de la frase
- Más palabras importantes tiene la frase → más posibilidades de retenerla
- Más larga es la frase → más palabras pertinentes → más posibilidades de retenerla

- **Resumenes constituidos generalmente de frases largas**

Cálculo de la Frecuencia F

Palabras

Frases

	$i=1$	$i=2$	$i=3$	$i=4$
$\mu=1$	1	1		
$\mu=2$	1	10	1	
$\mu=3$	1		1	2

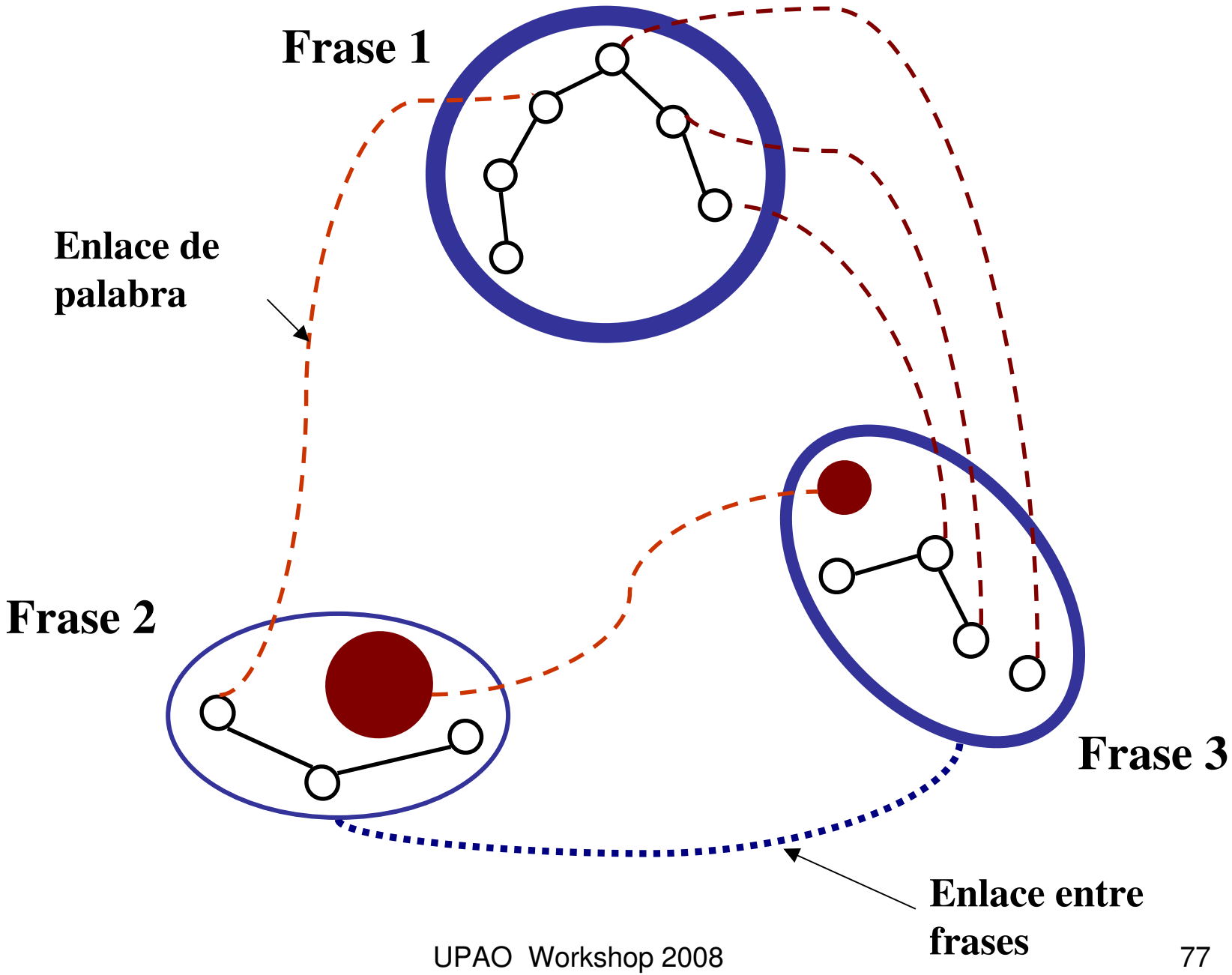
F
2
12
4

p	3/18	11/18	2/18	2/18

$\Sigma = T = 18$

Interacción de frases *I*

- Cortex es un sistema que crea redes de palabras
  - Palabras de la misma familia
  - Presencia de estas palabras en otras frases del texto


Frase 1
Enlace de palabra
Frase 2
Frase 3
Enlace entre frases

Cálculo de *Interacciones*

Palabras

Frases

	i=1	*i=2*	*i=3*	*i=4*
μ=1	1	1		
μ=2	1	1	1	
μ=3	1		1	1

Interacc.
3
4
3

$$I^{\mu} = I^{1} + I^{2} + I^{3} = 2 + 1 + 1 = 4$$

Frecuencias y rareza

- La frase X contiene la misma palabra 10 veces
- Esa palabra no está en ninguna otra frase
- La frase X puede ser seleccionada por su alta frecuencia ($F = 10$)
- Esto es grave: esta palabra podria no ser importante

Frecuencias y rareza

- Sea otra palabra que aparece solo 1 vez en la frase X, pero que está en otras frases (una vez en cada frase)
- Ademas esta palabra se encuentra en el título (**tema principal)**
- Consecuencia : la métrica *F* escogerá la frase X en lugar de otras frases que contienen por lo menos una palabra del título

Frecuencias y rareza

Tener una frecuencia *F=10 es positivo*, pero

¿ realmente la frase X merece tener una alta ponderación ?

¿ Qué hacer ?

- **Re-ponderar** la frecuencia de las palabras por su rareza

Calculo Frecuencia x rareza

Rareza palabra 1 = 3; Rareza palabra 2 = 1;
Rareza palabra 3 =1; Rareza palabra 4=1

Frases

	i=1	*i=2*	*i=3*	*i=4*
μ=1	1			
μ=2	1	10		
μ=3	5		1	1
	3	**1**	**1**	**1**

Rareza

F
1
1 + 10 = 11
5 + 1 + 1 = 7

F x Rareza
1 x 3 = 3
1x3 + 10x1 = 13
5x3 + 1x1 + 1x1 = 17

Entropía

- **Entropía** : medida que depende de la probabilidad p_i de una palabra
- Derivada de la teoría de la información
- Entre mas elevada sea p_i de las palabras seran mas favorizadas las frases que las contienen

$$E^{\mu} = -\sum_{\substack{i=0 \\ \xi_i^{\mu} \neq 0}}^{N_L - 1} p_i \log_2 p_i$$

Cálculo de *Entropía*

Palabras

Frases

	i=1	*i=2*	*i=3*	*i=4*
μ=1	1	1		
μ=2	1	1	1	
μ=3	1		1	1

F
0.67

p	3/18	11/18	2/18	2/18

$$E = -\,[\, p(1)\, log\, p(1) + p(2)\, log\, p(2) + p(3)\, log\, p(3)\,]$$

$$= 0.129 + 0.30 + 0.244 = \mathbf{0.67}$$

Matriz de Hamming H

Un autor varia el vocabulario: utiliza **sinónimos**

Idea : Bonificar la frase donde solo **una** de 2 palabras está presente

La matriz de Hamming trata de evaluar hasta qué punto 2 palabras son utilizadas individualmente:

- Localizar los sinónimos pertinentes

Cálculo : *Matriz de Hamming*

	1	i=2	i=3	4
1	1	1	0	0
2	1	1	1	0
3	1	0	1	1

Palabras

H

	i=1	i=2	i=3	i=4
i=1	X	1	1	2
i=2		X	2	3
i=3			X	1

Palabras

H

Palabra 3 = **0 1 1**

Palabra 2 = **1 1 0**

$H(2,3) = 2$

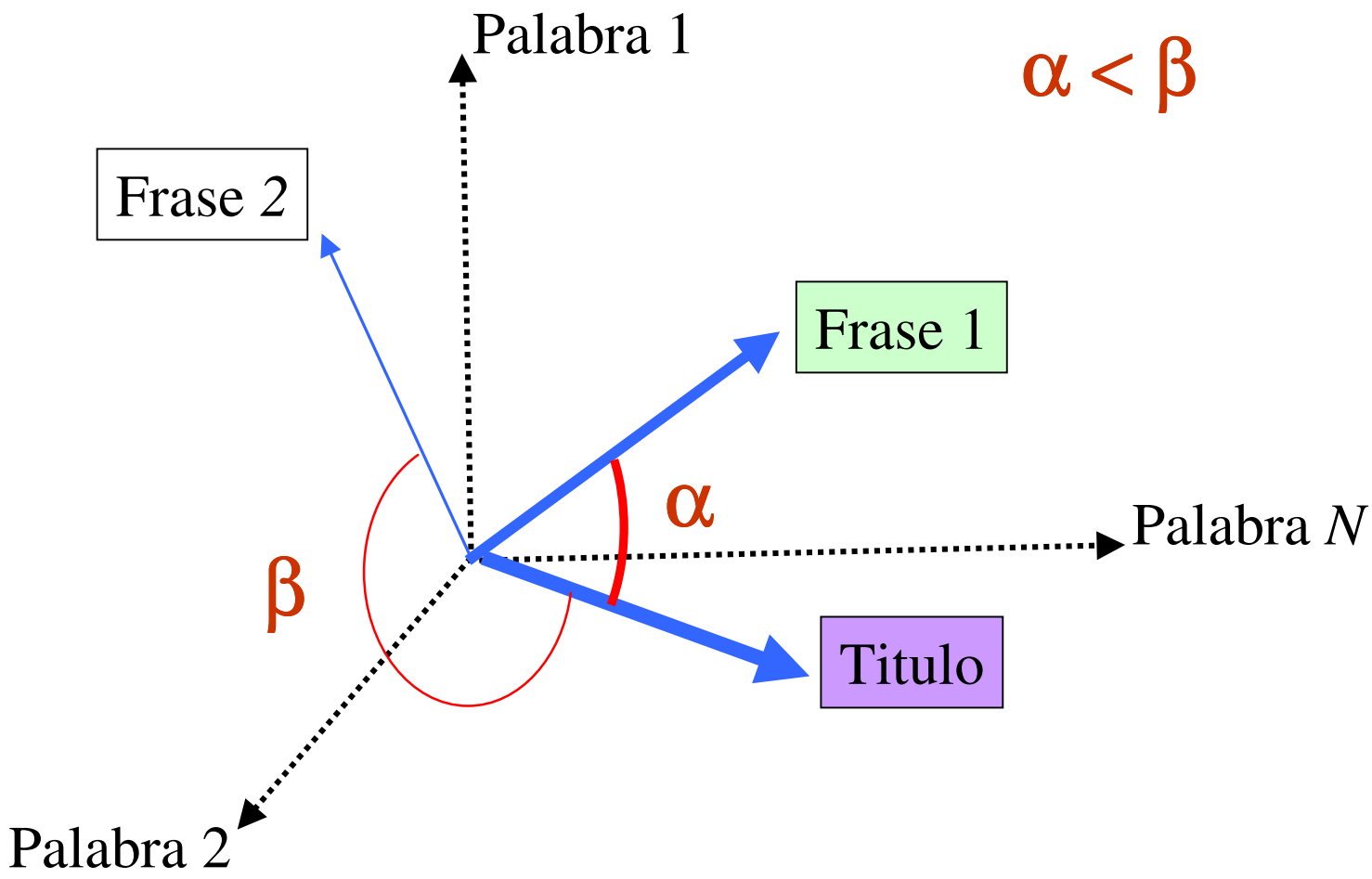
Título

- Los textos tienen un título
- **Bonificar** frases que hablen del tema en el título
- Comparar palabra a palabra **frase y título**
- **Producto escalar**: la frase y el título son vectores de dimension *N*

$$\cos(\alpha) = \left(\frac{< \gamma, \mathrm{Titre} >}{|\gamma| \cdot |\mathrm{Titre}|} \right)$$

$$\Theta^{\mu} = \arccos\left(\cos(\alpha)\right)$$

Título : calculo de angulos



Cálculo de angulo del *Título*

0	0	0	1

Título

Palabras

Frases

	i=1	*i=2*	*i=3*	*i=4*
μ=1	1	1		
μ=2	1	10	1	
μ=3	1		1	2

Angulo

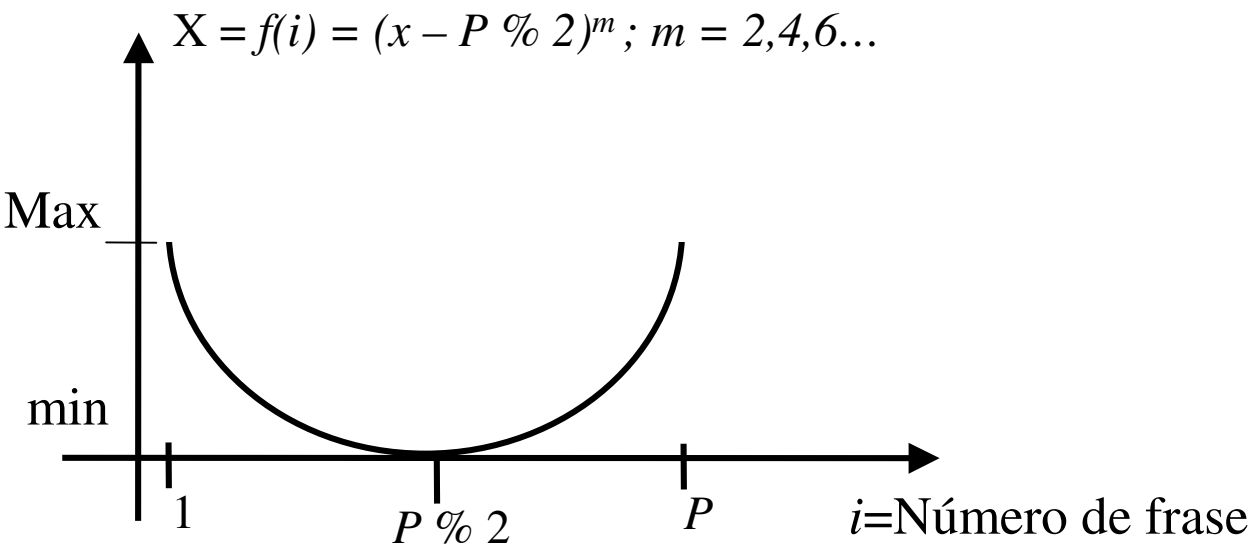
cos Θ
0
0
0.81

$$\textit{Título}^{\mu} = \textit{Título}(1)\ \gamma(3,1) + \textit{Título}(2)\ \gamma(3,2) + \textit{Título}(3)\ \gamma(3,2) + \textit{Título}(4)\ \gamma(3,4)$$

$$= 0 + 0 + 0 + 2/\mathrm{sqr}(6) = \mathbf{0.81}$$

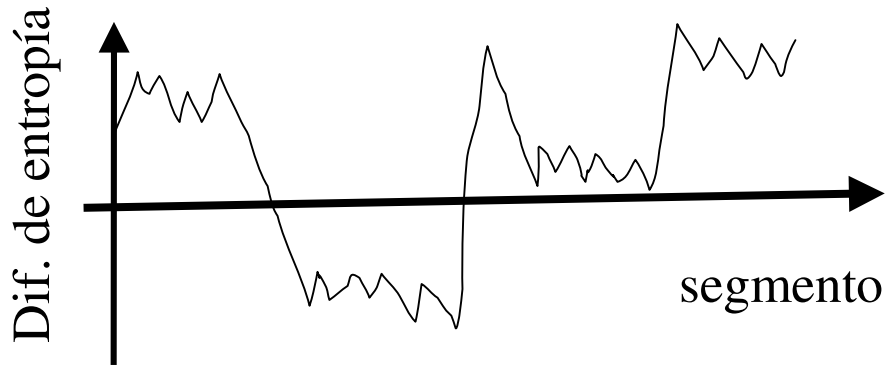
Posición de la frase

- Modeliza contenido semántico según su posición
- Modelo de artículos periodísticos



Otras métricas (plug in)

- Longitud de frases
- Diferencia de entropía condicional entre frases



- Entidades nombradas
- N-gramas
- Otras medidas de distancia

Normalización de las métricas

- Desviaciones importantes entre las métricas
- Normalizacion entre [0,1]

$$\|\lambda_\mu\| = \frac{\lambda_\mu - m}{M - m}$$

Algoritmo de decisión

- Compilar las métricas para dar una nota final a cada frase
- Idea : calcular una **media sofisticada**
  - Tendencia positiva $\lambda > 0{,}5$
  - Tendencia negativa $\lambda < 0{,}5$
  - $\lambda = 0{,}5$ es ignorado

➢ **Convergencia**

➢ **Amplificacion**

Generación del resumen

Post-procesamiento superficial

- A partir de la selección de frases pertinentes
- Tamaño definido por el usuario (%)
- Presentación
  - XML
  - Texto

Pruebas

- 14 Personas (nivel universitario) : escoger los segmentos que les parecieran pertinentes

- CORTEX
- Autoresumen MS-Word
- Summarizer Minds New Mexico State University **messene.nmsu.edu/minds**
- Pertinence **www.pertinence.net**
- Copernic **www.copernic.com**

Análisis del texto « PUCES »

- **Artificialmente ambiguo**
  - **Dos temas : « puces INFO » + « puces BIO »**
- ***N* = 279 palabras**
- ***P* = 30 frases**
- **Frases segmentadas por « : », « . », « ? », « ! »**
- **Extracto**
  - **25% del tamano = 8 segmentos**

Invasion de puces et de poux - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.admin.ch/cp/f/1997Sep10.064053.8237@idz.bfi.admin.ch.html Go
Google Search Web Search Site PageRank Page Info Up Highlight

PUCES BIO

Invasion de puces et de poux

Communiqué aux agences de presse cu 9 septembre 1997

Avenches: Invasion de puces et de poux à l'école de recrues d'infanterie d'exploration et de transmission 213

Les cantonnements de la compagnie IV de l'école de recrues d'infanterie d'exploration et de transmission 213, stationnée à Avenches, sont envahis par les puces et les poux. Des piqûres de puces ont été relevées sur plus d'un tiers des militaires. On a aussi retrouvé des cadavres de poux sur 3 militaires. Des mesures d'urgence ont été prises en conséquence.

Des piqûres de puces ont été diagnostiquées sur plus d'un tiers des 155 hommes de la compagnie IV de l'école de recrues d'infanterie d'exploration et de transmission 213. Des cadavres de poux, mais aucun oeuf, ont également été décelés sur 3 militaires. Ces insectes sont transmis par contact personnel. La cause de cette invasion n'est pas claire; ces insectes semblent toutefois avoir essaimé à partir du local de garde.

Le médecin de troupe a donné immédiatement les soins nécessaires aux militaires concernés et il a ordonné les mesures d'hygiène qui s'imposaient. Des produits spéciaux ont été remis pour les soins corporels. Tout le matériel personnel de la compagnie a été emballé hermé-tiquement et apporté à l'arsenal cantonal de Fribourg. La troupe sera déplacée dans un complexe industriel. Une section d'hygiène de l'école de recrues d'hôpital 268, stationnée à Moudon, va désinfecter tous ces cantonnements. On estime qu'avec ces mesures sanitaires appropriées la troupe pourra réintégrer ses cantonnements vendredi au plus tard.

Renseignements complémentaires: Colonel EMG Peter Meier, commandant de l'école de recrues d'infanterie d'exploration et de transmission 213 Téléphone 026-350 82 00

Done Internet
Start Voyages-sn... Invasion ... 3½ Floppy (... poster arco2001_p... Microsoft P... EN 20:27

par marques choisies - Wanadoo

Fichier Edition Affichage Favoris Outils ?

Précédente Rechercher Favoris Média

Adresse http://www.geocities.com/gaelmail/result3.html OK

PUCES INFO

Et si l ordinateur pouvait fonctionner un jour

sans électricité ou presque

La démarche de chercheurs américains de l université de Notre Dame

dans l Indiana

montre que l on peut manipuler des électrons pour construire des circuits élémentaires avec des quantités d énergie infimes

Leurs expériences

relatées dans l édition du 9 avril du magazine Science

ouvrent la voie à des composants capables de fonctionner à des fréquences 10 à 100 fois plus élevées que celles des puces actuelles qui sont bridées par des problèmes de dissipation de chaleur

Les travaux de l équipe dirigée par Greg Snider portent sur

le puits quantique

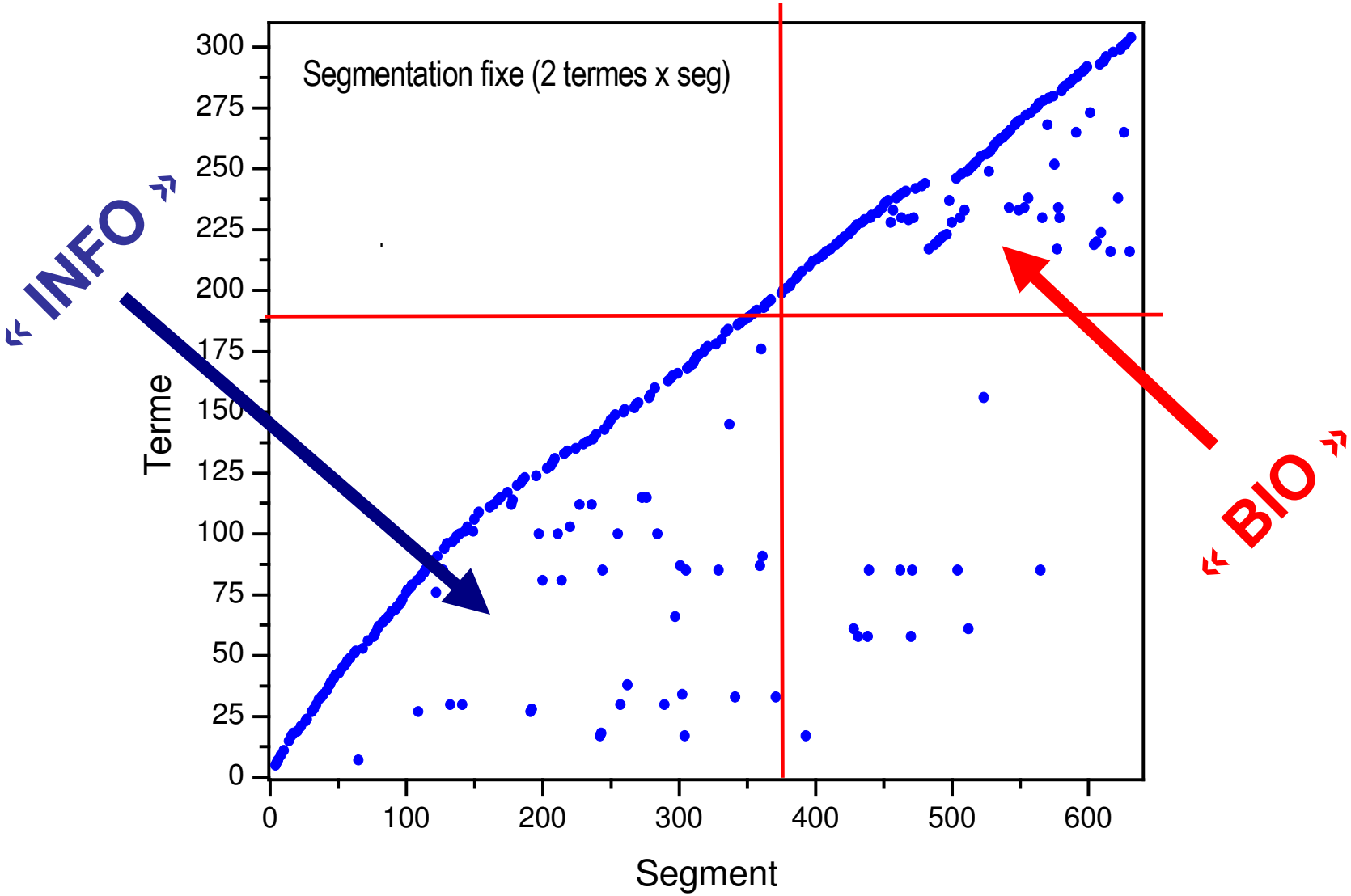
un piège infinitésimal dans lequel un électron peut être enfermé

Terminé Internet

PUCES

Et si l'ordinateur pouvait fonctionner un jour, sans électricité ou presque? La démarche de chercheurs américains de l'université de Notre Dame, dans l'Indiana, montre que l'on peut manipuler des électrons pour construire des circuits élémentaires avec des quantités d'énergie infimes. Leurs expériences, relatées dans l'édition du 9 avril du magazine Science, ouvrent la voie à des composants capables de fonctionner à des fréquences 10 à 100 fois plus élevées que celles des ***puces*** *actuelles qui sont bridées par des problèmes de dissipation de chaleur. Les travaux de l'équipe dirigée par Greg Snider portent sur le puits quantique, un piège infinitésimal dans lequel un électron peut être enfermé. Les scientifiques ont créé ...*

Les cantonnements de la compagnie IV de l'école de recrues d'infanterie d'exploration et de transmission 213, stationnée à Avenches, sont envahis par les **puces** et les poux. Des piqûres de **puces** ont été relevées sur plus d'un tiers des militaires. On a aussi retrouvé des cadavres de poux sur 3 militaires. Des mesures d'urgence ont été prises en conséquence. Des piqûres de **puces** ont été diagnostiquées sur plus d'un tiers des 155 hommes de la compagnie IV de l'école de recrues d'infanterie d'exploration et de transmission 213. Des cadavres de poux, mais aucun oeuf, ont également été décelés sur 3 militaires. Ces insectes sont transmis par contact personnel. La cause de cette invasion n'est pas claire; ces insectes semblent toutefois avoir essaimé à partir du local de garde. Le médecin de troupe a donné...


Segmentation fixe (2 termes x seg)
« INFO »
« BIO »
Terme
Segment
0
25
50
75
100
125
150
175
200
225
250
275
300
0
100
200
300
400
500
600

Texto PUCES.TXT

puces - Bloc-notes

Fichier Edition Format Affichage ?

Informatique et puces.

Et si l'ordinateur pouvait fonctionner un jour, sans électricité ou presque? La démarche de chercheurs américains de l'université de Notre Dame, dans l'Indiana, montre que l'on peut manipuler des électrons pour construire des circuits élémentaires avec des quantités d'énergie infimes. Leurs expériences, relatées dans l'édition du 9 avril du magazine Science, ouvrent la voie à des composants capables de fonctionner à des fréquences 10 à 100 fois plus élevées que celles des puces actuelles qui sont bridées par des problèmes de dissipation de chaleur. Les travaux de l'équipe dirigée par Greg Snider portent sur le puits quantique, un piège infinitésimal dans lequel un électron peut être enfermé. Les scientifiques ont créé des cellules carrées formées de quatre puits quantiques, dans laquelle ils ont introduit une paire d'électrons. Les forces de répulsion provoquent le déplacement des électrons qui trouvent leur équilibre lorsqu'ils se trouvent placés aux deux extrémités de l'une ou l'autre des diagonales de la cellule. La première représente l'état 0, tandis que l'autre indique le 1. Chaque cellule représente donc un bit, la plus petite quantité d'information que l'on peut manipuler dans les ordinateurs. Tout déplacement d'un électron sous l'effet d'une force extérieure provoque automatiquement le déplacement du second électron de manière à retrouver l'équilibre, et donc le basculement de la cellule entre les états 0 et 1. L'utilisation d'une cellule unique ne prouve rien. Les chercheurs américains ont réussi à en assembler plusieurs, provoquant, suivant leurs besoins, le déplacement des électrons sans devoir fournir d'énergie, ou presque. Dans les transistors actuels, le passage de l'état 0 à l'état 1 n'est possible qu'au prix du déplacement de plusieurs milliers d'électrons, ce qui génère un important flux de chaleur. En regroupant cinq cellules élémentaires, les chercheurs ont mis au point un circuit baptisé "majoritaire" capable de réaliser les deux fonctions logiques de base, ET et OU, à la demande. Ils ont ensuite vérifié son bon fonctionnement et espèrent assembler plusieurs de ces circuits pour effectuer des additions et des multiplications sur des nombres. En cas de succès, la technique des cellules logiques quantiques pourrait permettre d'entasser des centaines de milliards de circuits dans une seule puce électronique. Pour l'instant, le dispositif fonctionne seulement à une température voisine du zéro absolu, mais les chercheurs ne désespèrent pas de parvenir à le réchauffer tout en maîtrisant son comportement.

Les cantonnements de la compagnie IV de l'école de recrues d'infanterie d'exploration et de transmission 213, stationnée à Avenches, sont envahis parles puces et les poux. Des piqûres de puces ont été relevées sur plus d'untiers des militaires. On a aussi retrouvé des cadavres de poux sur 3 militaires. Des mesures d'urgence ont été prises en conséquence. Des piqûres de puces ont été diagnostiquées sur plus d'un tiers des 155 hommes de la compagnie IV de l'école

Segmentacion PUCES.XML

C:\cygwin\home\pollito\Projet\cortex\SortiesManuel\puces\Tmp\puces_objt.xml - Wanadoo

Fichier Edition Affichage Favoris Outils ?

Précédente Rechercher Favoris Média

Adresse C:\cygwin\home\pollito\Projet\cortex\SortiesManuel\puces\Tmp\puces_objt.xml OK


```
<?xml version="1.0" encoding="UTF-8" ?>
- <Texte Langue="Fra" Titre="Informatique et puces">
  <Phrase Indice="0">Et si l'ordinateur pouvait fonctionner un jour, sans électricité ou presque?
    </Phrase>
  <Phrase Indice="1">La démarche de chercheurs américains de l'université de Notre Dame, dans
    l'Indiana, montre que l'on peut manipuler des électrons pour construire des circuits
    élémentaires avec des quantités d'énergie infimes.</Phrase>
  <Phrase Indice="2">Leurs expériences, relatées dans l'édition du 9 avril du magazine Science,
    ouvrent la voie à des composants capables de fonctionner à des fréquences 10 à 100 fois
    plus élevées que celles des puces actuelles qui sont bridées par des problèmes de
    dissipation de chaleur.</Phrase>
  <Phrase Indice="3">Les travaux de l'équipe dirigée par Greg Snider portent sur le puits
    quantique, un piège infinitésimal dans lequel un électron peut être enfermé.</Phrase>
  <Phrase Indice="4">Les scientifiques ont créé des cellules carrées formées de quatre puits
    quantiques, dans laquelle ils ont introduit une paire d'électrons.</Phrase>
  <Phrase Indice="5">Les forces de répulsion provoquent le déplacement des électrons qui
    trouvent leur équilibre lorsqu'ils se trouvent placés aux deux extrémités de l'une ou l'autre
    des diagonales de la cellule.</Phrase>
  <Phrase Indice="6">La première représente l'état 0, tandis que l'autre indique le 1.</Phrase>
  <Phrase Indice="7">Chaque cellule représente donc un bit, la plus petite quantité d'information
    que l'on peut manipuler dans les ordinateurs.</Phrase>
  <Phrase Indice="8">Tout déplacement d'un électron sous l'effet d'une force extérieure
    provoque automatiquement le déplacement du second électron de manière à retrouver
    l'équilibre, et donc le basculement de la cellule entre les états 0 et 1.</Phrase>
```

Filtrado PUCES-FIL.XML

C:\cygwin\home\pollito\Projet\cortex\SortiesManuel\puces\Tmp\puces_fonc.xml - Wanadoo

Fichier Edition Affichage Favoris Outils ?

Précédente Rechercher Favoris Média

Adresse C:\cygwin\home\pollito\Projet\cortex\SortiesManuel\puces\Tmp\puces_fonc.xml OK


```
<?xml version="1.0" encoding="UTF-8" ?>
- <Texte Langue="Fra" Titre="informatique puces">
  <Phrase Indice="0">ordinateur pouvait fonctionner jour électricité</Phrase>
  <Phrase Indice="1">démarche chercheurs américains université notre_dame indiana montre
    peut manipuler électrons construire circuits élémentaires quantités énergie
    infimes</Phrase>
  <Phrase Indice="2">expériences relatées édition 9 avril magazine science ouvrir_voie
    composants capables fonctionner fréquences 10 100 élevées puces actuelles sont bridées
    problèmes dissipation chaleur</Phrase>
  <Phrase Indice="3">travaux équipe dirigée greg snider portent puits_quantique piège
    infinitésimal électron enfermé</Phrase>
  <Phrase Indice="4">scientifiques ont créé cellules carrées formées quatre puits_quantique ont
    introduit paire électrons</Phrase>
  <Phrase Indice="5">forces répulsion provoquent déplacement électrons trouvent équilibre
    trouvent placés deux extrémités diagonales cellule</Phrase>
  <Phrase Indice="6">première représente état 0 indique 1</Phrase>
  <Phrase Indice="7">cellule représente bit quantité information peut manipuler
    ordinateurs</Phrase>
  <Phrase Indice="8">déplacement électron effet force extérieure provoque automatiquement
    déplacement second électron manière retrouver équilibre basculement cellule états 0
    1</Phrase>
  <Phrase Indice="9">utilisation cellule unique prouve</Phrase>
  <Phrase Indice="10">chercheurs américains ont réussi assembler provoquant besoins
    déplacement électrons devoir fournir énergie</Phrase>
```

Lematizacion PUCES-LEM.XML

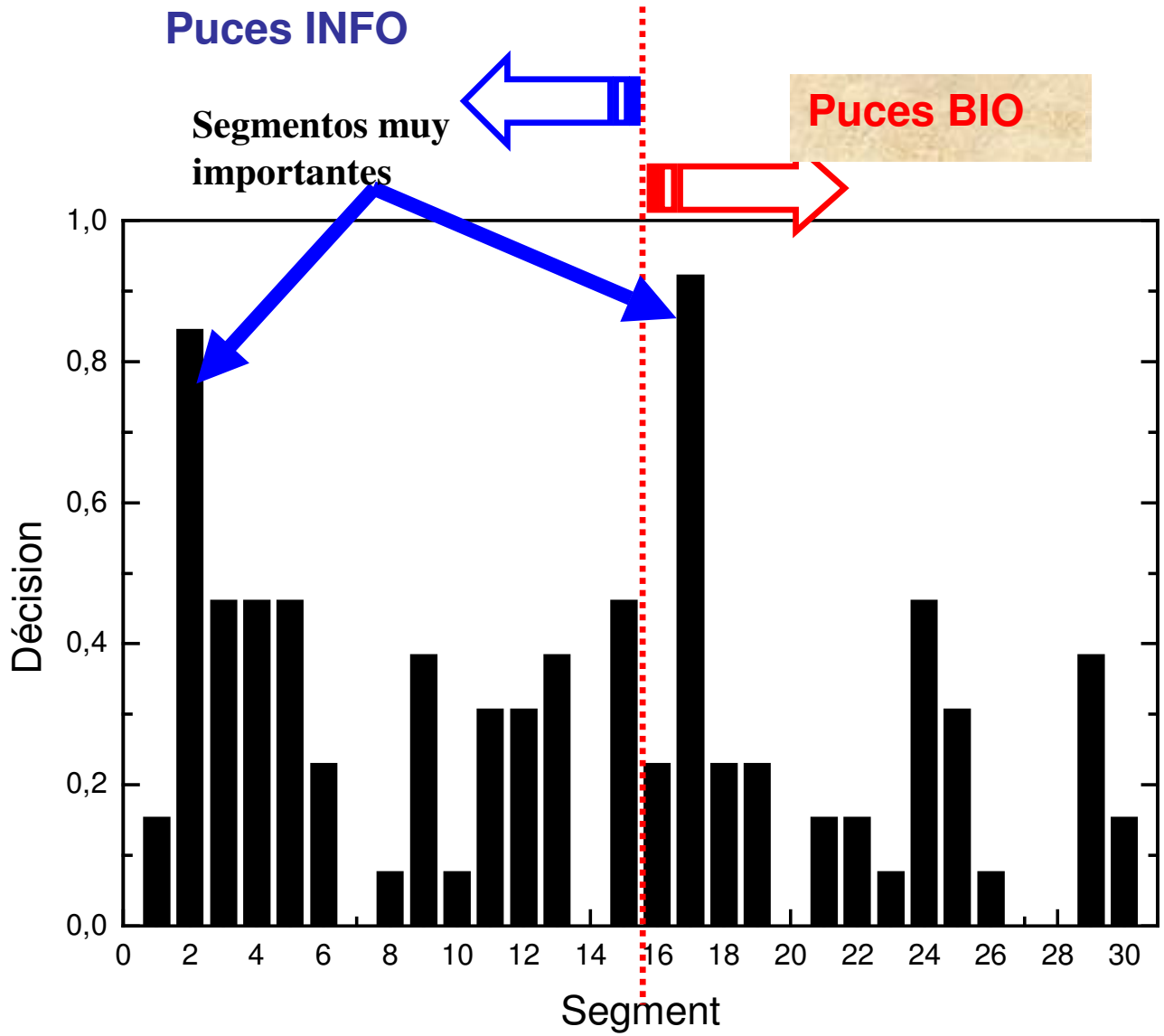
C:\cygwin\home\pollito\Projet\cortex\SortiesManuel\puces\Tmp\puces_vfnc.xml - Wanadoo

Fichier Edition Affichage Favoris Outils ?

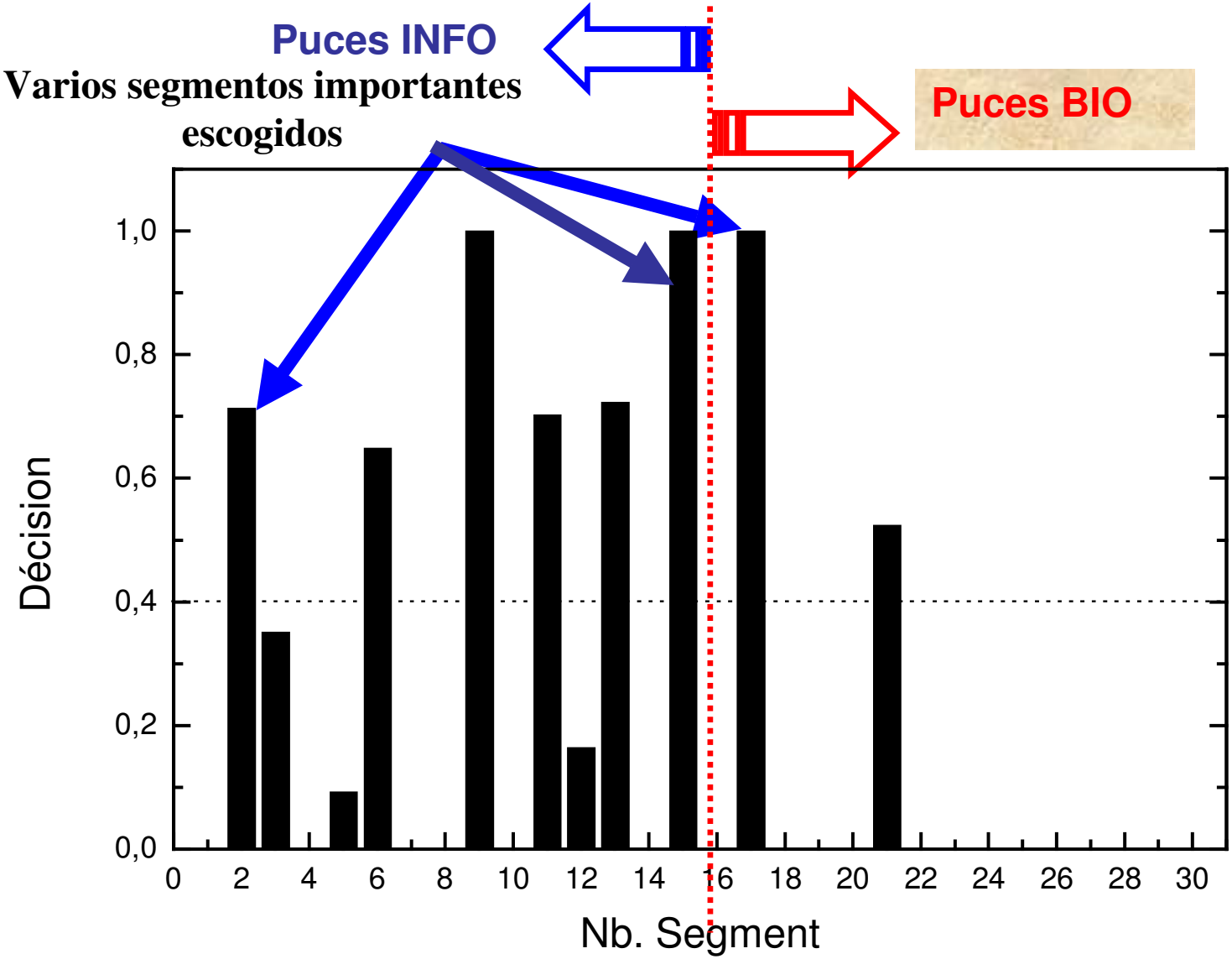
Précédente Rechercher Favoris Média

Adresse C:\cygwin\home\pollito\Projet\cortex\SortiesManuel\puces\Tmp\puces_vfnc.xml OK

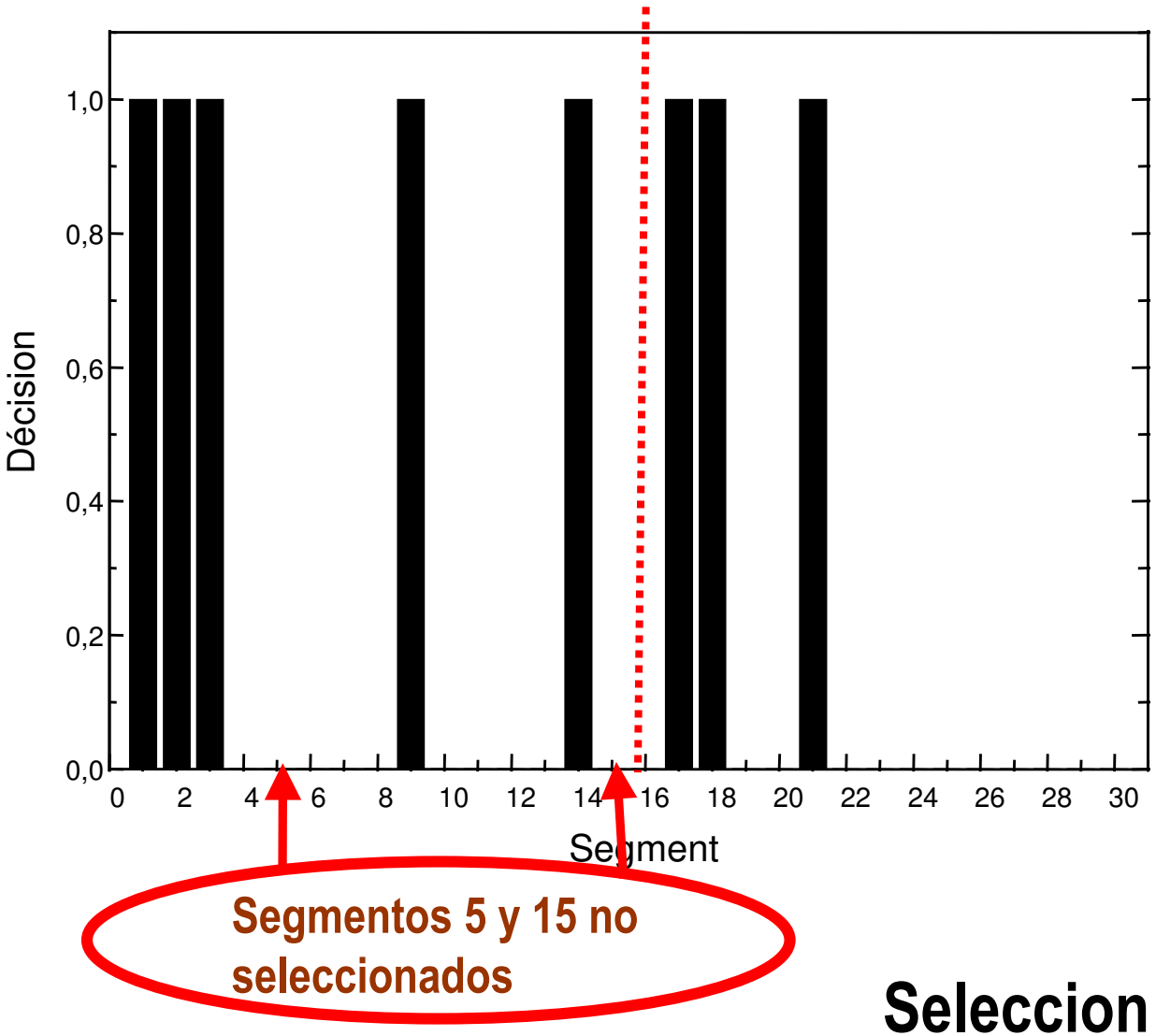

```xml
<?xml version="1.0" encoding="UTF-8" ?>
- <Texte Langue="Fra" Titre="informatique puce">
  <Phrase Indice="0">ordinateur fonctionner jour électricité</Phrase>
  <Phrase Indice="1">démarche chercheur américain université notre_dame indiana montrer
    manipuler électron construire circuit élémentaire quantité énergie infime</Phrase>
  <Phrase Indice="2">expérience relater éditer avril magazine science ouvrir_voie composant
    capable fonctionner fréquence élever puce actuel bridé problème dissipation
    chaleur</Phrase>
  <Phrase Indice="3">travailler équiper diriger greg snider porter puits_quantique piéger
    infinitésimal électron enfermer</Phrase>
  <Phrase Indice="4">scientifique créer cellule carrer former puits_quantique introduire pair
    électron</Phrase>
  <Phrase Indice="5">forcer répulsion provoquer déplacer électron trouver équilibrer trouver
    placer extrémité diagonal cellule</Phrase>
  <Phrase Indice="6">représenter état indiquer</Phrase>
  <Phrase Indice="7">cellule représenter bit quantité informer manipuler ordinateur</Phrase>
  <Phrase Indice="8">déplacer électron effet forcer extérieur provoquer automatique déplacer
    électron manière retrouver équilibrer basculer cellule état</Phrase>
  <Phrase Indice="9">utiliser cellule unique prouver</Phrase>
  <Phrase Indice="10">chercheur américain réussir assembler provoquer besoin déplacer
    électron fournir énergie</Phrase>
  <Phrase Indice="11">transistor passer état état possible prix déplacer électron générer
    important flux chaleur</Phrase>
  <Phrase Indice="12">regrouper cellule élémentaire chercheur mis_au_point circuit baptiser
```



Seleccion de los segmentos por las personas



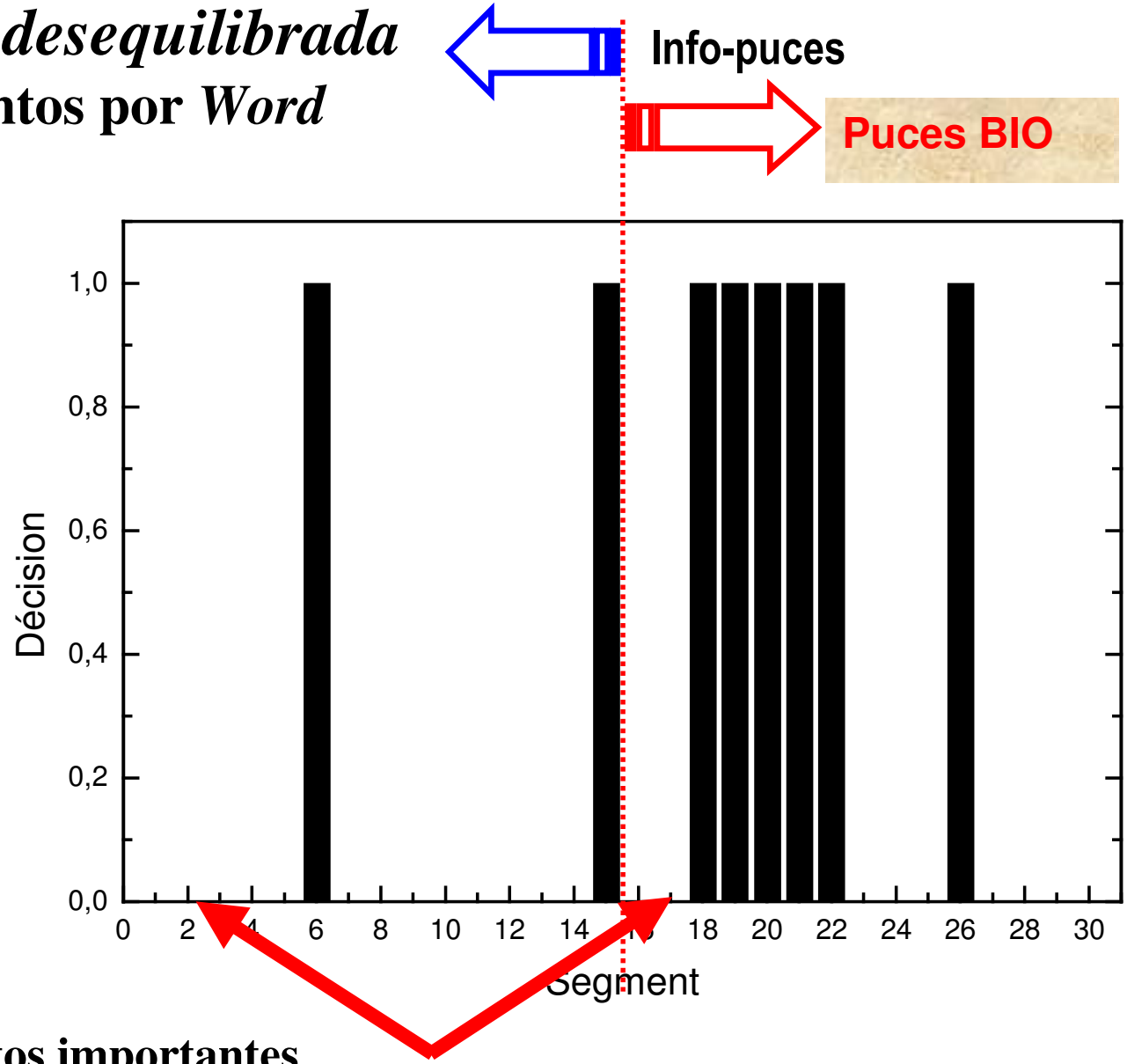
Seleccion de segmentos pertinentes por CORTEX



PUCES

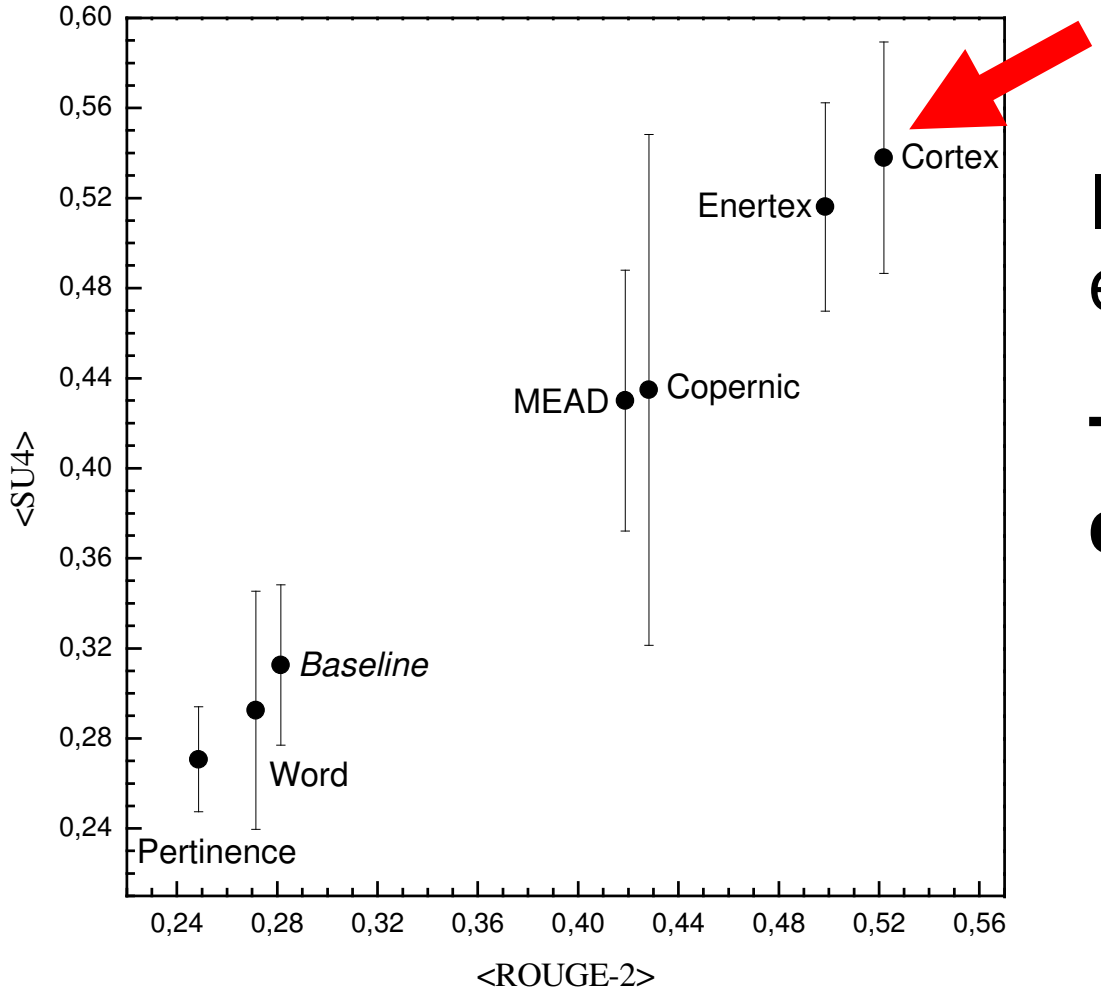
Seleccion de segmentos por Minds

Selección *desequilibrada* de segmentos por *Word*



Segmentos importantes no seleccionados

Resultados : resúmen genérico



Inglés, francés y español

Textos compuestos

Rappel moyen Rouge (Lin, 04)

Evaluation multi-documentos

NIST – DUC 2005-06

45 temas / 25 grupos de documentos en inglés

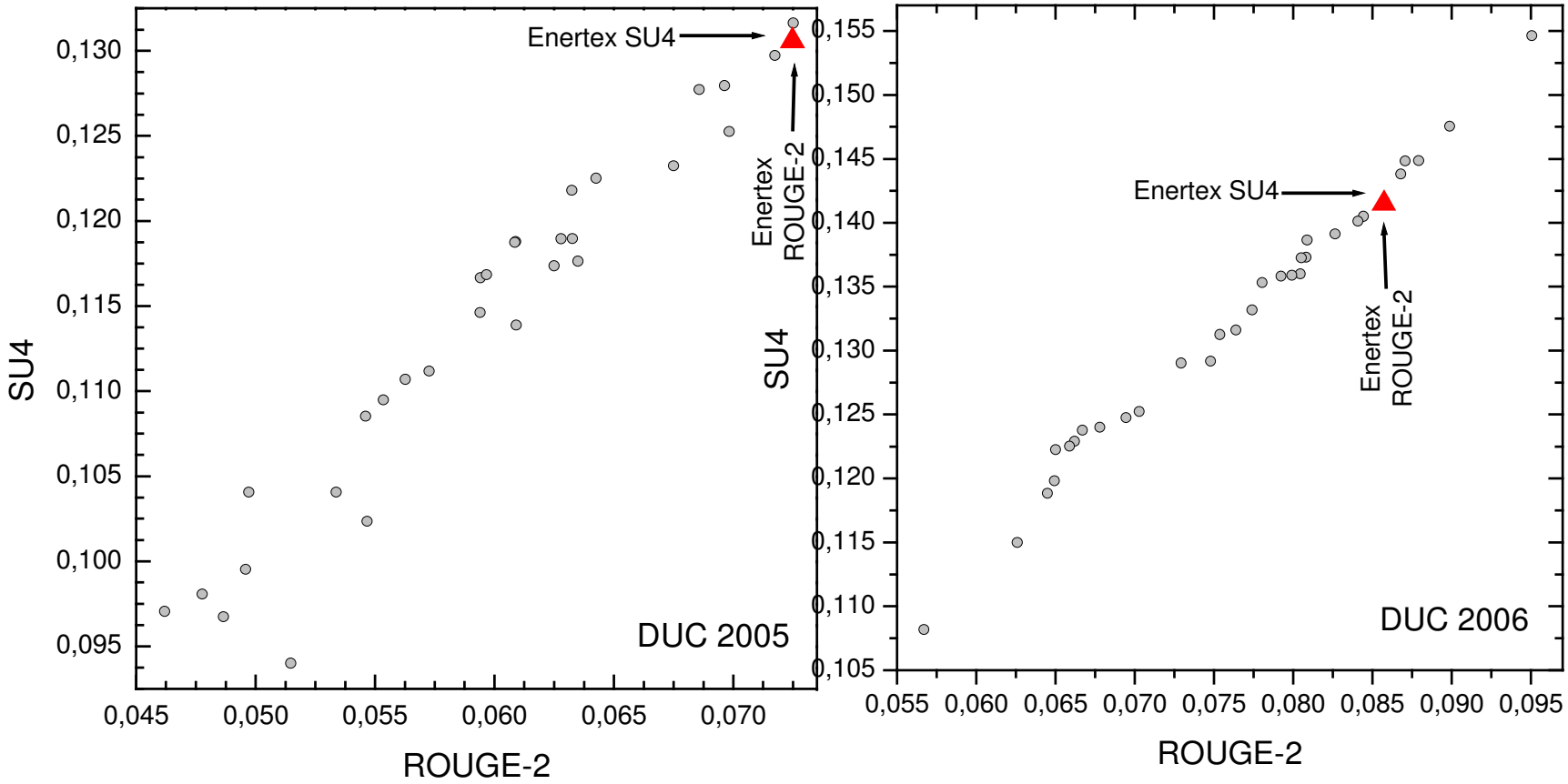
Generar resumenes de 250 palabras que responden a las preguntas del tema

Dos estrategias adicionales:

Éliminacion de la redundancia

Diversificar el contenido

Resultados : resúmen guiado por tema



Conclusión

- Algoritmo poderoso de resumen de textos
- Métricas independientes, modulares (plug in)
- Algoritmo de decision
- Independiente del tamaño del texto
- Independiente del tema
- Multilingüe
- Extractos equilibrados
- Extractos genéricos o personalizados
- Rapido

Perspectivas

Pre-tratamiento fino

- Modulo lingüístico

Post-tratamiento fino

- Modulo lingüístico
- Generación de texto

Mejora de los condensados :

- Aprendizaje supervisado
- Redes de neuronas
- Cálculo del erreur
- **Compresion de frases**

Compresion automatica de frases

Texto

- Trujillo, la bella capital de Libertad, goza, segun veo, de un clima verdaderamente excepcional.
- Por ese hecho, los inteligentes estudiantes de la UPAO, son recompensados en sus enormes efuerzos.
- Esperemos entonces que les ha gustado este curso tan interesante !

Resumen por compresion

- ***Trujillo capital de Libertad goza de un clima excepcional***
- ***Los estudiantes de la UPAO son recompensados en sus efuerzos***
- ***Esperemos que les ha gustado este curso!***

Tasa de compresion palabras : 26/40 = 66 %

Lecturas

American National Standards for Writing Abstracts. ANSI Inc., USA.

Balpe J., Lelu A., Papy F., and Saleh I. (1996). Techniques avancées pour l'hypertexte. Éditions Hermès, Paris.

Deerwester S., Dumais D., Furnas T., Launder G., and Harshman T. (1990). Indexing by latent semantic analysis. Journal of the Amer. Soc for Infor. Science, 6(41):391-407.

Huot F. (2000). Copernic summarizer ou la tentation de l'impossible. Québec Micro, 6.12(12):61-64.

Leloup C. (1997). Moteurs d'indexation et de Recherche. Eyrolles.

Meunier J.-G. and Nault G. (1997). Approche connexioniste au problème de l'extraction de connaissances terminologiques à partir de textes. In Les Techniques d'intelligence artificielle appliquées aux Technologies de l'Information, pages 62--76. Les Cahiers scientifiques ACFAS 90.

Morris A., Kasper G., and Adams D. (1999). The effects and limitations of automated text condensing on reading comprehension performance. In Advances in automatic text summarization, pp 305-323. The MIT Press, U.S.A.

Saggion H. and Lapalme G. (2000). Concept identification and presentation in the context of technical text summarization. In Automatic Summarization Workshop, pages 1--10, Seattle. ANLP/NAACL.

Salton G. (1971). The SMART Retrieval System - Experiments un Automatic Document Processing. Englewood Cliffs.

Salton G. and McGill M. (1983). Introduction to Modern Information Retrieval. McGraw-Hill.

Seffah A. and Meunier J.-G. (1996). Aladin: an integrated object-oriented environment for computer assited text analisys. Cahiers de recherche 96.1, LANCI-UQAM.

Torres-Moreno, J.M, Velázquez-Morales, P. et Meunier, J.G., *Condensés de textes par des méthodes numériques.* JADT 2002, Vol. (2):723-734, A. Morin & P. Sébillot éditeurs, IRISA/INRIA, France 2002.

Torres-Moreno, J.M, Velázquez-Morales, P. et Meunier, J.G., *Cortex : un algorithme pour la condensation automatique des textes*. ARCo 2001, Lyon, pp 365 + vol 2. ISC-Lyon, pp 65-5, Décembre 2001.

Torres-Moreno J.-M., Velazquez-Morales P., and Meunier J. (9-11 Mars 2000). Classphères : un réseau incrémental pour l'apprentissage non supervisé appliqué à la classification de textes. In JADT 2000, pages 365--372, Lausanne. EPFL M. Rajman & J.-C. Chappelier éditeurs.

Merci! C'est fini!

Avez-vous des questions?
¿Tienen preguntas?

Juan-Manuel Torres

juan-manuel.torres@univ-avignon.fr

Laboratoire Informatique d'Avignon (France)

UNIVERSITÉ D'AVIGNON
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE