Estructuración de resúmenes de artículos biomédicos para una mejor comprensión del lector

Àlex Bravo Serrano

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje
upf. Universitat Pompeu Fabra Barcelona
taln
GOBIERNO DE ESPAÑA
MINISTERIO DE ENERGÍA, TURISMO Y AGENDA DIGITAL

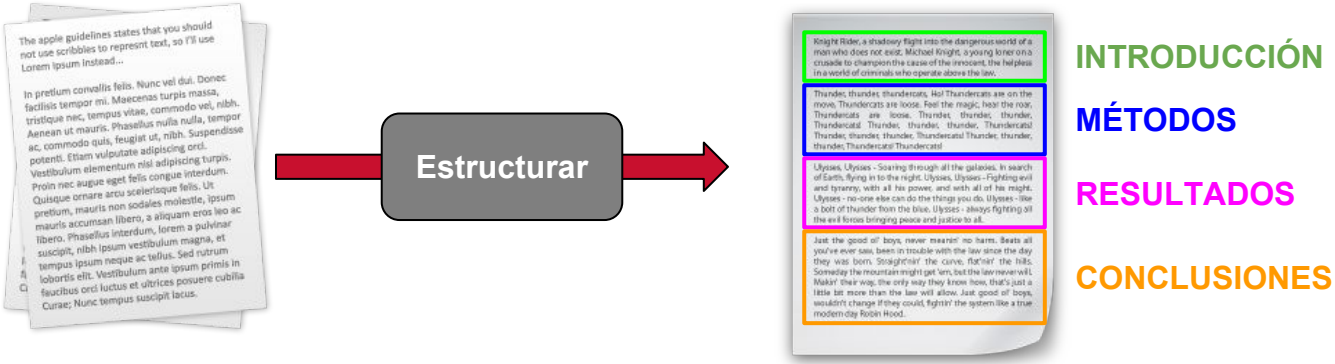
Introducción

Muchos artículos contienen resúmenes no estructurados.

→ Obliga al lector a leer su totalidad.

Scielo corpus

→ Informes, revisiones, opiniones → **NO ESTRUCTURADAS**



Contenido

1. Método
    a. Obtención de los Datos
    b. Word Embeddings
    c. Machine Learning
    d. Deeplearning
2. Resultados
3. Conclusiones y Rumbo

EXPLICACIÓN BASADA EN PROCESAMIENTO DE ESPAÑOL

upf. taln

Grup de Recerca en Tractament Automàtic del Llenguatge Natural

Automatic Natural Language Processing Research Group

Método

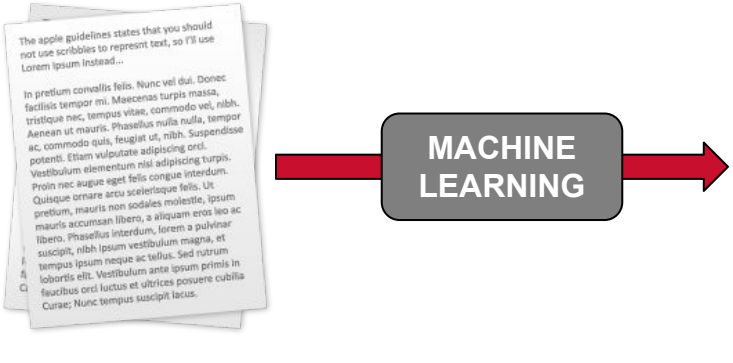
Método

Aprendizaje Automático (Machine Learning)

→ Datos para aprender

→ Datos para evaluar

→ Extracción de Características



INTRODUCCIÓN

MÉTODOS

RESULTADOS

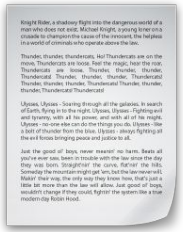
CONCLUSIONES

Método

Obtención de los Datos

Scielo Corpus

Artículos + Exp. Reg.

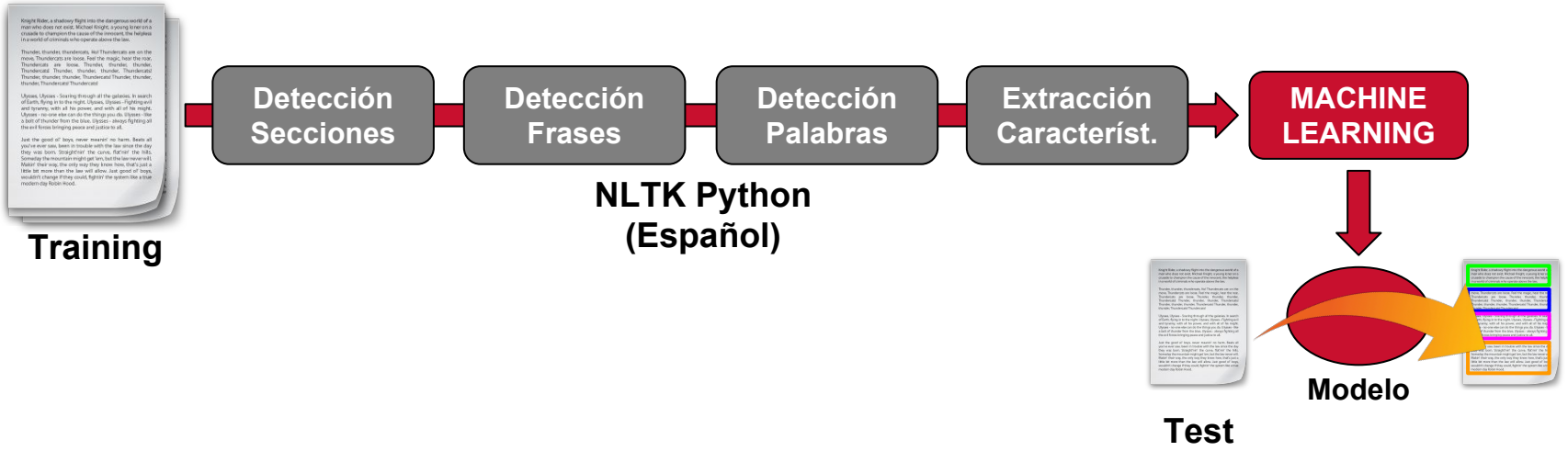


~2500 resúmenes estructurados (ES y EN)

Fundamentos Introducción Propósito Contexto	→	**INTRODUCCIÓN**
Objetivos Finalidad	→	**OBJETIVOS**
Metodología Estudio Escenario Desarrollo	→	**MÉTODOS**
Resultado Resultado	→	**RESULTADOS**
Conclusiones Debate Comentario	→	**CONCLUSIONES**

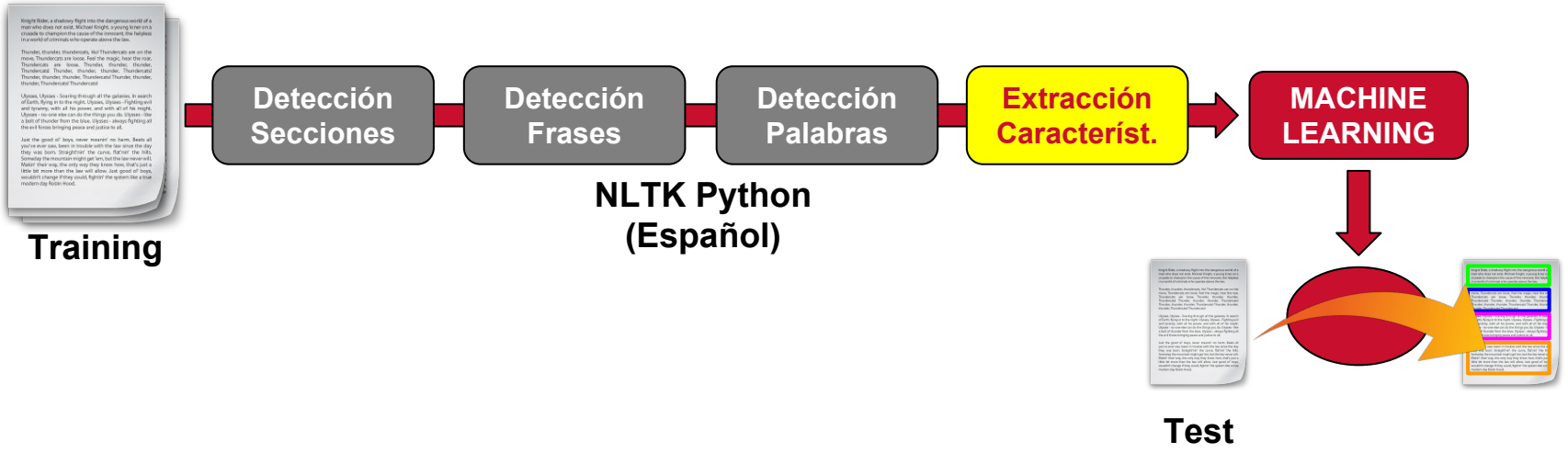
Método

Procesamiento de los Datos



Método

Procesamiento de los Datos



Método

Word Embeddings

→ **Transformaciones** de palabras en vectores (-1 a 1)

→ Capturan información **sintáctica y semántica**

→ **Palabras similares** → Cerca en el espacio del embedding

→ **Deeplearning** y otros sistemas de Machine Learning

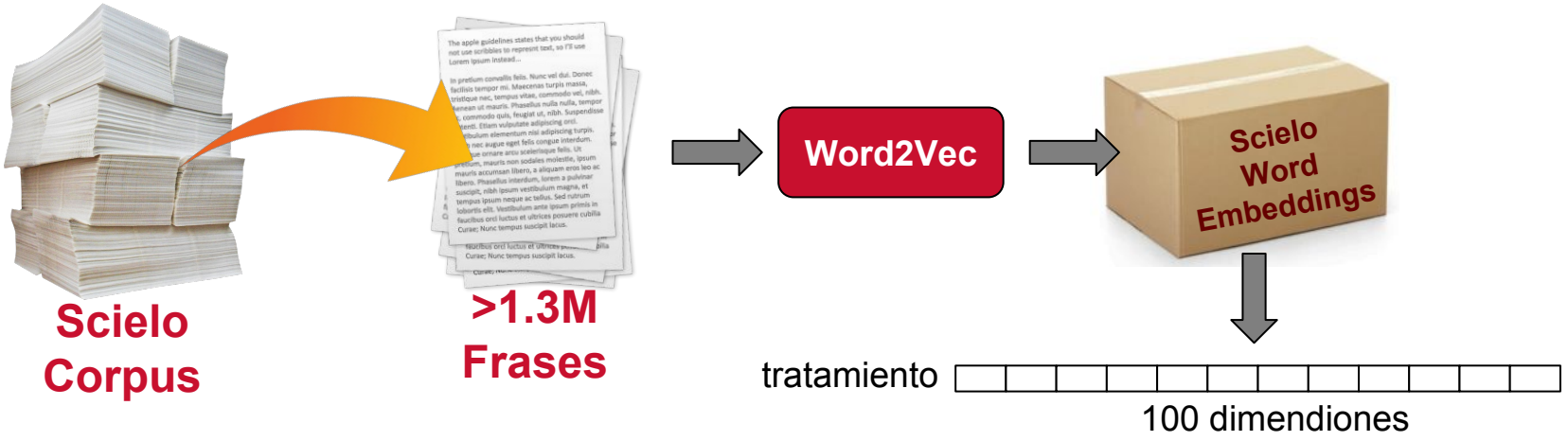
→ Pre-Trained Word Embeddings:

- Google News (EN)
- Wikipedia (Varias Lenguas)
- BalbelNET (Varias Lenguas)
- PubMed & PMC (EN)
- **Biomédico en Español?**

Método

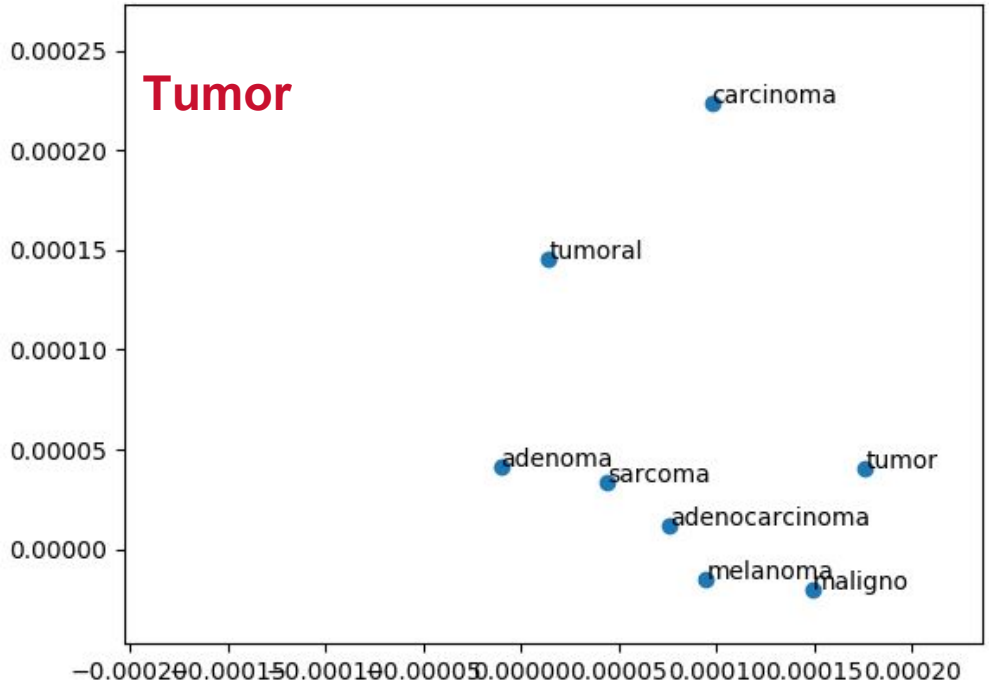
Scielo Word Embeddings

→ Word2Vec (https://code.google.com/archive/p/word2vec/)



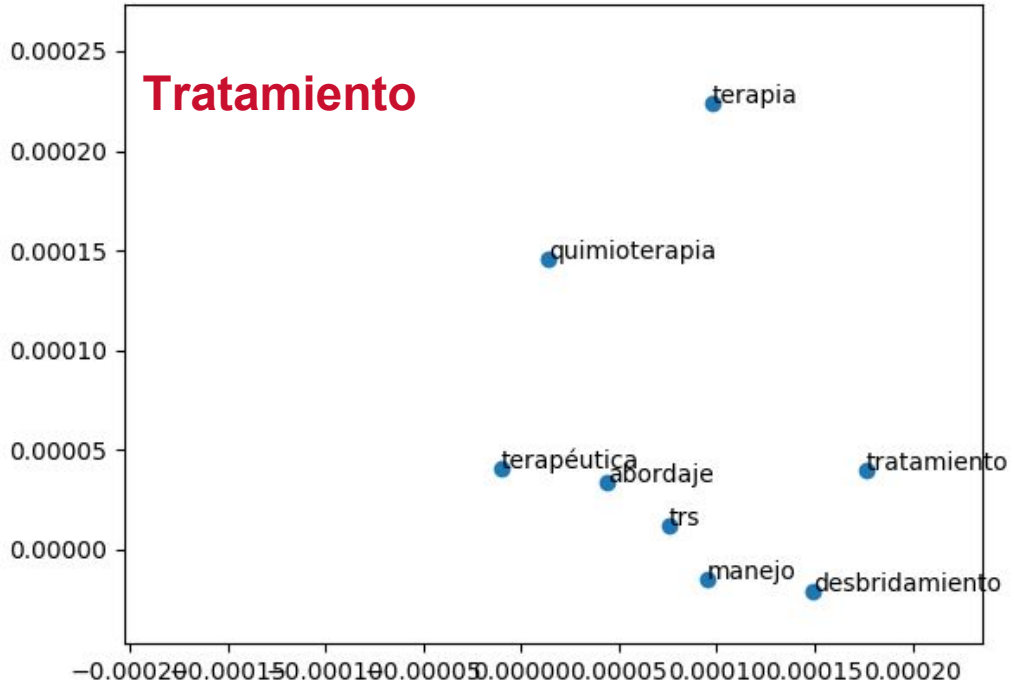
Método

Scielo Word Embeddings



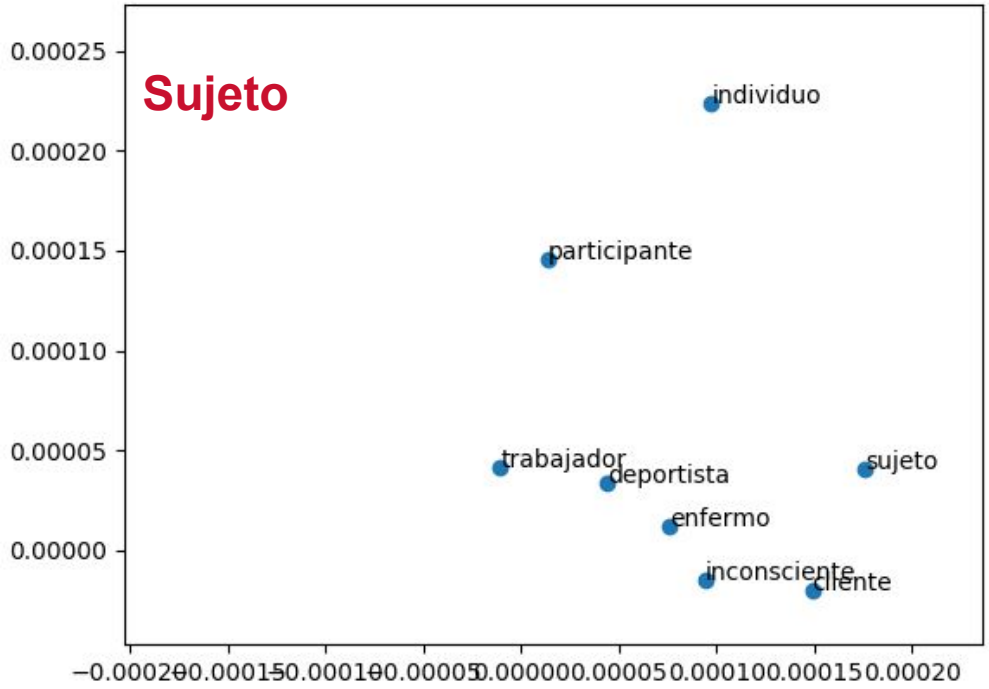
Método

Scielo Word Embeddings



Método

Scielo Word Embeddings



Método

Machine Learning basado en Word Embeddings

→ WEKA

→ Support Vector Machine (SVM)

Frase → Clasificar en su categoría

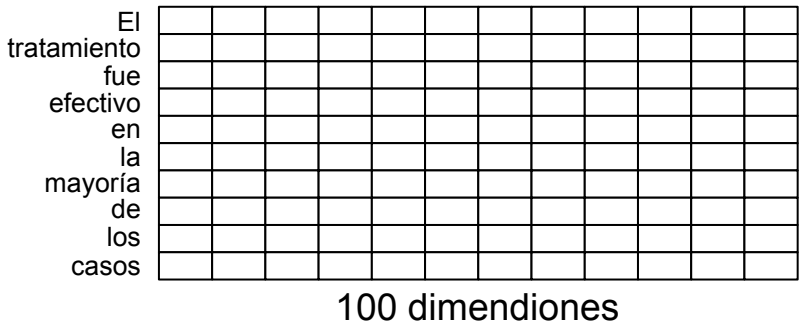
Método

Machine Learning basado en Word Embeddings

→ WEKA

→ Support Vector Machine (SVM)

Frase → Clasificar en su categoría



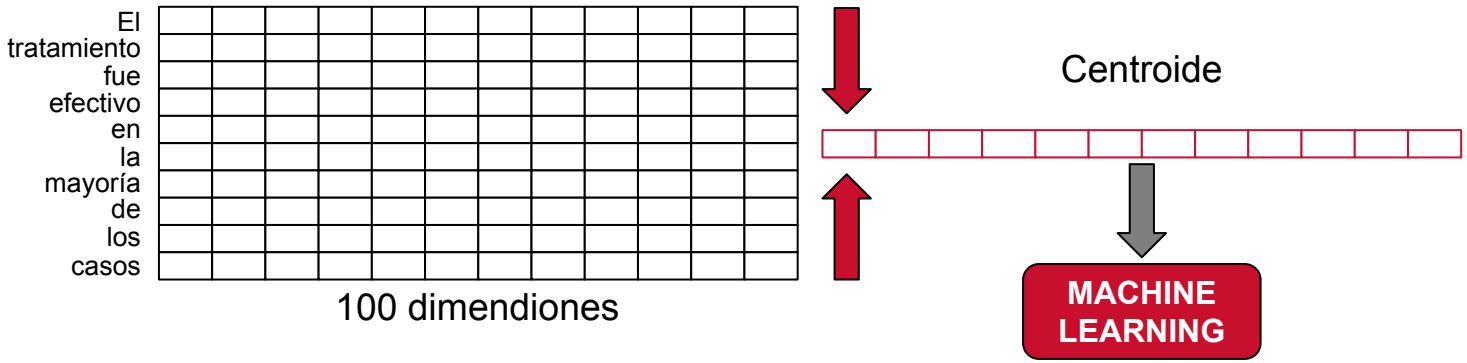
Método

Machine Learning basado en Word Embeddings

→ WEKA

→ Support Vector Machine (SVM)

Frase → Clasificar en su categoría



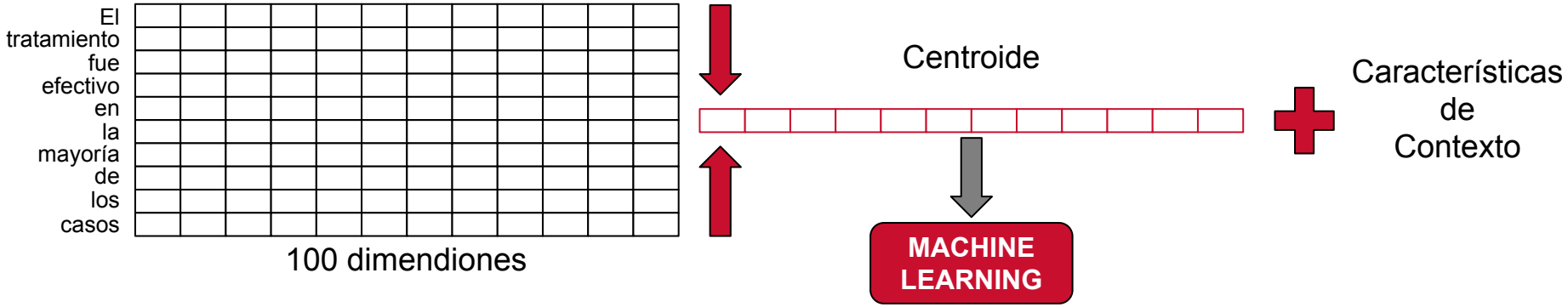
Método

Machine Learning basado en Word Embeddings

→ WEKA

→ Support Vector Machine (SVM)

Frase → Clasificar en su categoría



Método

Deeplearning (Redes Neuronales)

→ Problemas complejos de aprendizaje

→ Aplicadas en **PLN**

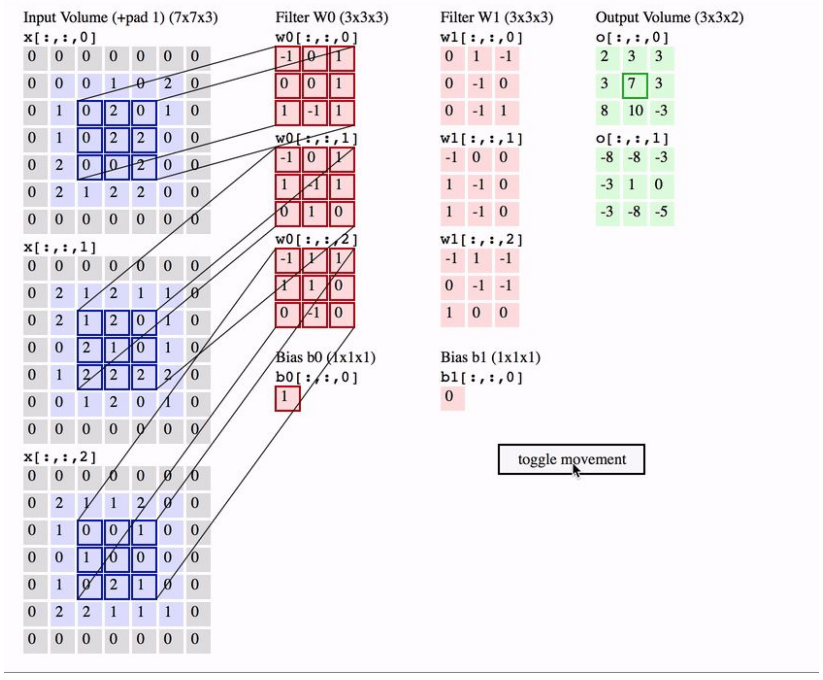
Convolutional Neural Network (CNN)

→ Convoluciones, pooling y fully-connected

→ Procesamiento de Imágen

→ Visión por Computador

→ **PLN**



Método

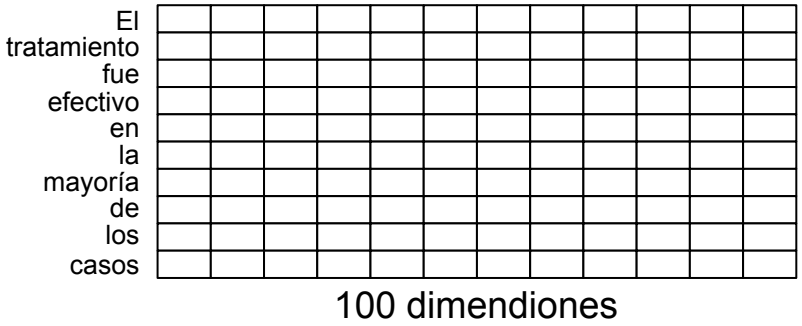
Convolución en PLN

→ Decodificar la frase → Características de más alto nivel

→ Una convolución → aplicar un filtro para obtener un valor en cada “paso”

→ Puede aplicar varios filtros

→ Pueden aplicar filtros de diferente tamaño



Método

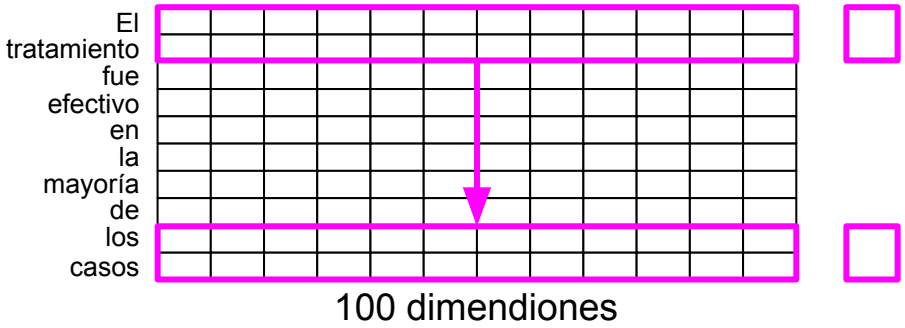
Convolución en PLN

→ Decodificar la frase → Características de más alto nivel

→ Una convolución → aplicar un filtro para obtener un valor en cada “paso”

→ Puede aplicar varios filtros

→ Pueden aplicar filtros de diferente tamaño



Método

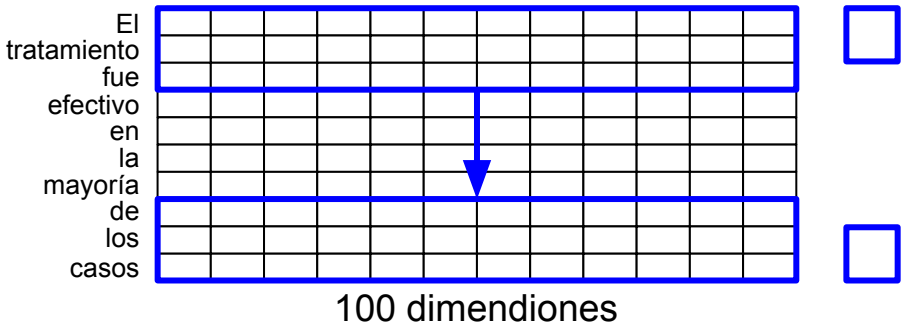
Convolución en PLN

→ Decodificar la frase → Características de más alto nivel

→ Una convolución → aplicar un filtro para obtener un valor en cada “paso”

→ Puede aplicar varios filtros

→ Pueden aplicar filtros de diferente tamaño



Método

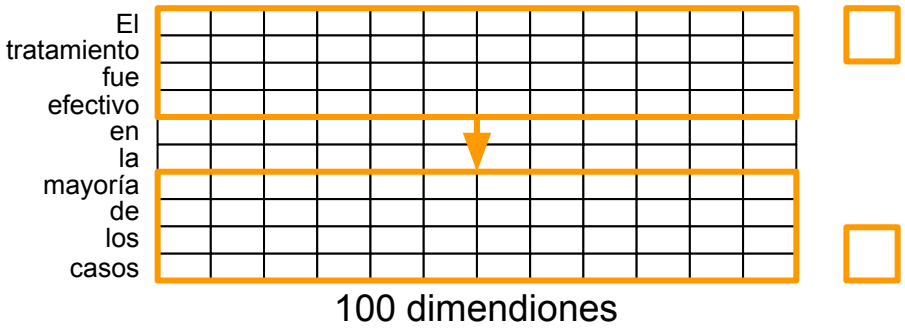
Convolución en PLN

→ Decodificar la frase → Características de más alto nivel

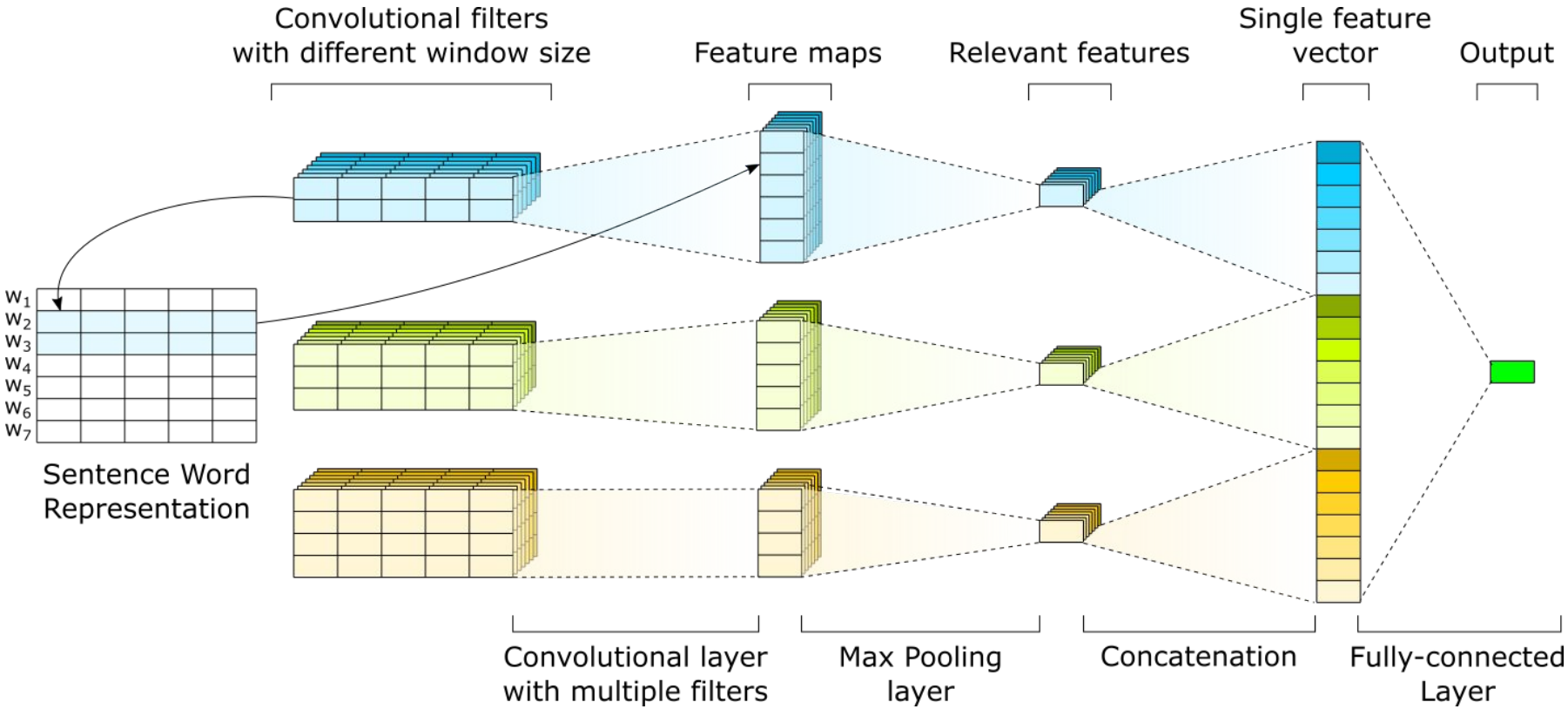
→ Una convolución → aplicar un filtro para obtener un valor en cada “paso”

→ Puede aplicar varios filtros

→ Pueden aplicar filtros de diferente tamaño



Método


Convolutional filters
with different window size
Feature maps
Relevant features
Single feature
vector
Output
w1
w2
w3
w4
w5
w6
w7
Sentence Word
Representation
Convolutional layer
with multiple filters
Max Pooling
layer
Concatenation
Fully-connected
Layer

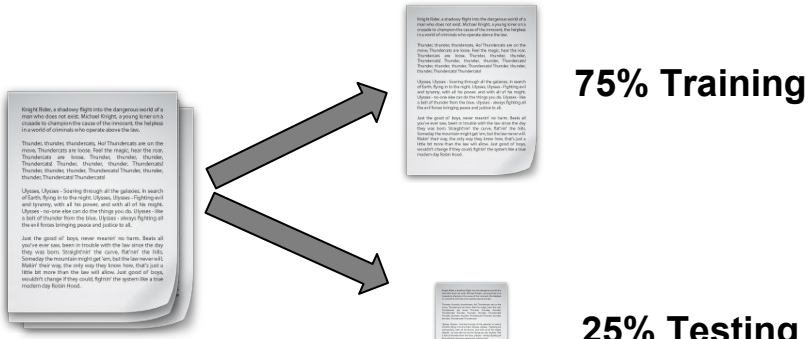
upf. taln

Grup de Recerca en Tractament Automàtic del Llenguatge Natural

Automatic Natural Language Processing Research Group

Resultados

Resultados



~2500
resúmenes
estructurados
(ES)

75% Training

25% Testing

Resultados

Support Vector Machine

→ Solo con Word Embeddings (centroide):

- Precisión: 70.7%
- Recall: 71.1%
- **F-Score: 70.7%**

→ Añadiendo Información de Contexto:

- Precisión: 88.2%
- Recall: 88.3%
- **F-Score: 88.2%**

Resultados

Deeplearning

→ CNNs de 3, 4 y 5 → **2,3,4 y 5**
→ Cada una con 100 Filtros → **>200**
→ 1 Iteración/batch (250) → **~25**
→ 50 Épocas → **> 100**

→ Solo con Word Embeddings:
- Precisión: 69.2%
- Recall: 68.5%
- **F-Score: 68.8%**

Resultados

Ejemplo de Resultado

S0211-57352014000400002

Características sociodemográficas de las personas con conducta acumuladora/trastorno por acumulación (S. de Diógenes) en la ciudad de Madrid: serie de casos

El trastorno de conducta por acumulación origina significativos riesgos para la salud del acumulador y para la Salud Pública, problemas de convivencia en el entorno familiar y vecinal y amenazas para la seguridad motivada por el riesgo de incendios. El presente estudio tiene como objetivo describir las características sociodemográficas de los acumuladores en la ciudad de Madrid. Fueron seleccionados 295 casos de las 1147 solicitudes de intervención por posibles situaciones de insalubridad que entraron desde el 1 de enero de 2009 al 31 de diciembre de 2012 a la Unidad Técnica de Entorno Urbano y Vivienda de Madrid Salud. Los casos cumplían todos o alguno de los criterios diagnósticos del acumulador patológico descritos por Randy O Frost et al en 1993(1). Los acumuladores tienen una edad media de 64.77 años, son hombres en el 55.9% de los casos y el 47% son mayores de 65 años. Son españoles el 95.6%, y pensionistas 65.1%. En 129 casos se retiraron un total de 260.346 kgrs de basura y enseres. El trastorno por acumulación, aunque poco frecuente, provoca un grave problema a la persona que lo sufre, un riesgo para la Salud Pública y para la seguridad y un alto coste por la elevada cantidad de recursos que consume su atención y resolución. Nuestro trabajo añade argumentos para considerar el trastorno de acumulación como una entidad independiente como ya se clasifica en la DSM-5 de mayo de 2013.

Resultados

Ejemplo de Resultado

S0211-57352014000400002

Características sociodemográficas de las personas con conducta acumuladora/trastorno por acumulación (S. de Diógenes) en la ciudad de Madrid: serie de casos

INTRODUCCIÓN:
El trastorno de conducta por acumulación origina significativos riesgos para la salud del acumulador y para la Salud Pública, problemas de convivencia en el entorno familiar y vecinal y amenazas para la seguridad motivada por el riesgo de incendios.

OBJETIVO:
El presente estudio tiene como objetivo describir las características sociodemográficas de los acumuladores en la ciudad de Madrid.

RESULTADOS:
Fueron seleccionados 295 casos de las 1147 solicitudes de intervención por posibles situaciones de insalubridad que entraron desde el 1 de enero de 2009 al 31 de diciembre de 2012 a la Unidad Técnica de Entorno Urbano y Vivienda de Madrid Salud. Los casos cumplían todos o alguno de los criterios diagnósticos del acumulador patológico descritos por Randy O Frost et al en 1993(1). Los acumuladores tienen una edad media de 64.77 años, son hombres en el 55.9% de los casos y el 47% son mayores de 65 años. Son españoles el 95.6%, y pensionistas 65.1%. En 129 casos se retiraron un total de 260.346 kgrs de basura y enseres.

CONCLUSION:
El trastorno por acumulación, aunque poco frecuente, provoca un grave problema a la persona que lo sufre, un riesgo para la Salud Pública y para la seguridad y un alto coste por la elevada cantidad de recursos que consume su atención y resolución. Nuestro trabajo añade argumentos para considerar el trastorno de acumulación como una entidad independiente como ya se clasifica en la DSM-5 de mayo de 2013.

Resultados

Ejemplo de Resultado

S1135-57272008000400004

Tendencias de la mortalidad por enfermedades cardiovasculares en Andalucía entre 1975 y 2004</title>

Las enfermedades cardiovasculares están entre las primeras causas de mortalidad en los países industrializados. El objetivo de este trabajo es conocer las tendencias de la mortalidad por enfermedades isquémicas del corazón (EIC) y enfermedades cerebrovasculares (ECV) en Andalucía entre 1975 y 2004. Con las defunciones por EIC y ECV de las estadísticas oficiales y las correspondientes poblaciones se calcularon las tasas brutas (TB), ajustadas por edad (TS) y Años Potenciales de Vida Pedidos (APVP). Para cuantificar las tendencias y sus puntos de cambio se realizó el análisis con regresión joinpoint. En las mujeres las muertes por EIC pasaron de 2.086 en el año 1975 a 3.336 en el 2004, la TS descendió de 74,29 defunciones por 100.000 mujeres a 50,94, y los APVP pasaron de 173,65 años por 100.000 mujeres a 90,56. En los hombres el número de defunciones pasó de 2.854 defunciones en el 1975 a 4085 en el año 2004, la TS descendió de 147, 67 defunciones por 100.000 hombres a 104,96. Igual comportamiento tuvieron los APVP entre el primer y último año de la serie con unos valores de 716,46 años por 100.000 hombres y 460,04. Para las ECV en las mujeres las defunciones en números absolutos pasaron de 4.712 a 4.221, la TS pasó de 166,00 defunciones por 100.000 mujeres a 62,08 y los APVP de 338,08 años por 100.000 mujeres a 87,63. En los hombres el número de defunciones pasó de 3.714 a 2.951, la TS de 206,88 defunciones por 100.000 en 1975 a 76,12 en 2004, y la tasa de APVP de 533,12 años por 100.00 hombres a 182,38. La tendencia de la mortalidad por EIC no fue constante ni en las mujeres ni en los hombres aunque en ambos siempre ha sido descendente, siendo el descenso estadísticamente significativo. La disminución de las ECV fue tan importante que tanto los números absolutos como las tasas brutas fueron menores en los últimos años de la serie estudiados que en los primeros, a pesar el envejecimiento de la población andaluza.

Resultados

Ejemplo de Resultado

S1135-57272008000400004

Tendencias de la mortalidad por enfermedades cardiovasculares en Andalucía entre 1975 y 2004</title>

INTRODUCTION:
Las enfermedades cardiovasculares están entre las primeras causas de mortalidad en los países industrializados. El objetivo de este trabajo es conocer las tendencias de la mortalidad por enfermedades isquémicas del corazón (EIC) y enfermedades cerebrovasculares (ECV) en Andalucía entre 1975 y 2004.

MÉTODOS:
Con las defunciones por EIC y ECV de las estadísticas oficiales y las correspondientes poblaciones se calcularon las tasas brutas (TB), ajustadas por edad (TS) y Años Potenciales de Vida Pedidos (APVP). Para cuantificar las tendencias y sus puntos de cambio se realizó el análisis con regresión joinpoint.

RESULTADOS:
En las mujeres las muertes por EIC pasaron de 2.086 en el año 1975 a 3.336 en el 2004, la TS descendió de 74,29 defunciones por 100.000 mujeres a 50,94, y los APVP pasaron de 173,65 años por 100.000 mujeres a 90,56. En los hombres el número de defunciones pasó de 2.854 defunciones en el 1975 a 4085 en el año 2004, la TS descendió de 147, 67 defunciones por 100.000 hombres a 104,96. Igual comportamiento tuvieron los APVP entre el primer y último año de la serie con unos valores de 716,46 años por 100.000 hombres y 460,04. Para las ECV en las mujeres las defunciones en números absolutos pasaron de 4.712 a 4.221, la TS pasó de 166,00 defunciones por 100.000 mujeres a 62,08 y los APVP de 338,08 años por 100.000 mujeres a 87,63. En los hombres el número de defunciones pasó de 3.714 a 2.951, la TS de 206,88 defunciones por 100.000 en 1975 a 76,12 en 2004, y la tasa de APVP de 533,12 años por 100.00 hombres a 182,38.

CONCLUSIONES:
La tendencia de la mortalidad por EIC no fue constante ni en las mujeres ni en los hombres aunque en ambos siempre ha sido descendente, siendo el descenso estadísticamente significativo. La disminución de las ECV fue tan importante que tanto los números absolutos como las tasas brutas fueron menores en los últimos años de la serie estudiados que en los primeros, a pesar el envejecimiento de la población andaluza.

upf. taln

Grup de Recerca en Tractament Automàtic del Llenguatge Natural

Automatic Natural Language Processing Research Group

Conclusiones y Rumbo

Conclusiones y Rumbo

Realizado un sistema de estructuración de resúmenes para Español e Inglés

Se han probado diferentes técnicas de **Machine Learning**

→ **Support Vector Machine**

→ **Deeplearning con Convolución → TUNNING**

Se han obtenido resultados cercanos al 90% de F-score.

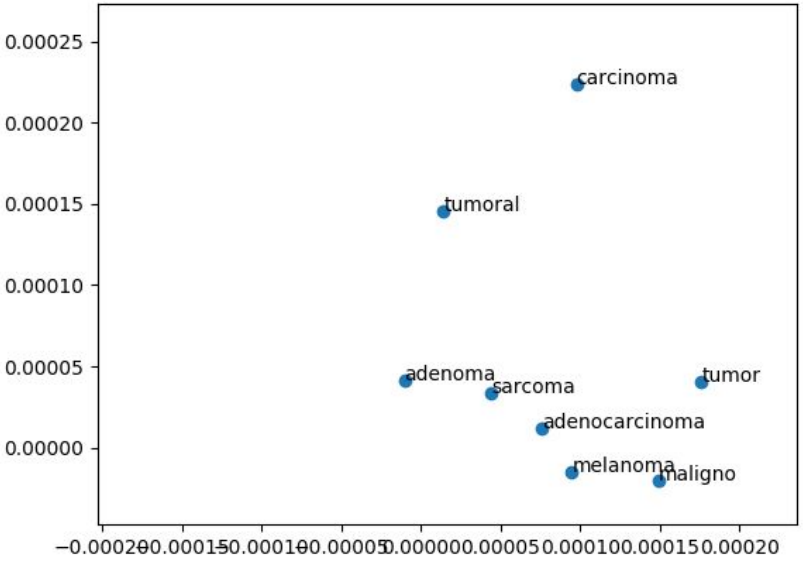
→ Mayoría de fallos: **INTRODUCCIÓN** y **OBJETIVO**

Conclusiones y Rumbo

Contribuido con **NUEVOS RECURSOS**

→ Scielo Word Embeddings

- Campo Biomédico
- Español
- Binario y Texto



Conclusiones y Rumbo

Contribuido con **NUEVOS RECURSOS**

→ Scielo Word Embeddings
- Campo Biomédico
- Español
- Binario y Texto

→ ~2500 resúmenes estructurados (Español)

→ ~2500 resúmenes estructurados (Inglés)

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>

<corpus>
<article id="S0211-57352009000200005">
<title>Presencia física de profesionales de Salud Mental en un Centro de Atención Primaria como forma alternativa de coordinación: Una experiencia piloto</title>
<abstract>
<section label="INTRODUCTION"> Problemas de coordinación entre los niveles de atención primaria y salud mental sugieren la necesidad de explorar nuevas vías que faciliten la comunicación entre ambos niveles asistenciales. </section>
<section label="METHODS"> Un Psiquiatra y un Psicólogo Clínico de Alcalá de Henares (Área 3 de la Comunidad Autónoma de Madrid) se han desplazado un día a la semana a un centro de atención primaria (centro experimental) para evaluar pacientes derivados por los médicos de ese centro y contrastar directamente con ellos sus valoraciones. Después de 15 meses de funcionamiento, se administró a los médicos un cuestionario acerca de su satisfacción con su relación con salud mental. Sus respuestas se compararon con las de médicos de otro centro en el que no tenía lugar la experiencia (centro control). </section>
<section label="RESULTS"> los médicos del centro experimental valoraban mejor la información recibida por salud mental y la cantidad de contactos que tenían con ellos, percibían al equipo de salud mental como más disponible, recibían más cursos específicos, se habían coordinado más con salud mental y, en general, valoraban mejor la coordinación atención primaria - salud mental que los del centro control. </section>
<section label="CONCLUSIONS"> La experiencia muestra cómo es posible una forma alternativa de coordinación con la que los médicos de primaria se muestran altamente satisfechos. Se plantean algunas cuestiones metodológicas y posibles vías para continuar futuros trabajos.
</section>
</abstract>
</article>

<article id="S0211-57352010000200002">
<title>Psicosis y diferencias sociales: Comparando la prevalencia de las psicosis en dos medios urbanos diferenciados</title>
<abstract>
<section label="OBJECTIVE"> Contribuir a la reflexión sobre la etiología y/o los factores de riesgo para las psicosis comparando la prevalencia en población general y población de riesgo de la esquizofrenia y otras psicosis en dos barrios de Barcelona (España). </section>
<section label="METHODS"> Nuestras aportaciones en este trabajo se apoyan sobre todo en un estudio descriptivo transversal de todos los pacientes con psicopatología detectados en la USM de Sant Martí-La Mina: un territorio geodemográfica y asistencialmente delimitado formado por 5 Áreas Básicas de Salud (103.615 habitantes. </section>
<section label="RESULTS"> Sobre un total de 21.536 pacientes con registro de casos abierto desde el año 1982 hasta el año 2000, se halló que 838 cumplían los criterios restrictivos para ser diagnosticados como "esquizofrénicos" (N=476) o "afectados por otras psicosis" (N=362). Sin embargo, las prevalencias de esquizofrenia y otras psicosis en el barrio sujeto a más factores de riesgo psicosociales eran alrededor de 2 veces mayores que las encontradas en el barrio colindante por el mismo equipo y en el mismo período temporal. </section>
```

Conclusiones y Rumbo

RUMBO

Estructuración de Textos Biomédicos:

→ **Historias Clínicas:** Datos paciente, antecedentes, medicación…

→ Conversiones de artículos en **PDF** a XML

Potenciar Scielo Word Embeddings:

→ **Enriqueciendo** con más Fuentes (>1.3M)

Clasificación de frases

→ Frases relevantes → Extracción de resúmenes

→ Detección de Patrones en Redes Sociales

Gracias por su Atención