Extracción automática de medicamentos en expedientes médicos electrónicos

Salvador Barcenas Valladolid

SEDESA - Problema a solucionar

Exhaustivo trabajo manual por parte de las personas de sedesa en la identificación de medicamentos, signos vitales, síntomas, etc, de pacientes con covid.

GOBIERNO DE LA
CIUDAD DE MÉXICO

Formato de la nota

Dificil de leer, ¿cierto?

Ahora imagínate pasar esto a una tabla en Excel, ¿cuánto trabajo crees que se necesite para miles de líneas como esta?

Objetivo:

1) Preprocesamiento y estructuración de los datos
2) Identificación de medicamentos
3) Poner el modelo en producción

Notas médicas pacientes SARS-CoV-2

1) Consta de **98** notas médicas
2) Errores en los atributos XML en **7** notas

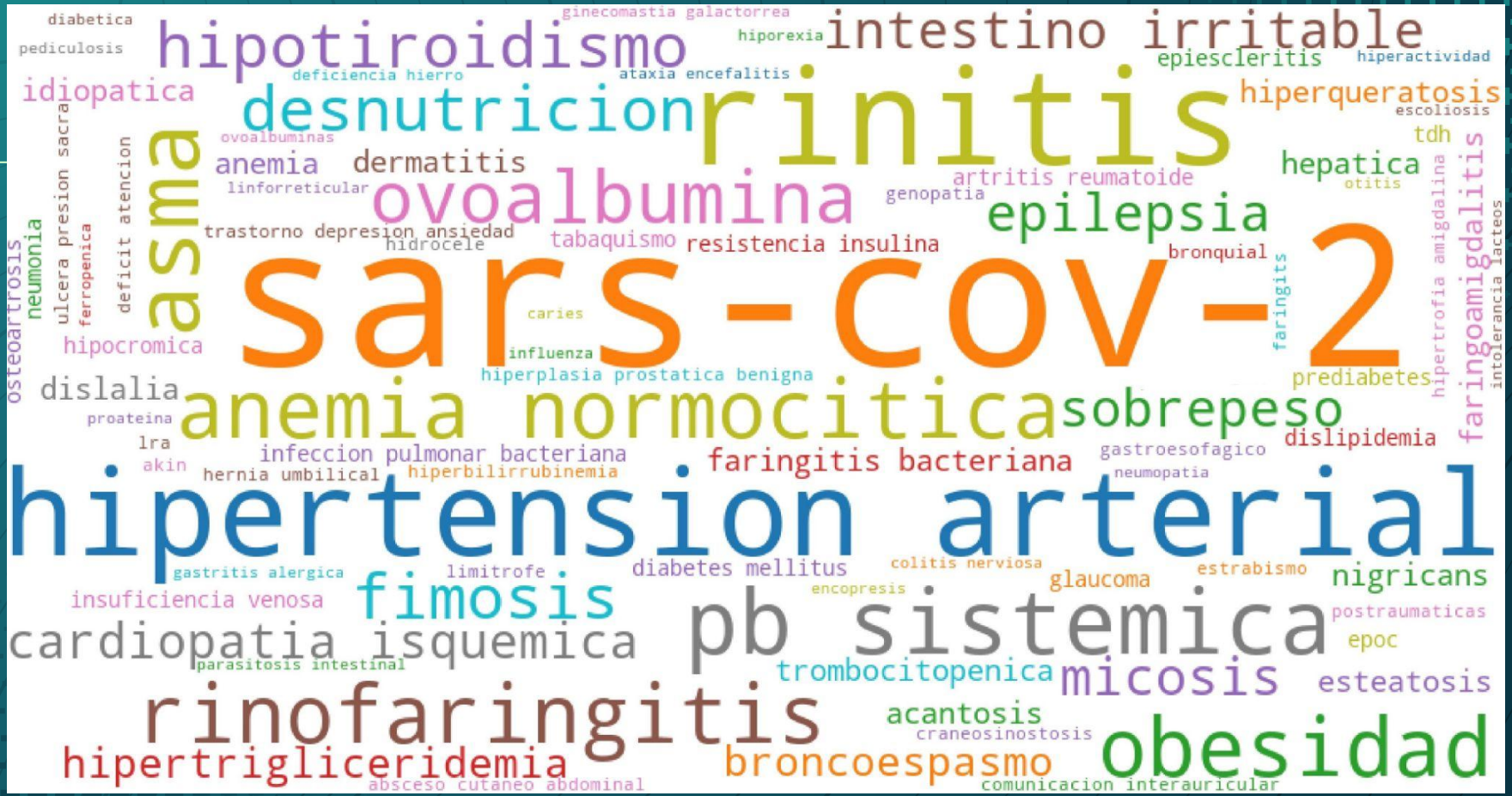
Ejemplo:

```
<assignedCustodian>
  <representedCustodianOrgganization>
    <id root="2.16.840.1.113883.2.38.1.10" />
    <name>HOSPITAL ABC</name>
  </representedCustodianOrganization>
</assignedCustodian>
```

Notas médicas otros pacientes

1) Consta de **100** notas médicas
2) Errores en los atributos XML en **6** notas
3) Notas médicas provienen de pacientes con distintas enfermedades

Ejemplos: (rinitis, asma, otitis, obesidad, pediculosis, dermatitis)


diabetica
pediculosis
ginecomastia galactorrea
hipotiroidismo
hiporexia
intestino irritable
epiescleritis
hiperactividad
deficiencia hierro
ataxia encefalitis
idiopatica
desnutricion
rinitis
hiperqueratosis
escoliosis
ovoalbuminas
tdh
asma
anemia
dermatitis
hepatica
artritis reumatoide
otitis
linforreticular
ovoalbumina
genopatia
epilepsia
ulcera presion sacra
deficit atencion
neumonia
trastorno depresion ansiedad
hidrocele
tabaquismo
resistencia insulina
bronquial
sars-cov-2
osteoartrosis
ferropenica
caries
faringitis
hipertrofia amigdalina
faringoamigdalitis
intolerancia lacteos
hipocromica
influenza
hiperplasia prostatica benigna
prediabetes
dislalia
anemia normocitica
sobrepeso
proateina
lra
dislipidemia
infeccion pulmonar bacteriana
faringitis bacteriana
gastroesofagico
akin
hernia umbilical
hiperbilirrubinemia
neumopatia
hipertension arterial
gastritis alergica
limitrofe
diabetes mellitus
colitis nerviosa
glaucoma
estrabismo
nigricans
insuficiencia venosa
fimosis
encopresis
pb sistemica
postraumaticas
cardiopatia isquemica
epoc
parasitosis intestinal
trombocitopenica
micosis
esteatosis
rinofaringitis
acantosis
craneosinostosis
obesidad
hipertrigliceridemia
broncoespasmo
absceso cutaneo abdominal
comunicacion interauricular

Limpieza de datos

Errores de codificación:

Datos guardados en otro tipo de codificación distinta a UTF-8.

Ejemplos: (“AnÃ¡lisis”, Análisis”), (“hipÃ³tesis”, “hipótesis”), (“aÃ±o”, “año”)

Limpieza de datos

Sustitución de caracteres alfanuméricos:

Existen caracteres que no aportan información a nuestro análisis.

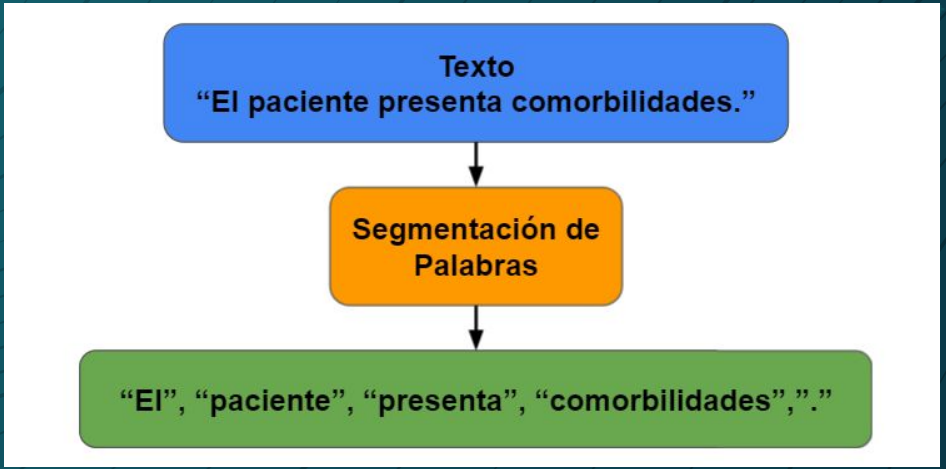
Ejemplos:

Caracteres alfabéticos	(A,B,C,...,Z) (a,b,c,...,z)
Caracteres numéricos	(1,2,...,9,0)
Caracteres especiales	(+ , - , * , / , ^ , . , ; , < , >, $,...)

Tokenización y etiquetado de datos

Tokenización:

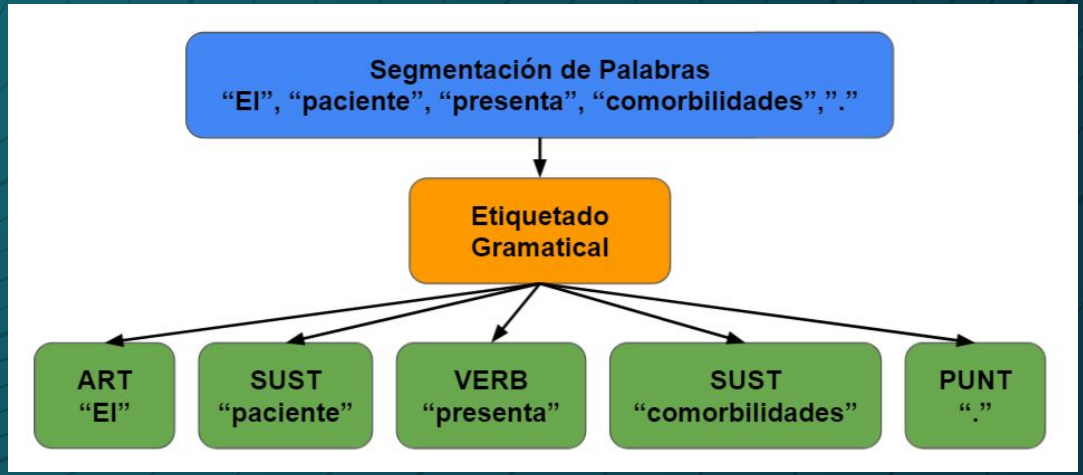
La tokenización es un proceso importante dentro del análisis de textos.



Tokenización y etiquetado de datos

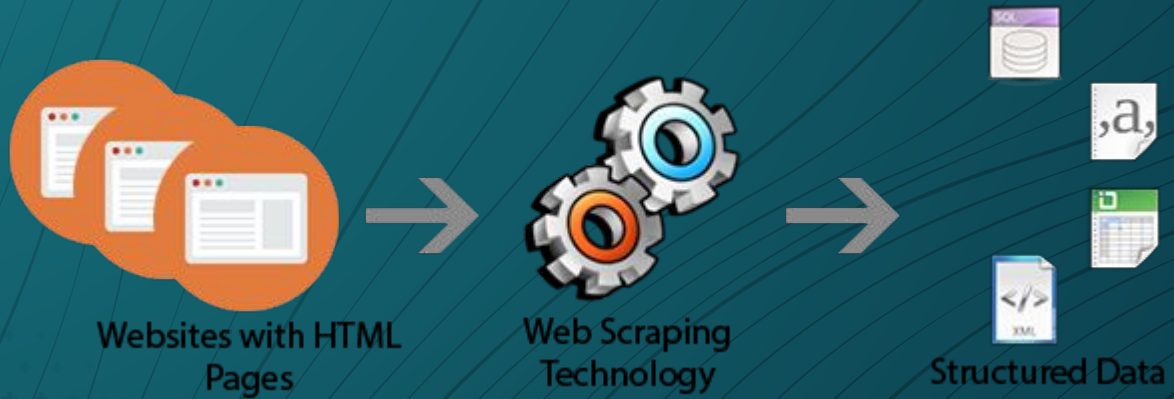
Etiquetado:

También conocido como POS-Tagging, nos ayuda a identificar con más facilidad las palabras de interés.



Primer acercamiento

Armar un listado de medicamentos, con ayuda de **Web Scrapping**[1].



[1] https://www.gob.mx/cms/uploads/attachment/file/441323/Listado_de_medicamentos_de_Referencia_.pdf

Problemas:

Errores ortográficos: Esto nos limita la identificación de medicamentos.

Ejemplo:("buthilhioscina","butilhioscina")

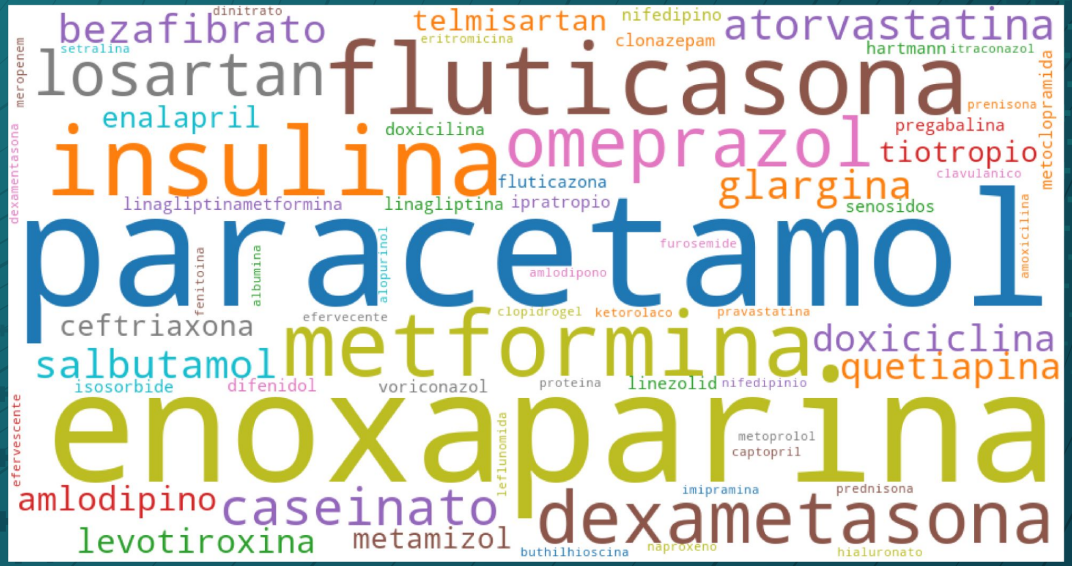
Solución: **Distancia de Levenshtein**

Es el número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra.

Por ejemplo, la distancia de Levenshtein entre "casa" y "calle" es de 3.

1. casa → cala (sustitución de 's' por 'l')
2. cala → calla (inserción de 'l' entre 'l' y 'a')
3. calla → calle (sustitución de 'a' por 'e')

Medicamentos/compuestos-frecuencia



Nota:

Este es un conteo de la aparición de los medicamentos en las notas, tomando en cuenta de que un medicamentos puede nombrarse dos o más veces en una misma nota.

Medicamentos/compuestos-frecuencia

EDAD	SEXO	Medicamentos_suministrados
58	Mujer	paracetamol,enoxaparina,omeprazol
53	Mujer	metformina,levotiroxina,dexametasona,paracetam...
56	Mujer	paracetamol,losartan
37	Mujer	paracetamol,enoxaparina
28	Mujer	paracetamol,enoxaparina
...	...	...
70	Hombre	dexmedetomidina,quetiapina,fluconazol,linezolid
23	Mujer	acido,paracetamol
61	Hombre	losartan,metformina,lorazepam,azitromicina,par...
72	Mujer	enalapril,metformina,metformina,paracetamol,en...
33	Mujer	paracetamol,enoxaparina,fluticasona,salbutamol

paracetamol	110
enoxaparina	94
metformina	40
dexametasona	33
losartan	25
fluticasona	24
omeprazol	20
calcio	20
acido	18
potasio	11
doxiciclina	9
levotiroxina	9
amlodipino	8
quetiapina	7
atorvastatina	7
salbutamol	7
enalapril	6
bromuro	5
bicarbonato	5
telmisartan	5

Resultados:

['paracetamol,enoxaparina,imipramina,tableta,dexametasona'

'paracetamol 500 mg,enoxaparina 40 mg,imipramina 25 mg,dexametasona 8mg']

****

['paracetamol,enoxaparina,enalapril,insulina,fluticasona,inhalado'

'paracetamol 1 gr,enoxaparina 40 mg,enalapril 10 mg,fluticasona 2 disparos']

****

Compuesto mal escrito	Compuesto bien escrito
nifedipinio	nifedipino
dexamentasona	dexametasona
doxicilina	doxiciclina
fluticazona	fluticasona
prenisona	prednisona
setralina	sertralina
tablletas	tabletas

Modelo Final

Nota médica (Formato XML) → Preprocesamiento

Preprocesamiento → Modelo

Modelo → **Medicamentos suministrados.**

Resultados:

- Preprocesamiento automático.
- Identificación de medicamentos con su respectiva dosis.
- Apoyamos un tema que actualmente requiere de mucha ayuda.
- No toda la tecnología se utiliza para generar mayor valor a los negocios, sino también para ayudar a la humanidad