Content In-context: Enriquecimiento automático de información para contextualización de noticias

Camilo Restrepo Arango

Asesora: PhD. Claudia Jiménez Guarín
Maestría en Ingeniería de Sistemas y Computación
Ingeniería de Información - Departamento de Ingeniería de Sistemas y Computación

10 de noviembre de 2016

Agenda

Agenda

Motivación

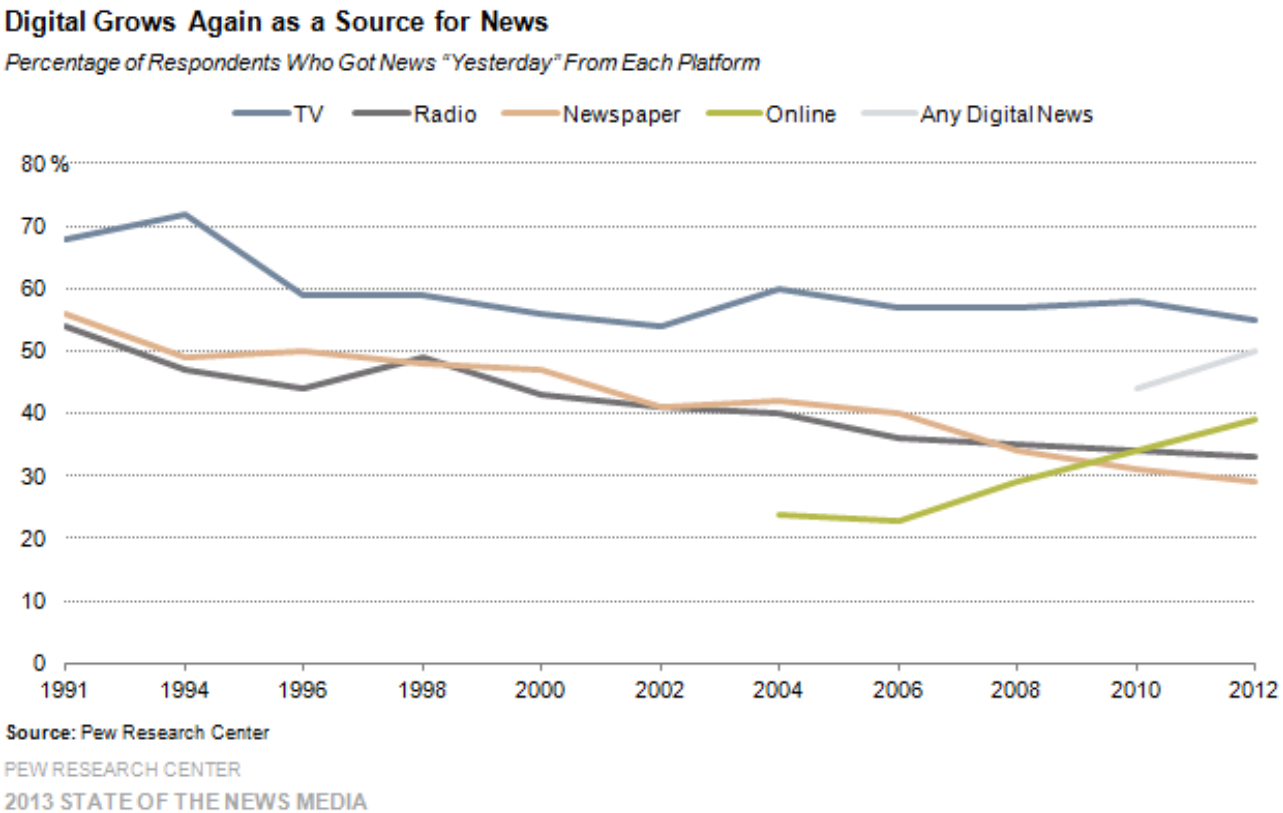


Figura 1: Consumo de noticias en Estados Unidos por medio de publicación.

Motivación

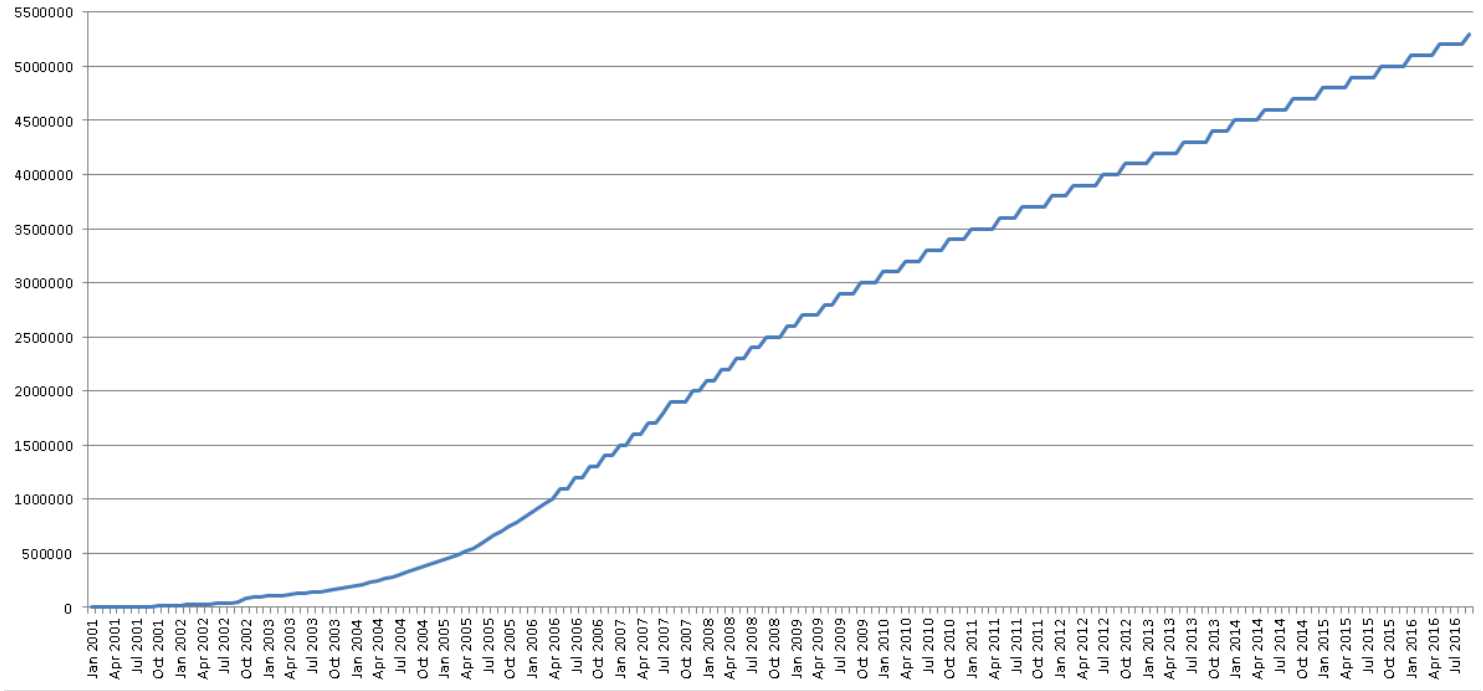


Figura 2: Evolución de la cantidad de articulos de Wikipedia.

Agenda

Retos

- Recursos y esfuerzo para hacer análisis detallados de las noticias.

Retos

- Recursos y esfuerzo para hacer análisis detallados de las noticias.
- Buscar y filtrar información.

Retos

- Recursos y esfuerzo para hacer análisis detallados de las noticias.
- Buscar y filtrar información.
- Entidades mencionadas.

Retos

- Recursos y esfuerzo para hacer análisis detallados de las noticias.
- Buscar y filtrar información.
- Entidades mencionadas.
- Relaciones dinámicas y no explicitas.

Objetivo general

Proponer una solución visualmente atractiva que permite construir contexto de forma automática por medio de un corpus documental para enriquecer el contenido disponible.

Objetivos específicos

- Proponer una arquitectura escalable de una solución de análisis de contenido utilizando técnicas de procesamiento de lenguaje natural.

Objetivos específicos

- Proponer una arquitectura escalable de una solución de análisis de contenido utilizando técnicas de procesamiento de lenguaje natural.
- Comparar modelos de Deep Learning y Machine Learning.

Objetivos específicos

- Proponer una arquitectura escalable de una solución de análisis de contenido utilizando técnicas de procesamiento de lenguaje natural.
- Comparar modelos de Deep Learning y Machine Learning.
- Relacionar contenido textual para encontrar relaciones significativas y enriquecer la información contenida en un corpus documental.

Objetivos específicos

- Proponer una arquitectura escalable de una solución de análisis de contenido utilizando técnicas de procesamiento de lenguaje natural.
- Comparar modelos de Deep Learning y Machine Learning.
- Relacionar contenido textual para encontrar relaciones significativas y enriquecer la información contenida en un corpus documental.
- Proponer una visualización apropiada para el usuario objetivo.

Objetivos específicos

- Proponer una arquitectura escalable de una solución de análisis de contenido utilizando técnicas de procesamiento de lenguaje natural.
- Comparar modelos de Deep Learning y Machine Learning.
- Relacionar contenido textual para encontrar relaciones significativas y enriquecer la información contenida en un corpus documental.
- Proponer una visualización apropiada para el usuario objetivo.
- Realizar una prueba de concepto de la solución desarrollada utilizando datos reales.

Agenda

Soluciones de análisis de noticias

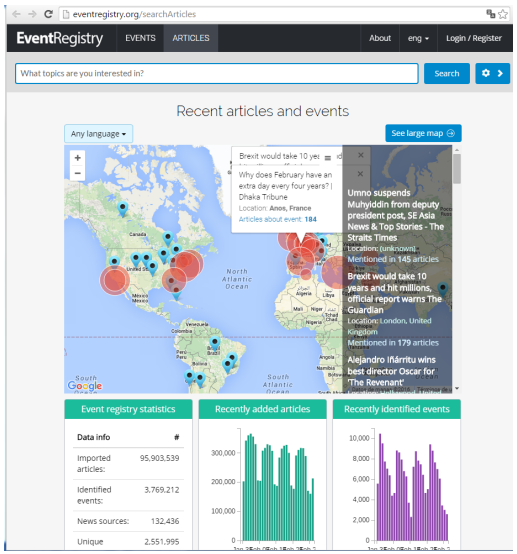


Figura 3: EventRiver

Soluciones de análisis de noticias

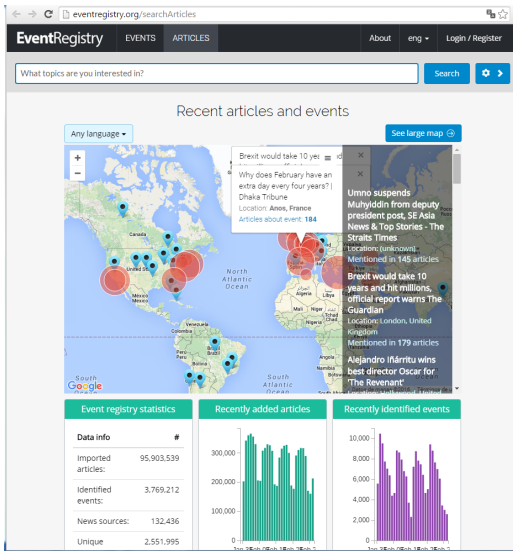


Figura 3: EventRiver

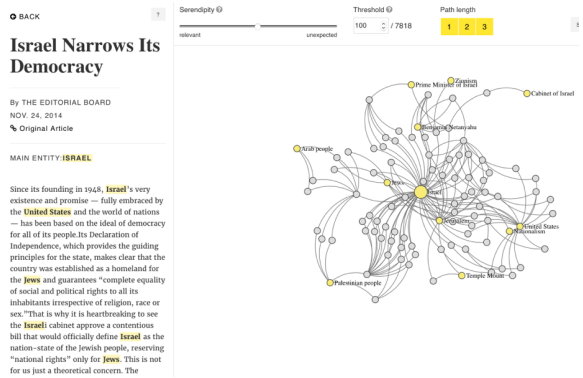


Figura 4: DaCena

Soluciones de análisis de noticias

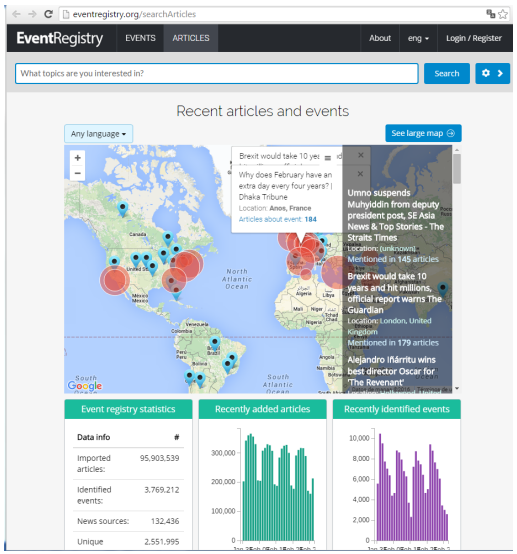


Figura 3: EventRiver

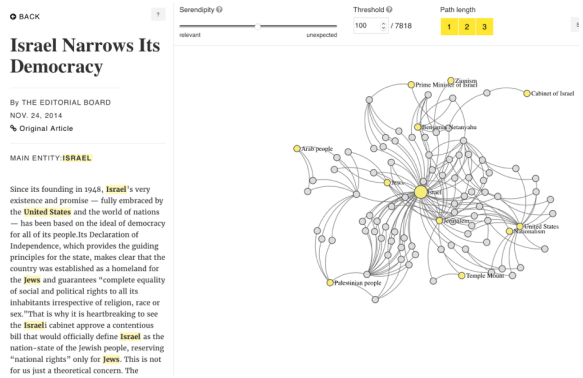


Figura 4: DaCena

Figura 5: Congreso Visible

Pasos de análisis de texto y enriquecimiento

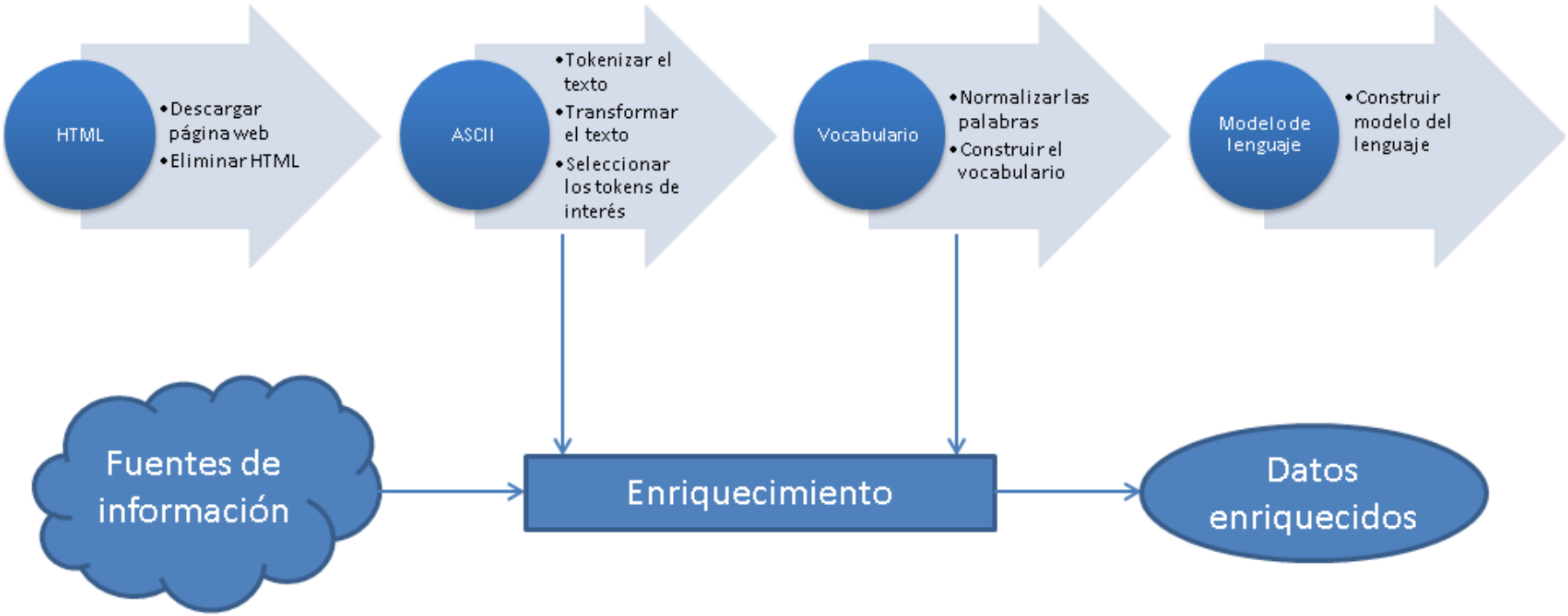


Figura 6: Flujo de procesamiento de texto.

Deep learning vs Machine learning

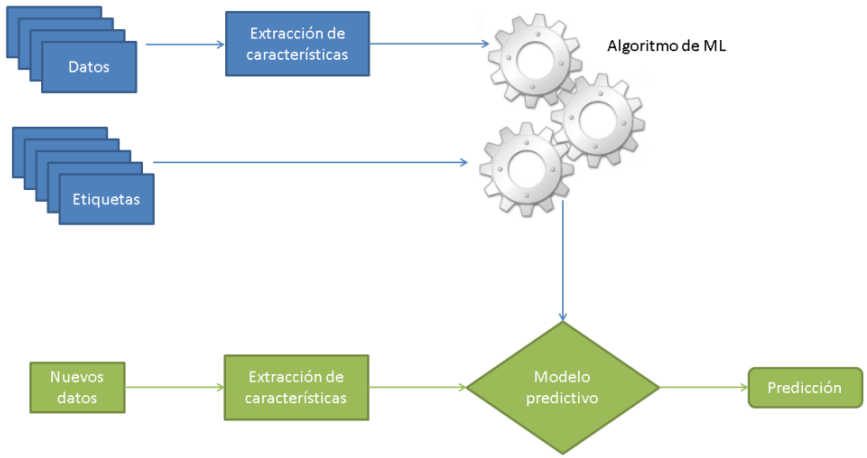


Figura 7: Flujo de trabajo de Machine learning.

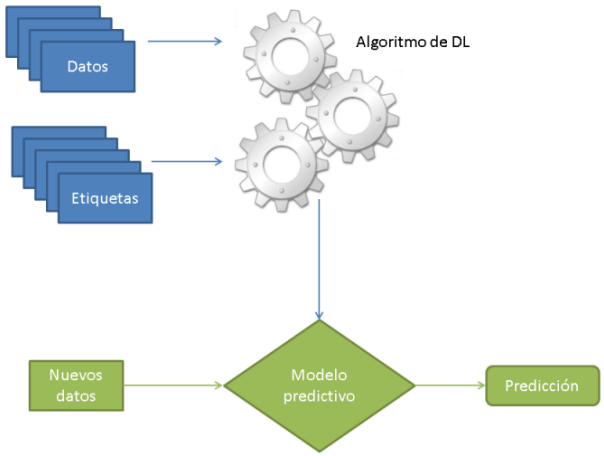


Figura 8: Flujo de trabajo de Deep learning.

Agenda

Aportes principales

Aportes principales

- Procesamiento de texto en español.

Aportes principales

- Procesamiento de texto en español.
- Corpus textual no controlado.

Aportes principales

- Procesamiento de texto en español.
- Corpus textual no controlado.
- Análisis de datos nuevos.

Aportes principales

- Procesamiento de texto en español.
- Corpus textual no controlado.
- Análisis de datos nuevos.

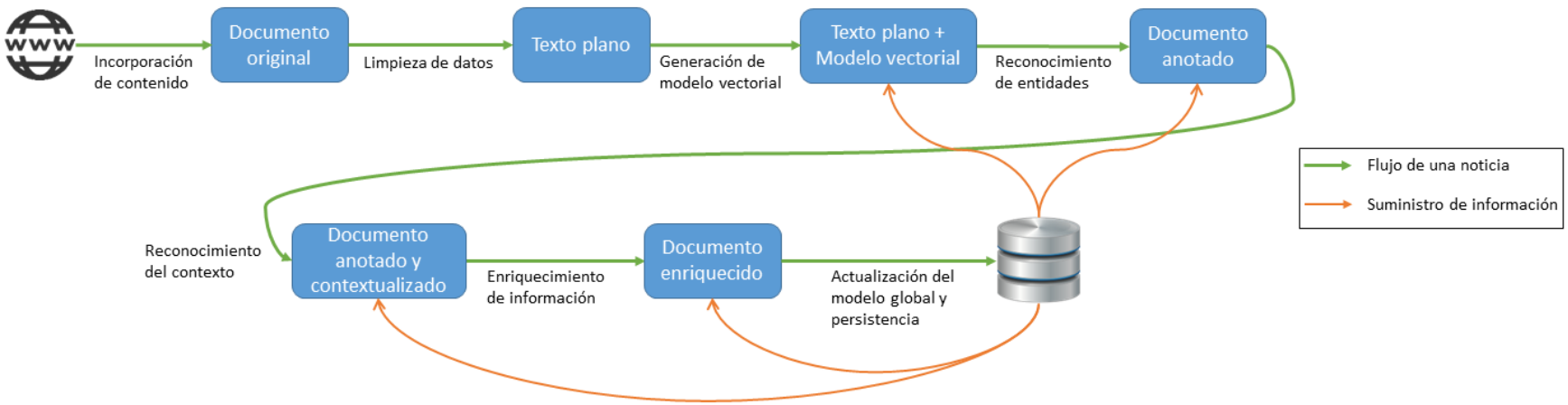


Figura 9: Flujo de procesamiento de una noticia.

Incorporación de contenido

Limpieza de datos

www.eltiempo.com/politica/justicia/caso-de-santiago-uribe-se-convierte-en-problema-poli

TIEMPO Captura de Santiago Uribe se convirtió en un problema político mayor

Captura de Santiago Uribe se convirtió en un problema político mayor

Hermano de expresidente Álvaro Uribe es acusado de homicidio y conformación de grupos paramilitares.

Por: POLITICA |
8:32 a.m. | 2 de marzo de 2016

Foto: Archivo particular
Dirigentes uribistas protestaron ante la Casa de Nariño y el Cap

259 COMPARTIDOS

PUBLICIDAD

La captura de Santiago Uribe Vélez (hermano del expresidente Álvaro Uribe Vélez), acusado de homicidio y conformación de grupos paramilitares, se convirtió en un problema político de marca mayor.

¡A las calles contra el Gobierno!, fue la instrucción de los líderes uribistas a sus bases para protestar por este hecho.
Y la orden no se hizo esperar. El mismo martes, un grupo de líderes del Centro Democrático (CD), con su excandidato presidencial Óscar Iván Zuluaga a la cabeza y los miembros de las bancadas en el Senado y la Cámara, se movilizó hasta la Casa de Nariño agitando banderines en los que se podía leer: "Santos, renuncie ya".

El punto crítico es que los seguidores del expresidente Uribe, jefe del principal sector de oposición, acusan al Gobierno del presidente Santos de influir en la decisión de la Fiscalía para ordenar la privación de la libertad

MÁS LEÍDO

Limpieza de datos

www.eltiempo.com/politica/justicia/caso-de-santiago-uribe-se-convierte-en-problema-poli

TIEMPO Captura de Santiago Uribe se convirtió en un problema político mayor

Captura de Santiago Uribe se convirtió en un problema político mayor

Hermano de expresidente Álvaro Uribe es acusado de homicidio y conformación de grupos paramilitares.

Por: POLITICA |
8:32 a.m. | 2 de marzo de 2016

Foto: Archivo particular
Dirigentes uribistas protestaron ante la Casa de Nariño y el Cap

259 COMPARTIDOS

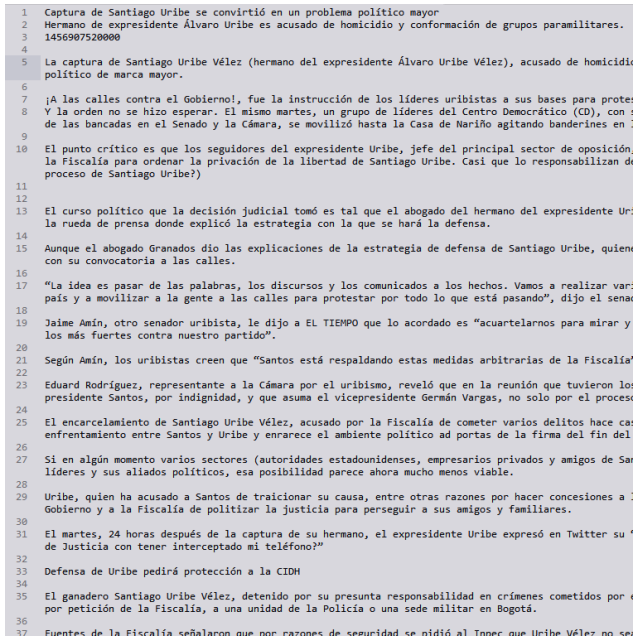
PUBLICIDAD

La captura de Santiago Uribe Vélez (hermano del expresidente Álvaro Uribe Vélez), acusado de homicidio y conformación de grupos paramilitares, se convirtió en un problema político de marca mayor.

¡A las calles contra el Gobierno!, fue la instrucción de los líderes uribistas a sus bases para protestar por este hecho.
Y la orden no se hizo esperar. El mismo martes, un grupo de líderes del Centro Democrático (CD), con su excandidato presidencial Óscar Iván Zuluaga a la cabeza y los miembros de las bancadas en el Senado y la Cámara, se movilizó hasta la Casa de Nariño agitando banderines en los que se podía leer: "Santos, renuncie ya".

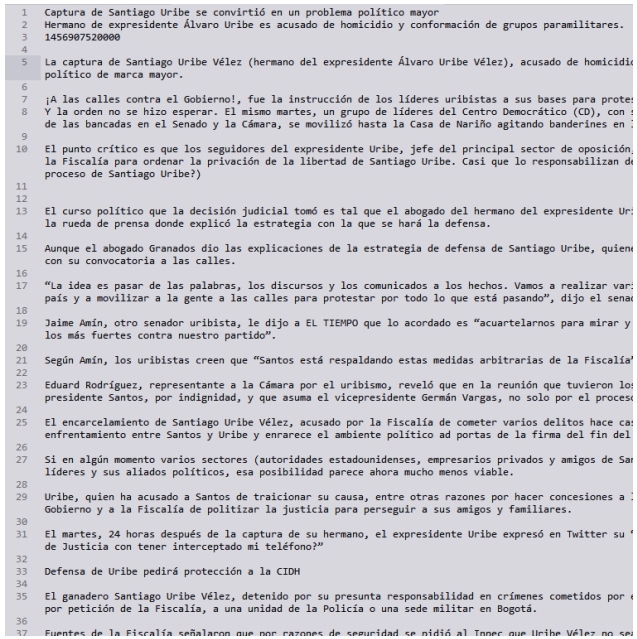
El punto crítico es que los seguidores del expresidente Uribe, jefe del principal sector de oposición, acusan al Gobierno del presidente Santos de influir en la decisión de la Fiscalía para ordenar la privación de la libertad

MÁS LEÍDO



```
Captura de Santiago Uribe se convirtió en un problema político mayor
Hermano de expresidente Álvaro Uribe es acusado de homicidio y conformación de grupos paramilitares.
1456907520000

La captura de Santiago Uribe Vélez (hermano del expresidente Álvaro Uribe Vélez), acusado de homicidio
político de marca mayor.

¡A las calles contra el Gobierno!, fue la instrucción de los líderes uribistas a sus bases para protes
Y la orden no se hizo esperar. El mismo martes, un grupo de líderes del Centro Democrático (CD), con
de las bancadas en el Senado y la Cámara, se movilizó hasta la Casa de Nariño agitando banderines en

El punto crítico es que los seguidores del expresidente Uribe, jefe del principal sector de oposición,
la Fiscalía para ordenar la privación de la libertad de Santiago Uribe. Casi que lo responsabilizan d
proceso de Santiago Uribe?)


El curso político que la decisión judicial tomó es tal que el abogado del hermano del expresidente Ur
la rueda de prensa donde explicó la estrategia con la que se hará la defensa.

Aunque el abogado Granados dio las explicaciones de la estrategia de defensa de Santiago Uribe, quien
con su convocatoria a las calles.

"La idea es pasar de las palabras, los discursos y los comunicados a los hechos. Vamos a realizar var
país y a movilizar a la gente a las calles para protestar por todo lo que está pasando", dijo el senad

Jaime Amín, otro senador uribista, le dijo a EL TIEMPO que lo acordado es "acuartelarnos para mirar y
los más fuertes contra nuestro partido".

Según Amín, los uribistas creen que "Santos está respaldando estas medidas arbitrarias de la Fiscalía"

Eduard Rodríguez, representante a la Cámara por el uribismo, reveló que en la reunión que tuvieron los
presidente Santos, por indignidad, y que asuma el vicepresidente Germán Vargas, no solo por el proceso

El encarcelamiento de Santiago Uribe Vélez, acusado por la Fiscalía de cometer varios delitos hace ca
enfrentamiento entre Santos y Uribe y enarece el ambiente político ad portas de la firma del fin del

Si en algún momento varios sectores (autoridades estadounidenses, empresarios privados y amigos de San
líderes y sus aliados políticos, esa posibilidad parece ahora mucho menos viable.

Uribe, quien ha acusado a Santos de traicionar su causa, entre otras razones por hacer concesiones a
Gobierno y a la Fiscalía de politizar la justicia para perseguir a sus amigos y familiares.

El martes, 24 horas después de la captura de su hermano, el expresidente Uribe expresó en Twitter su
de Justicia con tener interceptado mi teléfono?"

Defensa de Uribe pedirá protección a la CIDH

El ganadero Santiago Uribe Vélez, detenido por su presunta responsabilidad en crímenes cometidos por
por petición de la Fiscalía, a una unidad de la Policía o una sede militar en Bogotá.

Fuentes de la Fiscalía señalaron que por razones de seguridad se pidió al Inpec que Uribe Vélez no se
```

Limpieza de datos



Generación de un modelo vectorial

Generación de un modelo vectorial

- Estandar
  - Bolsa de palabras.
  - Binarización (one-hot enconding).
  - Dimensión de 20k a 13M.

tasa [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0]
casa [0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0]

Generación de un modelo vectorial

- Estandar
  - Bolsa de palabras.
  - Binarización (one-hot enconding).
  - Dimensión de 20k a 13M.

  tasa [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0]
  casa [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0]

- Representación distribuida
  - Modelo probabilísitico neuronal del lenguaje.
  - Utilizar una representación vectorial de acuerdo al contexto.
  - Dimensión de 50 a 1000.

  tasa [0.06 0.03 0.07 0.01 0.02 0.02]
  casa [0.03 0.05 0.07 0.09 0.11 0.13]

Generación de un modelo vectorial

Francia		Colombia		rojo		XBOX	
Palabra Relacionada	Distancia coseno	Palabra relacionada	Distancia coseno	Palabra relacionada	Distancia coseno	Palabra relacionada	Distancia coseno
Italia	0.70979	Ecuador	0.7912	azul	0.8649	PSOne	0.8327
Alemania	0.7017	Bogotá	0.7567	amarillo	0.8450	XboxDIGITO	0.8319
París	0.6790	Venezuela	0.7540	color	0.8258	DSiWare	0.8203
Bretaña	0.6744	colombiano	0.7236	verde	0.7813	PSX	0.8168
España	0.6664	Perú	0.7124	blanco	0.7765	WII	0.7979
Suiza	0.6389	Medellín	0.7080	anaranjado	0.7569	Snes	0.7959
Holanda	0.6378	Panamá	0.7064	negro	0.7564	Videoconsola	0.7957
Kerscamp	0.6326	Bolivia	0.7041	dorado	0.7562	Gamecube	0.7956
francés	0.6265	colombiana	0.7005	gris	0.7371	WiiU	0.7932

Tabla 1: Palabras relacionadas

Generación de un modelo vectorial

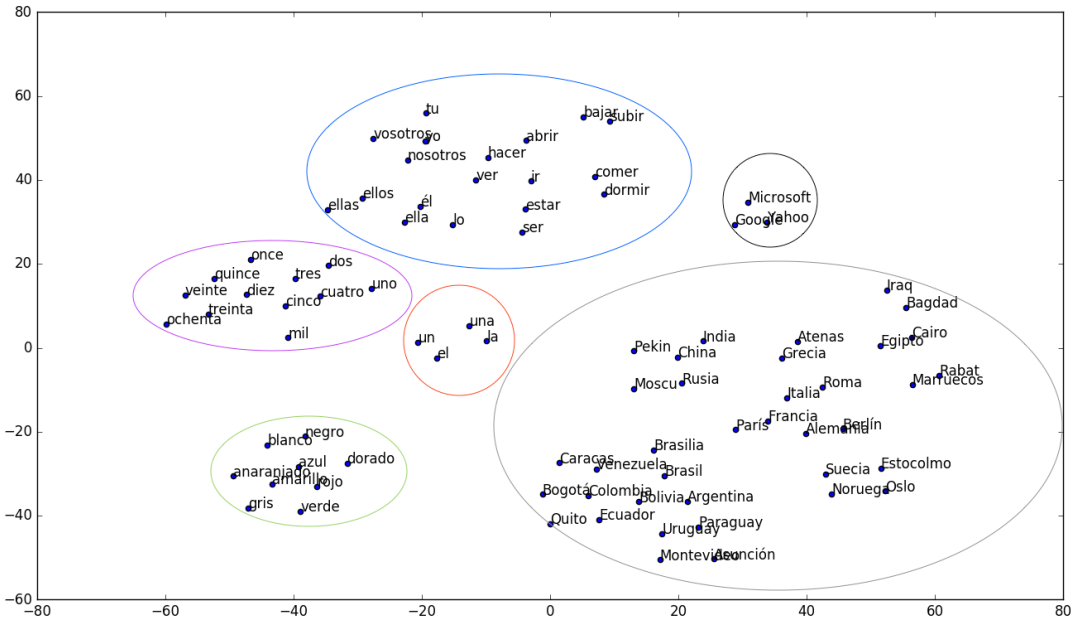


Figura 10: Modelo vectorial - TSNE

Generación de un modelo vectorial

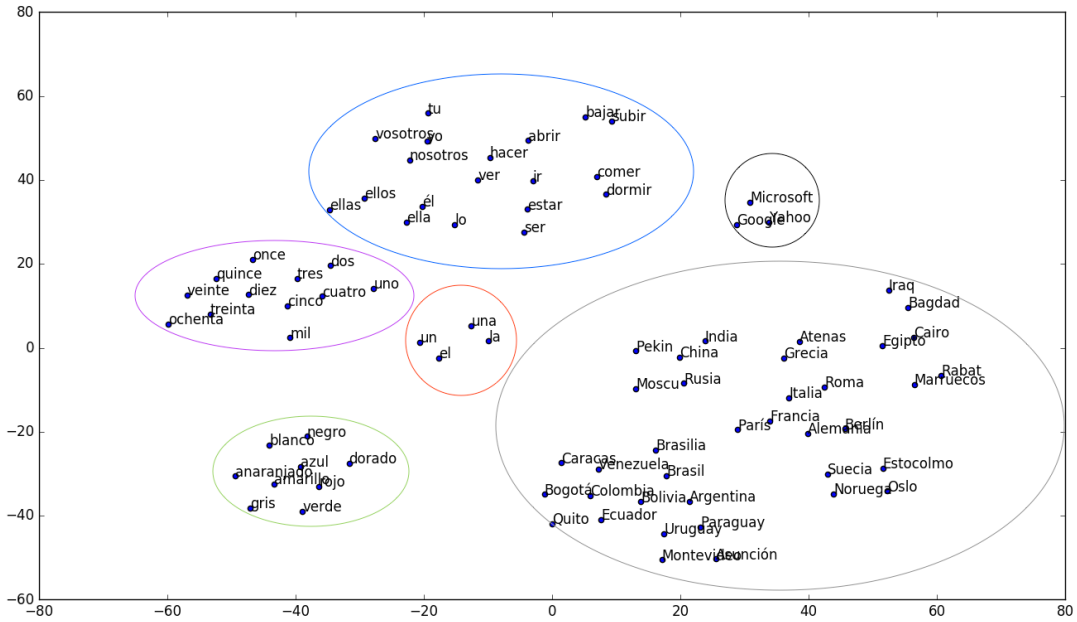


Figura 10: Modelo vectorial - TSNE

Modelo vectorial

La representación distribuida de las palabras permite tener similitud semántica.

Reconocimiento de entidades

Reconocimiento de entidades

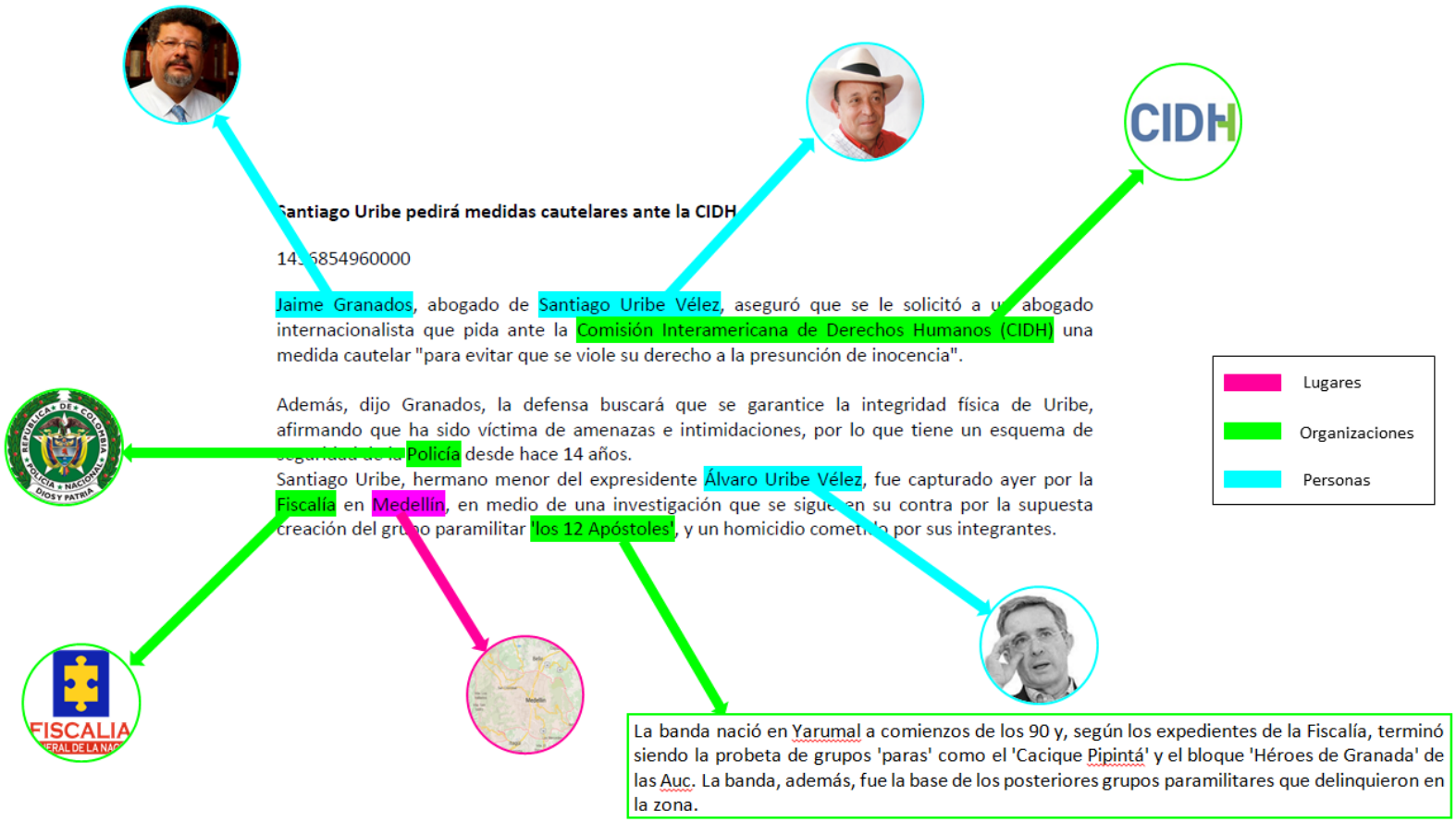


Figura 11: Reconocimiento de entidades

Reconocimiento de contexto

Reconocimiento de contexto

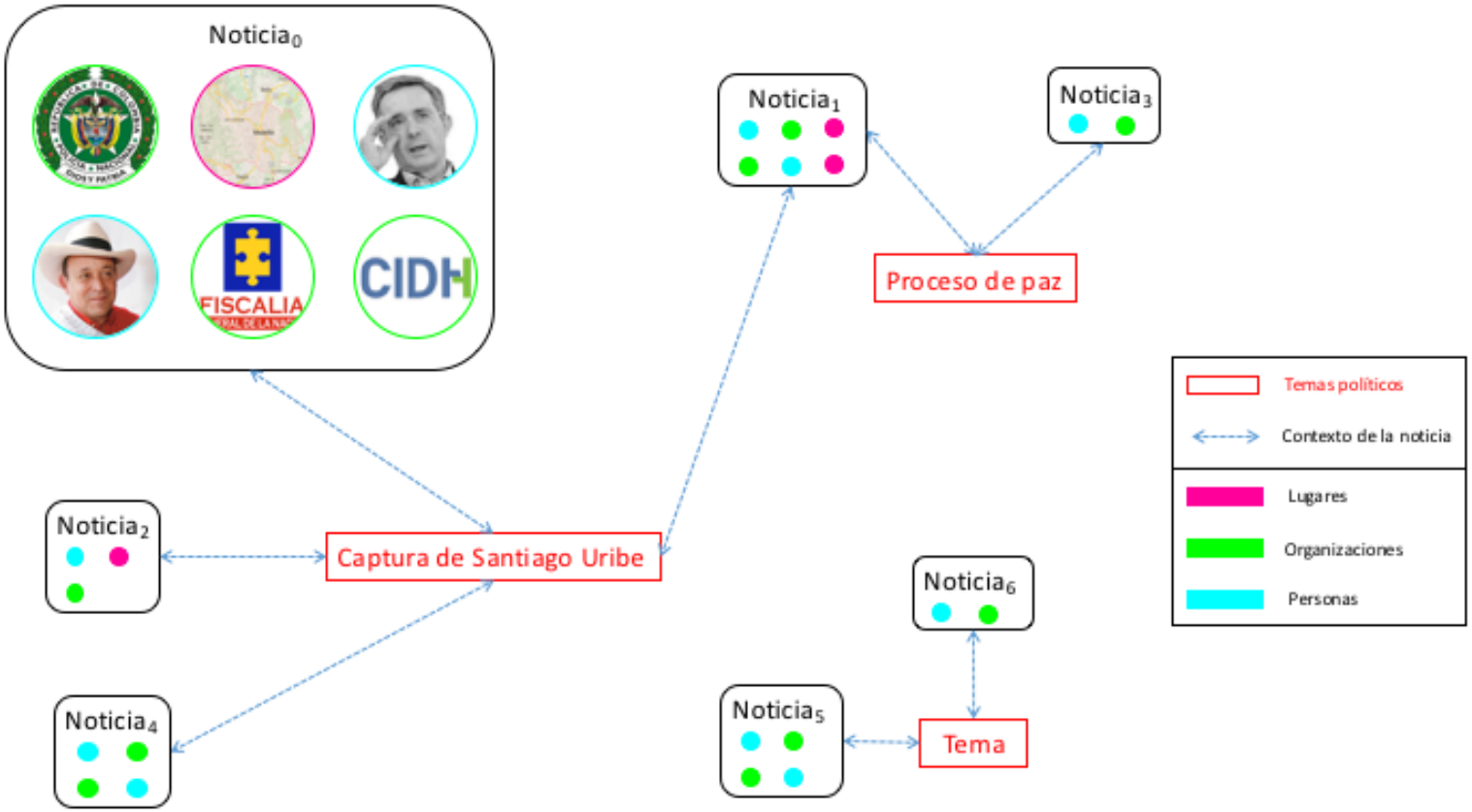


Figura 12: Reconocimiento de contexto

Enriquecimiento de Información

Enriquecimiento de Información

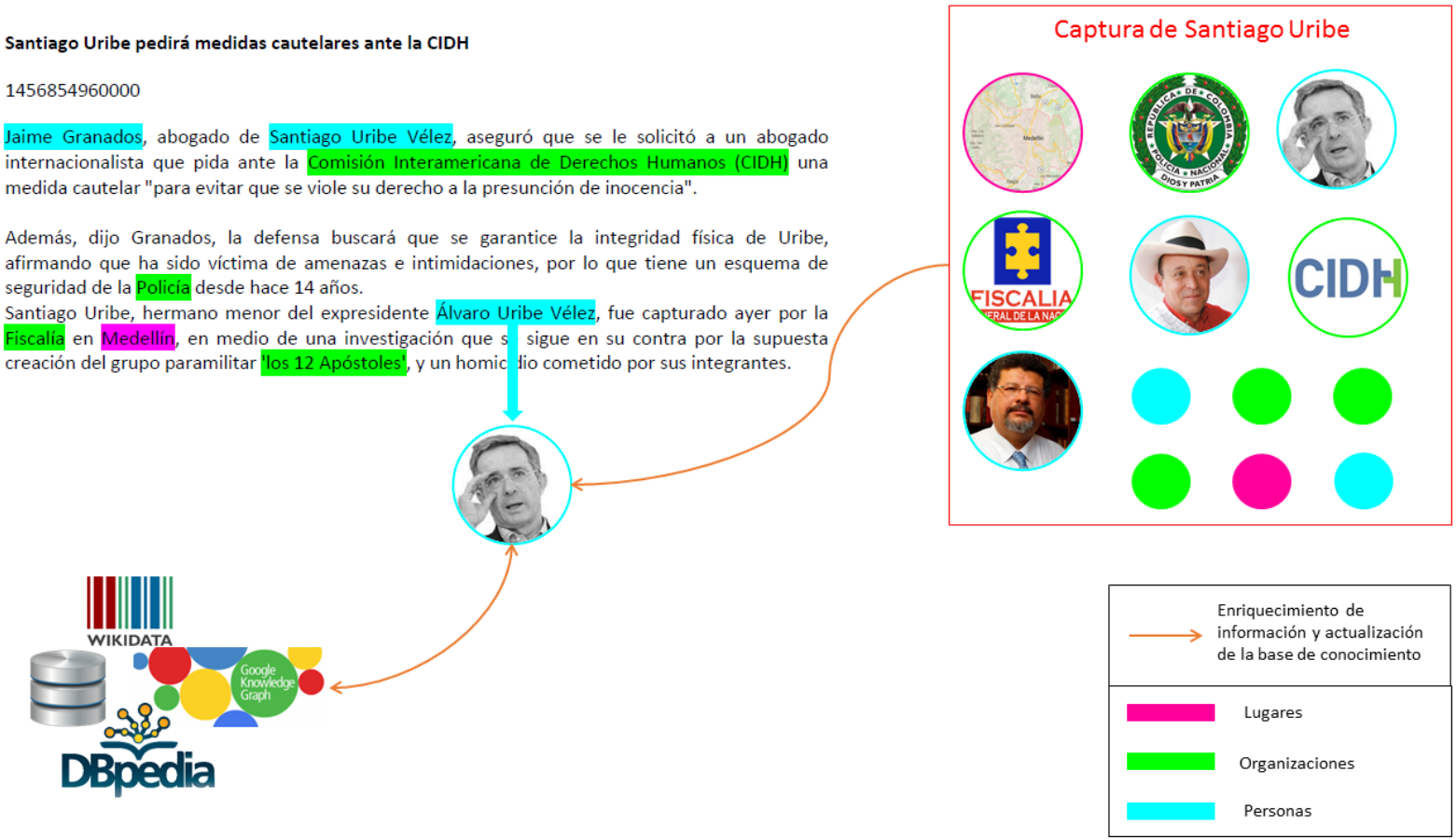


Figura 13: Enriquecimiento de Información

Resultado del proceso

Santiago Uribe pedirá medidas cautelares ante la CIDH

Jaime Granados, abogado de Santiago Uribe Vélez, aseguró que se le solicitó a un abogado internacionalista que pida ante la Comisión Interamericana de Derechos Humanos (CIDH) una medida cautelar "para evitar que se viole su derecho a la presunción de inocencia".

Además, dijo Granados, la defensa buscará que se garantice la integridad física de Uribe, afirmando que ha sido víctima de amenazas e intimidaciones, por lo que tiene un esquema de seguridad de la Policía desde hace 14 años.
Santiago Uribe, hermano menor del expresidente Álvaro Uribe Vélez, fue capturado ayer por la Fiscalía en Medellín, en medio de una investigación que s sigue en su contra por la supuesta creación del grupo paramilitar 'los 12 Apóstoles', y un homic dio cometido por sus integrantes.

Álvaro Uribe Vélez (Medellín; 4 de julio de 1952) es un político y abogado colombiano, Presidente de la República de Colombia en 2002, y reelegido en el año 2006. Uribe se graduó en derecho en la Universidad de Antioquia...

Figura 14: Resultado del proceso

Agenda

Diseño de la solución

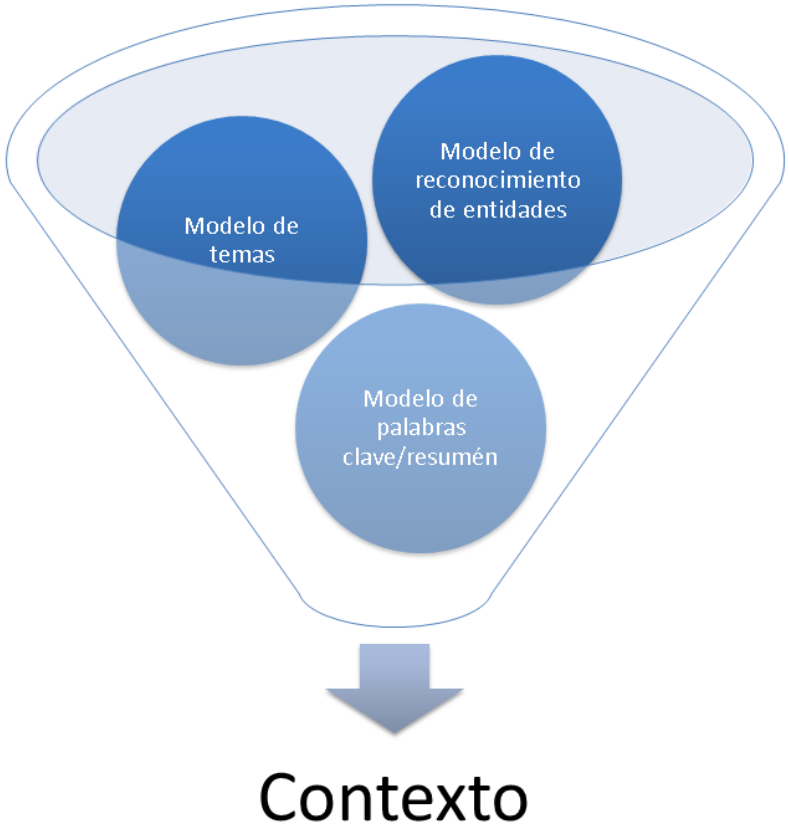


Figura 15: Modelos para reconocimiento de contexto

Diseño de la solución

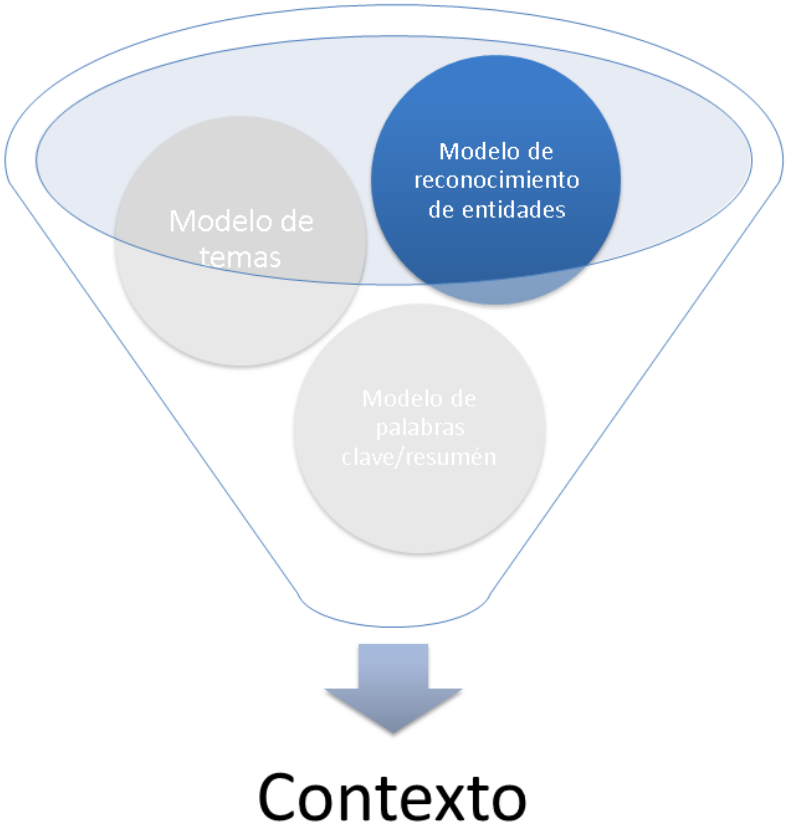
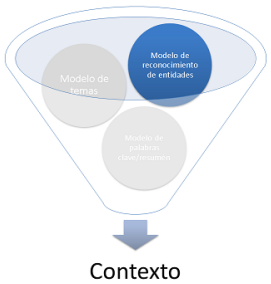


Figura 15: Modelos para reconocimiento de contexto

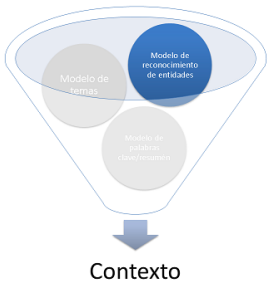
Reconocimiento de entidades

- Retos
- Soluciones



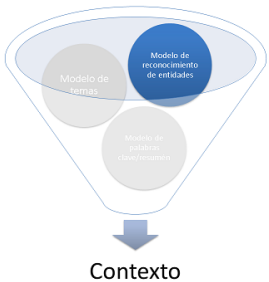
Reconocimiento de entidades

- Retos
  - Representación a trozos del texto.
- Soluciones
  - Tokenización + contexto.



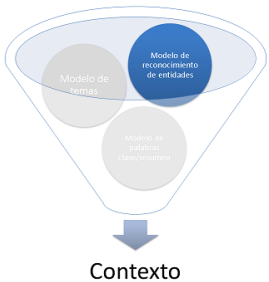
Reconocimiento de entidades

- Retos
  - Representación a trozos del texto.
  - Algoritmo de inferencia.
- Soluciones
  - Tokenización + contexto.
  - Clasificadores Bayesianos, HMM, CRF, DT, SVM, NN, CNN.



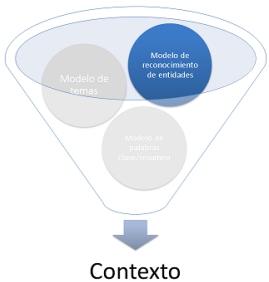
Reconocimiento de entidades

- Retos
  - Representación a trozos del texto.
  - Algoritmo de inferencia.
  - Modelado de dependencias no locales.
- Soluciones
  - Tokenización + contexto.
  - Clasificadores Bayesianos, HMM, CRF, DT, SVM, NN, CNN.
  - Contexto.



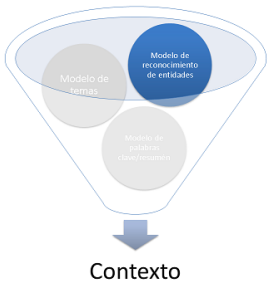
Reconocimiento de entidades

- Retos
  - Representación a trozos del texto.
  - Algoritmo de inferencia.
  - Modelado de dependencias no locales.
  - Recursos de conocimiento externo.
- Soluciones
  - Tokenización + contexto.
  - Clasificadores Bayesianos, HMM, CRF, DT, SVM, NN, CNN.
  - Contexto.
  - Ontologías.



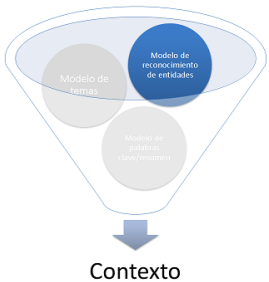
Reconocimiento de entidades

- Retos
  - Representación a trozos del texto.
  - Algoritmo de inferencia.
  - Modelado de dependencias no locales.
  - Recursos de conocimiento externo.
  - Generación de características.
- Soluciones
  - Tokenización + contexto.
  - Clasificadores Bayesianos, HMM, CRF, DT, SVM, NN, CNN.
  - Contexto.
  - Ontologías.
  - Machine learning vs Deep learning.



Reconocimiento de entidades

- Retos
  - Representación a trozos del texto.
  - Algoritmo de inferencia.
  - Modelado de dependencias no locales.
  - Recursos de conocimiento externo.
  - Generación de características.
  - Medida de desempeño.
- Soluciones
  - Tokenización + contexto.
  - Clasificadores Bayesianos, HMM, CRF, DT, SVM, NN, CNN.
  - Contexto.
  - Ontologías.
  - Machine learning vs Deep learning.
  - Precision, Recall, F1.



Machine learning vs Deep learning

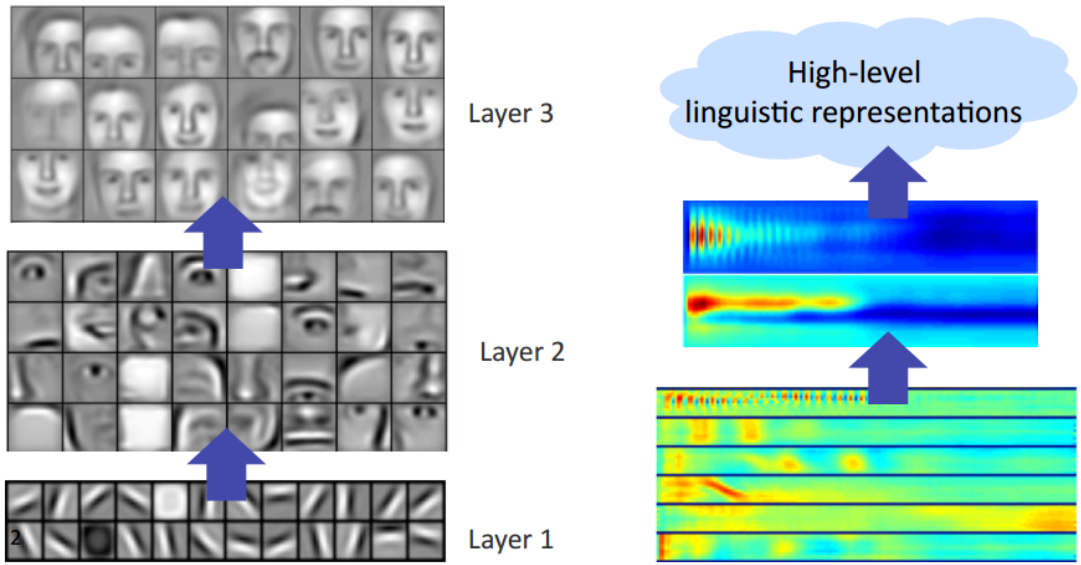
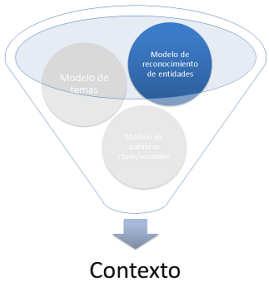


Figura 16: Aprendizaje de características [51]



Machine learning vs Deep learning

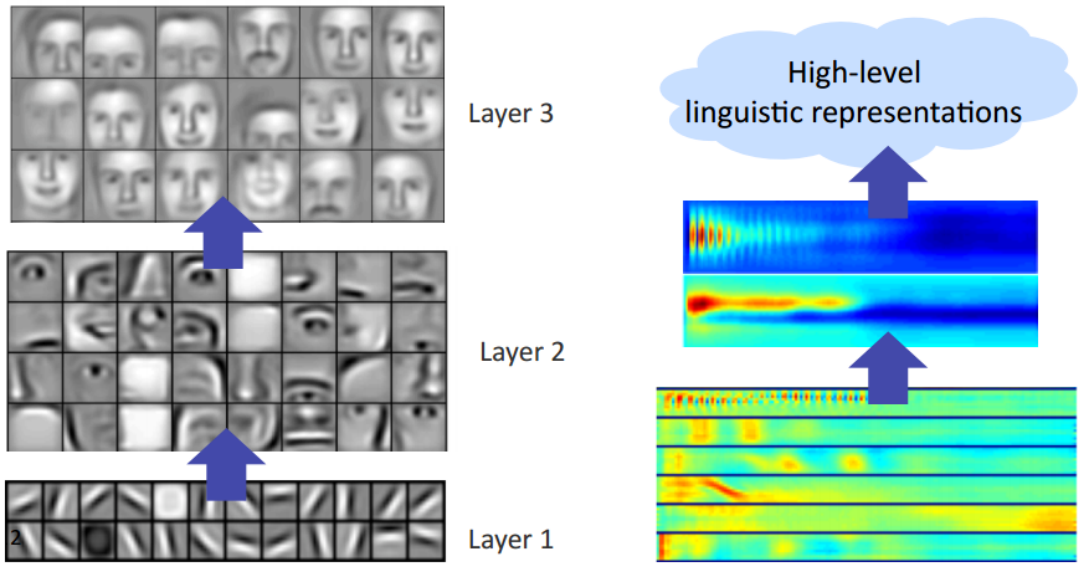
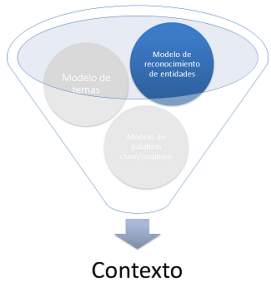


Figura 16: Aprendizaje de características [51]

Problema

Requiere gran cantidad de datos anotados.



Redes Neuronales

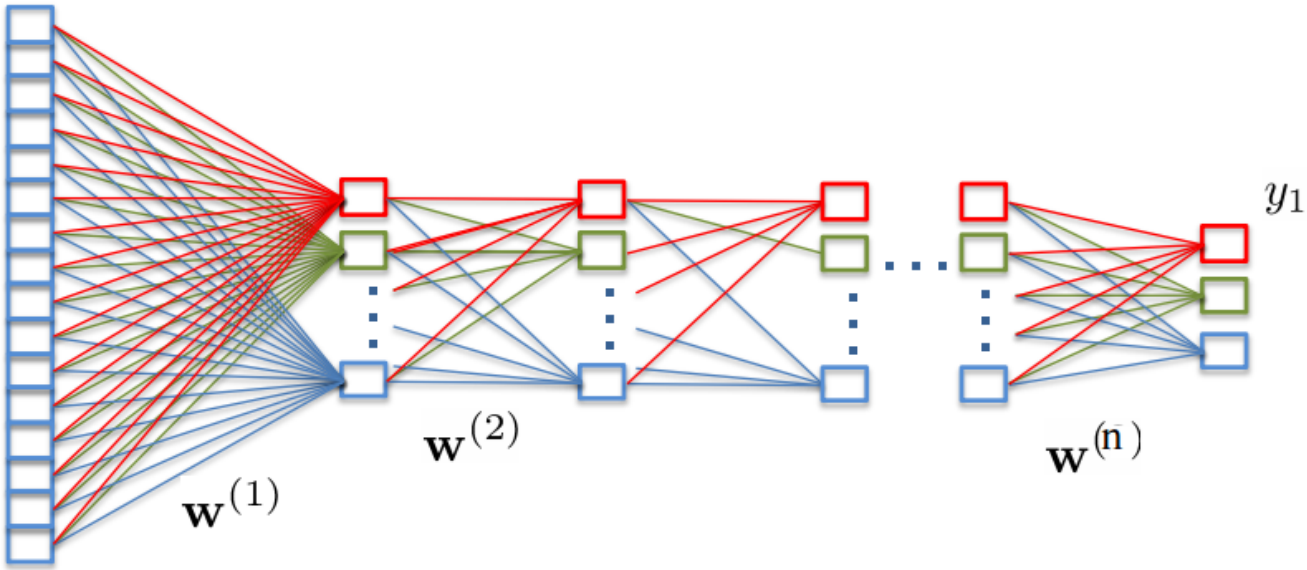


Figura 17: Red neuronal profunda

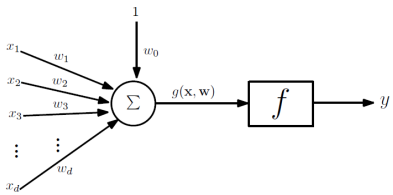
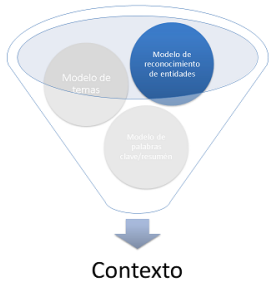


Figura 18: Modelo de una neurona



Propuesta - Red Neuronal

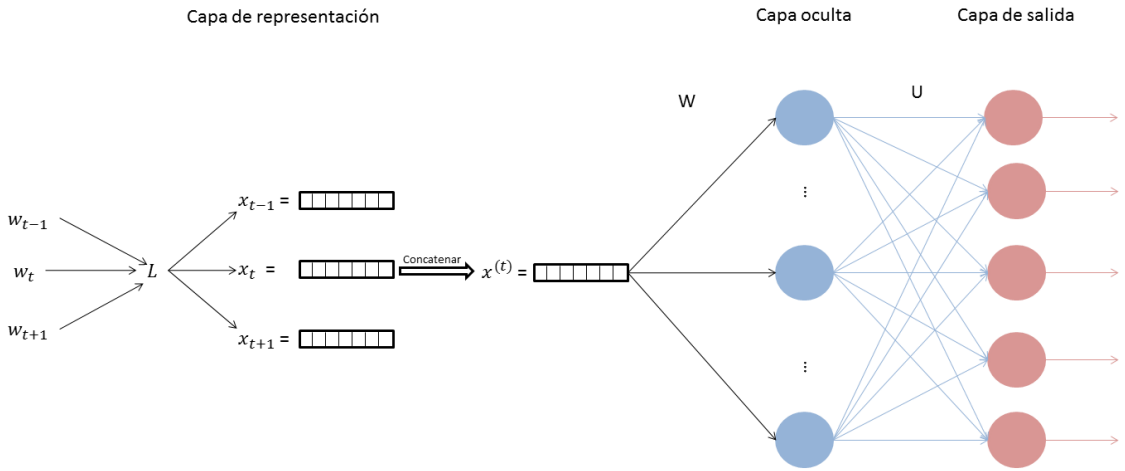


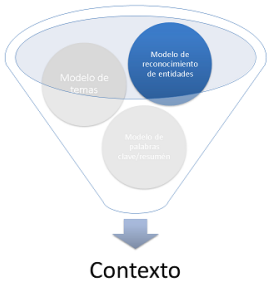
Figura 19: Modelo red neuronal profunda

$$x^{(t)} = [Lx_{(t-1)}, Lx_{(t)}, Lx_{(t+1)}], L \in \mathbb{R}^{|\mathrm{v}| \times \mathrm{d}}$$

$$h = tanh(Wx^{(t)} + b_1), W \in \mathbb{R}^{3\mathrm{d} \times \mathrm{c}}, \mathrm{b}_1 \in \mathbb{R}^{\mathrm{c}}$$

$$\hat{y} = softmax(Uh + b_2), U \in \mathbb{R}^{\mathrm{c} \times 5}, \mathrm{b}_2 \in \mathbb{R}^{5}$$

$$CE(y, \hat{y}) = -\sum_{c=1}^{C} y_c log(\hat{y}_c)$$



Redes Neuronales convolucionales

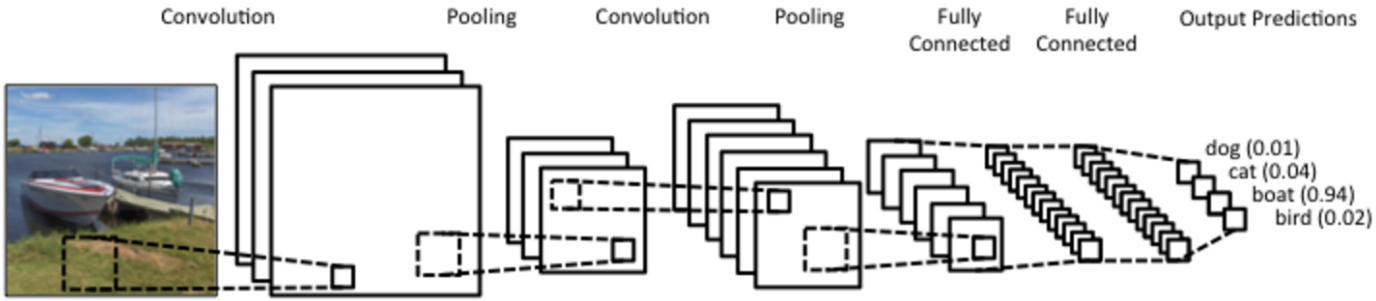


Figura 20: Red neuronal convolucional

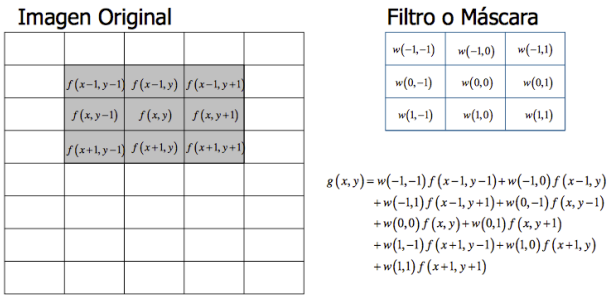


Figura 21: Producto de convolución

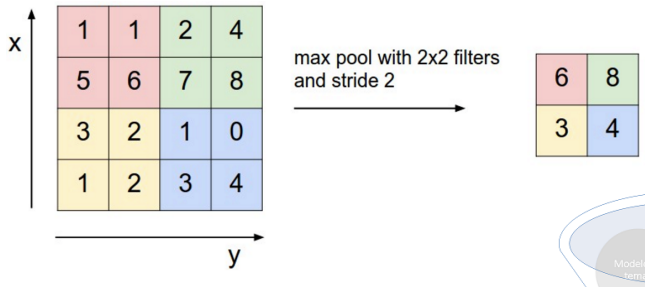
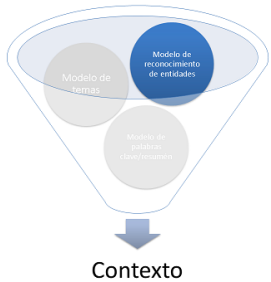


Figura 22: Max pooling



Propuesta - Red Neuronal Convolucional

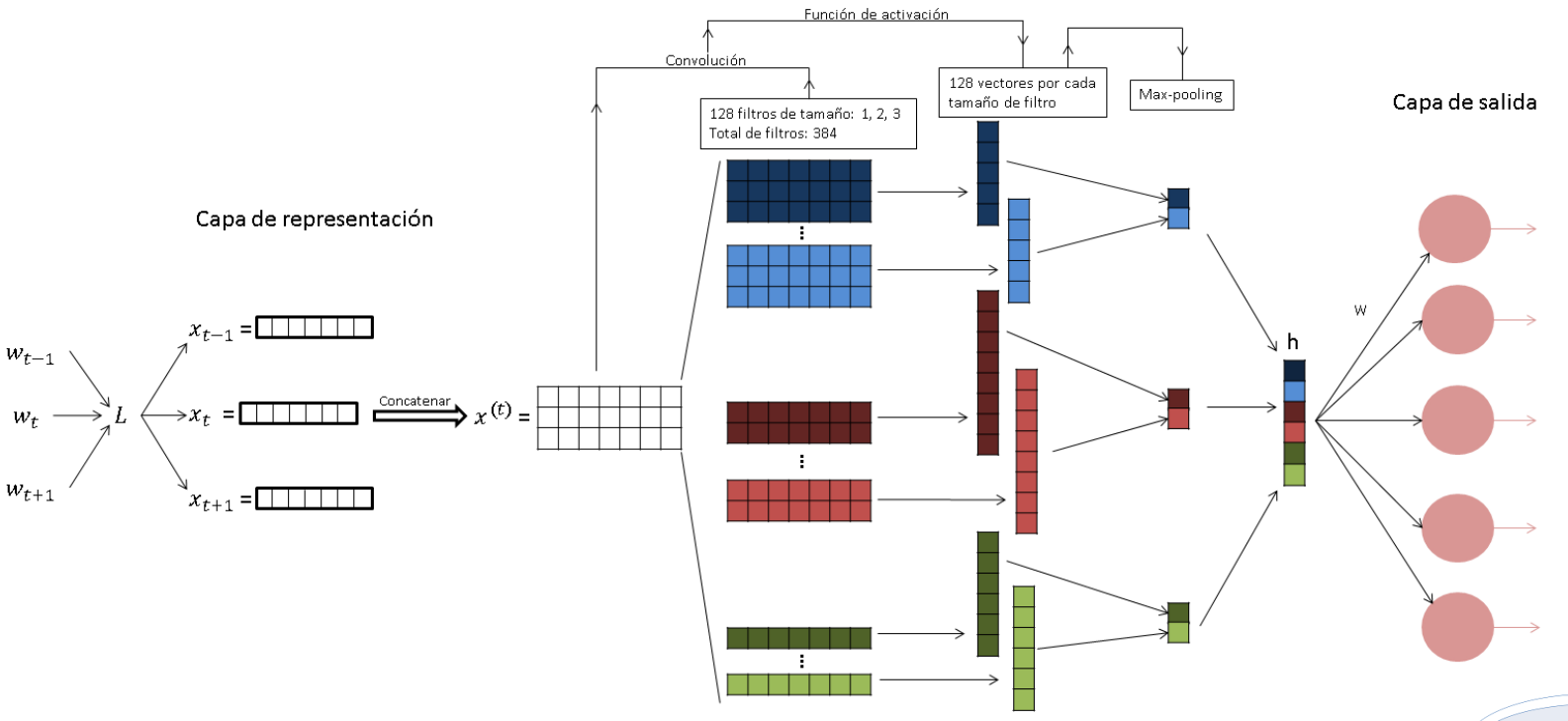
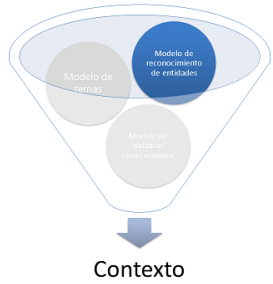
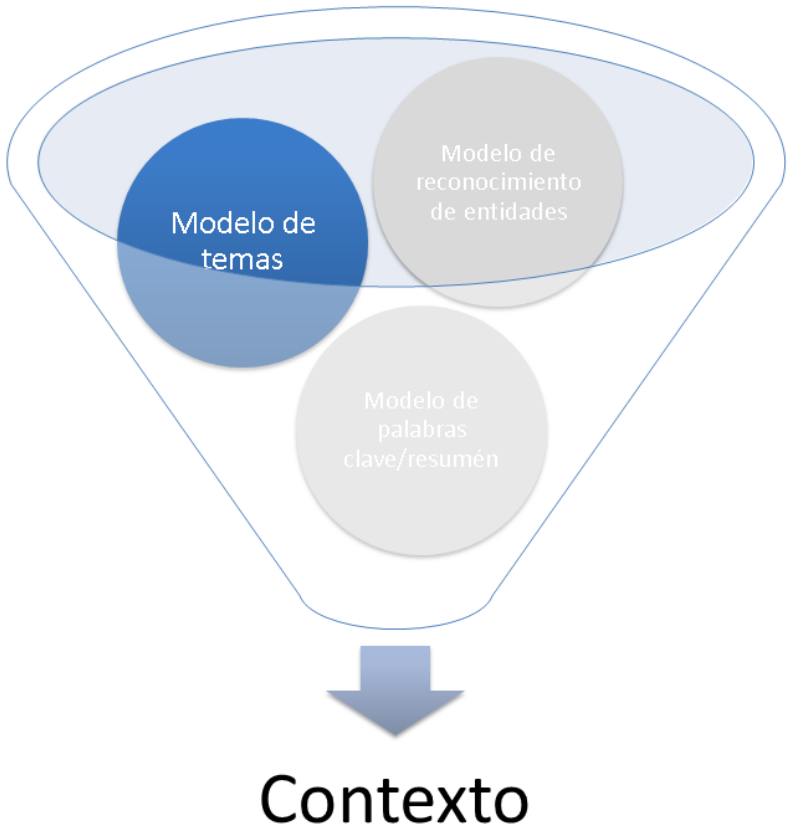


Figura 23: Modelo red neuronal convolucional

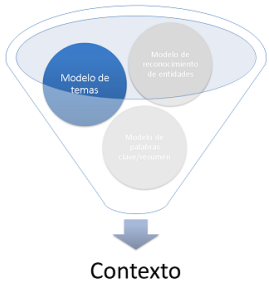


Diseño de la solución



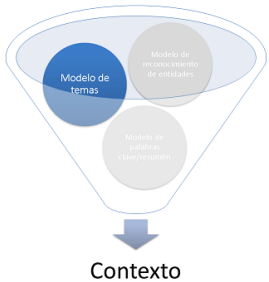
Modelos de temas

- Retos
- Soluciones



Modelos de temas

- Retos
  - Temas recurrentes no continuos en el tiempo.
- Soluciones
  - Modelos por rango de tiempo.



Modelos de temas

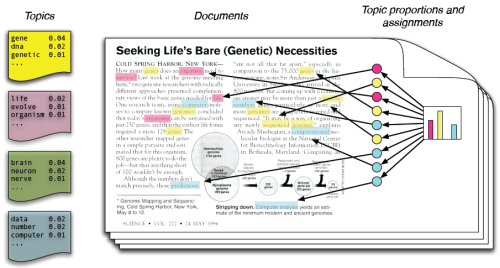
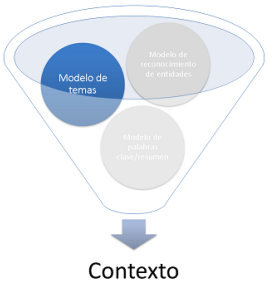


Figura 24: Representación LDA

- Modelos
  - Latent Dirichlet Allocation (LDA).



Modelos de temas

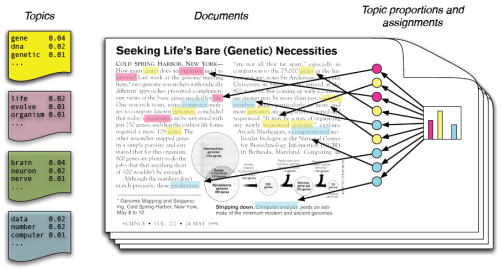


Figura 24: Representación LDA

- Modelos
  - Latent Dirichlet Allocation (LDA).
  - Clustering.

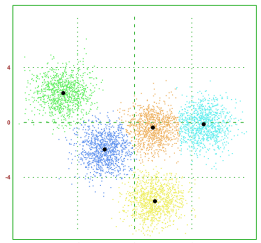
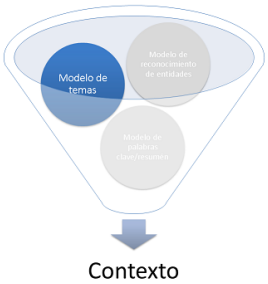


Figura 25: Representación clustering



Modelos de temas

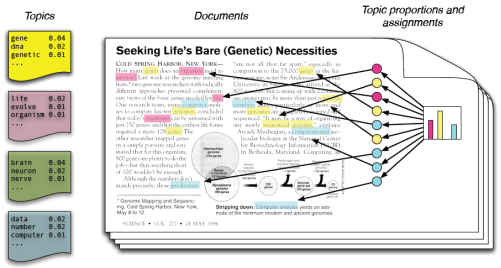


Figura 24: Representación LDA

- Modelos
  - Latent Dirichlet Allocation (LDA).
  - Clustering.
  - LSH/MinHash.

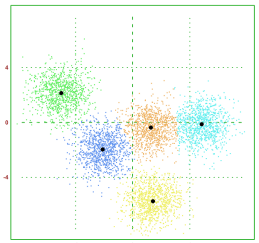
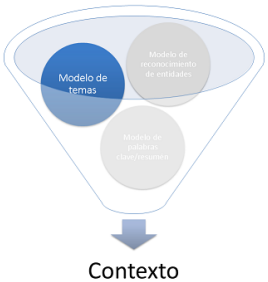


Figura 25: Representación clustering



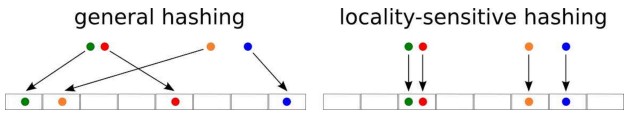
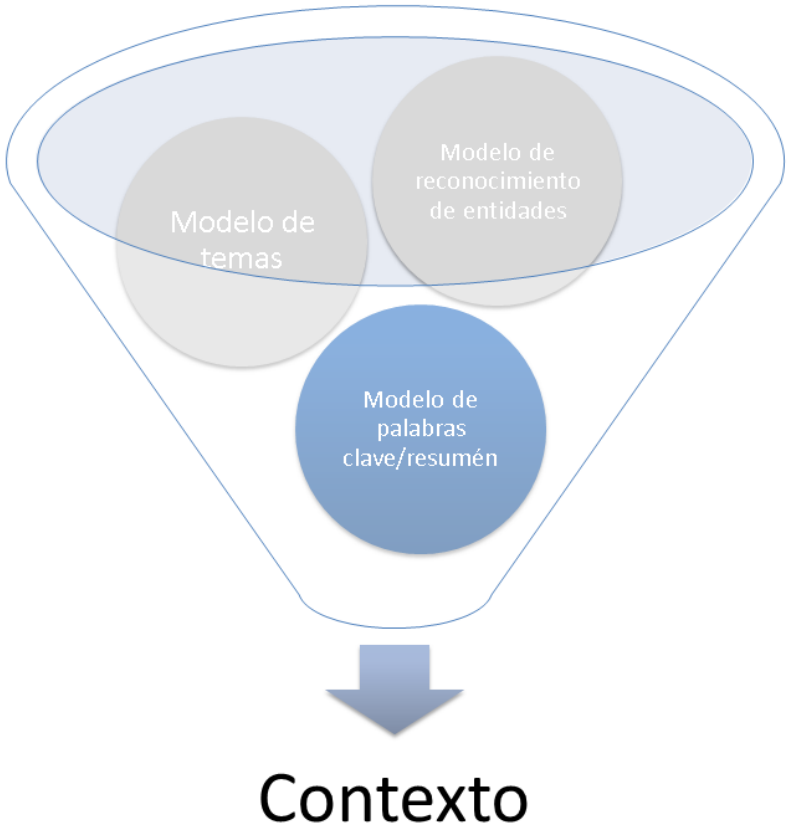


Figura 26: Representación LSH

Diseño de la solución



Modelos de palabras clave/resumen

- Modelos
  - Textrank.

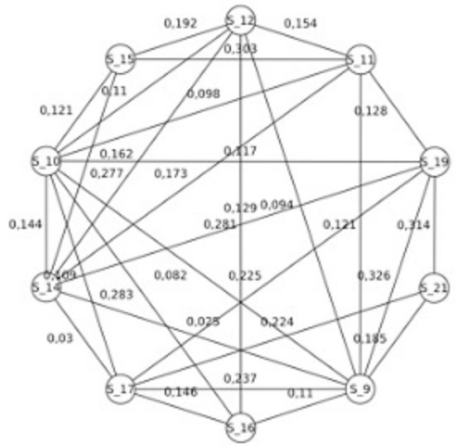
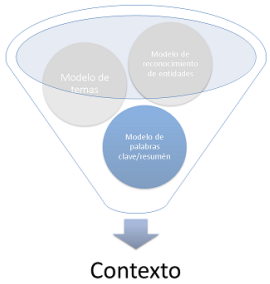


Figura 27: Representación Textrank



Modelos de palabras clave/resumen

- Modelos
  - Textrank.
  - Rake.

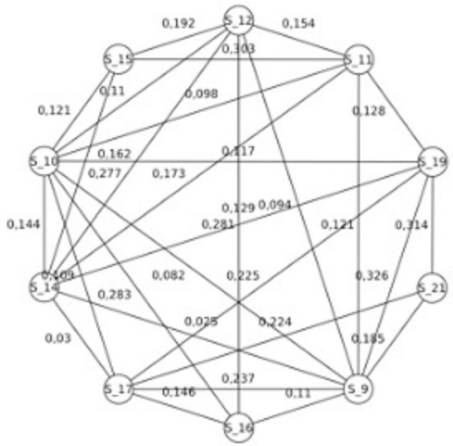
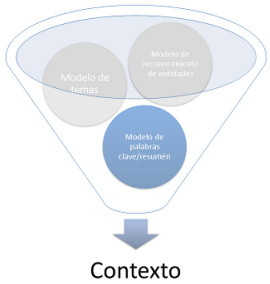


Figura 27: Representación Textrank



Agenda

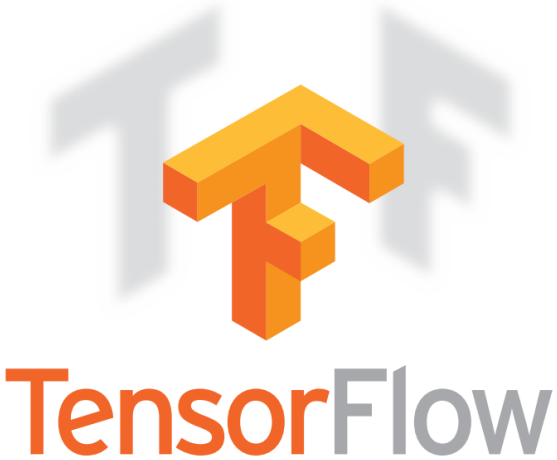
Implementación y validación

- Modelos de reconocimiento de entidades.
- Modelos de temas.
- Modelos de palabras clave/resumen.
- Aplicación de usuario final.

Modelos implementados

- Red Neuronal.
- Red neuronal convolucional.

Herramientas



NLTK

Red Neuronal

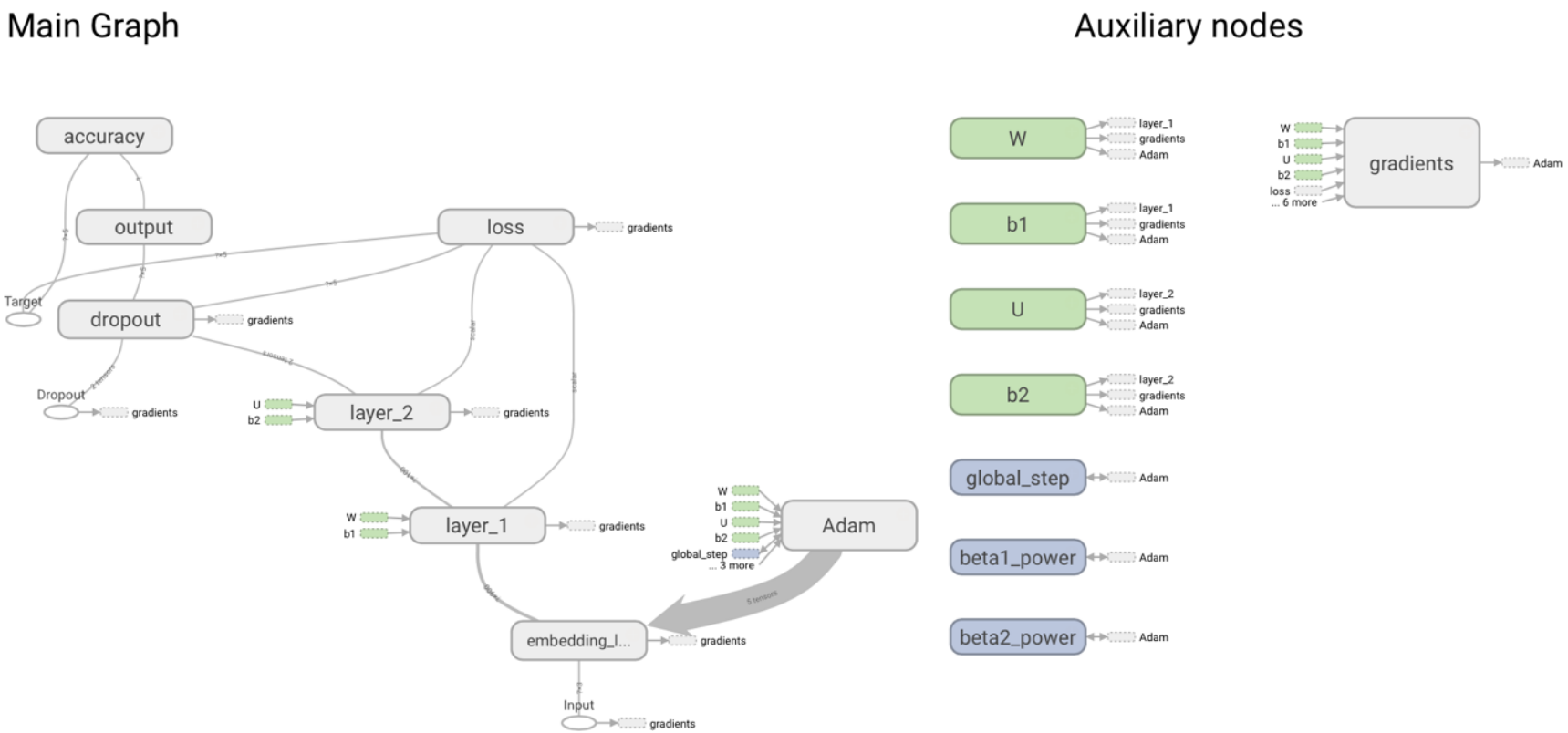


Figura 28: Implementación red neuronal

Red Neuronal

Hiperparmámetro	**Nombre del hiperparámetro**	**Valor**
dropout_keep_prob	Probabilidad de dropout	0.9
windows_size	Tamaño de la ventana	3
embed_size	Tamaño del vector de representación	300
hidden_size	Tamaño de la capa oculta	100
lr	Tasa de aprendizaje	0.001
label_size	Cantidad de clases	5
max_epochs	Épocas de entrenamiento	24

Cuadro 2: Hiperparámetros del modelo

Red Neuronal convolucional

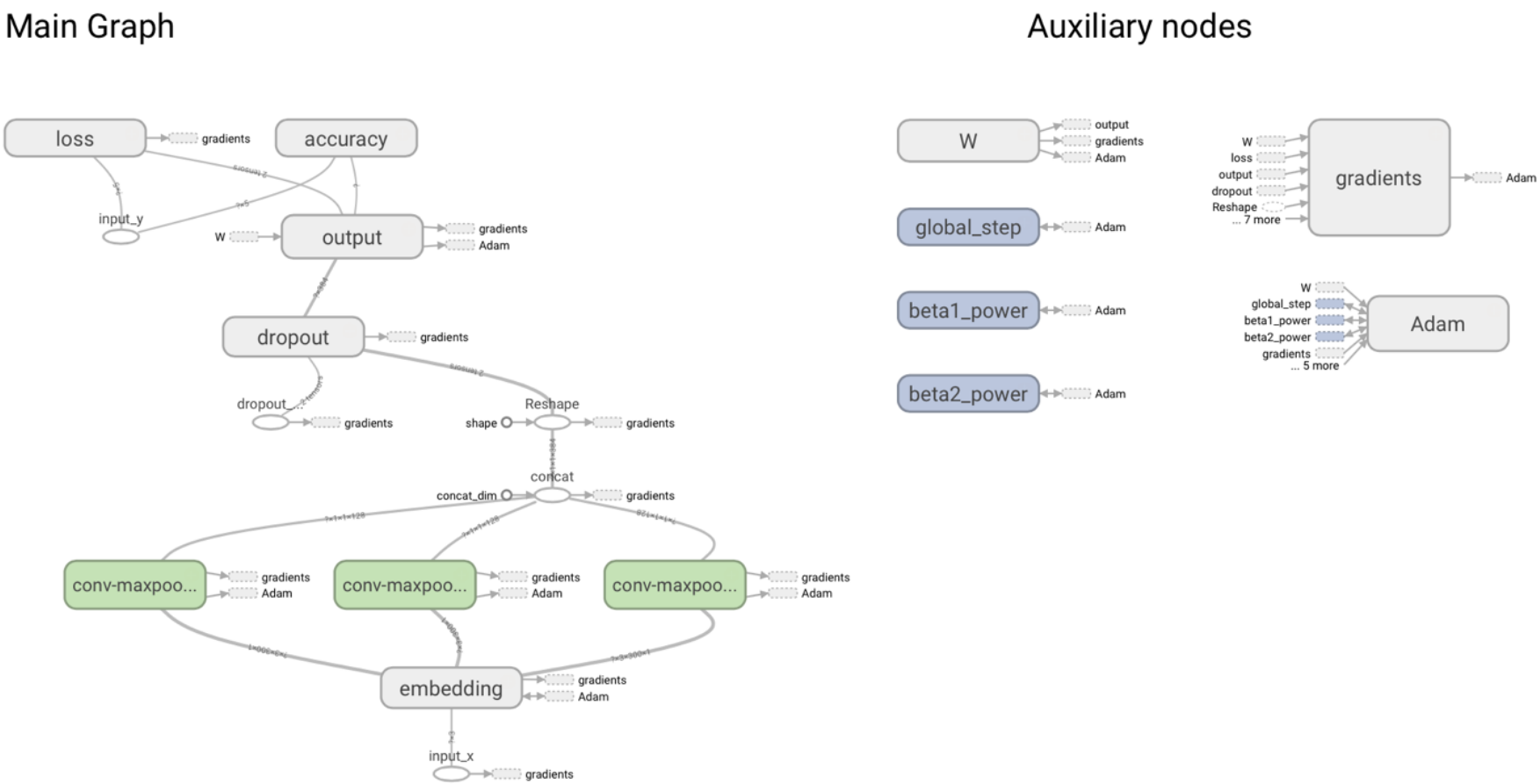


Figura 29: Implementación red neuronal convolucional

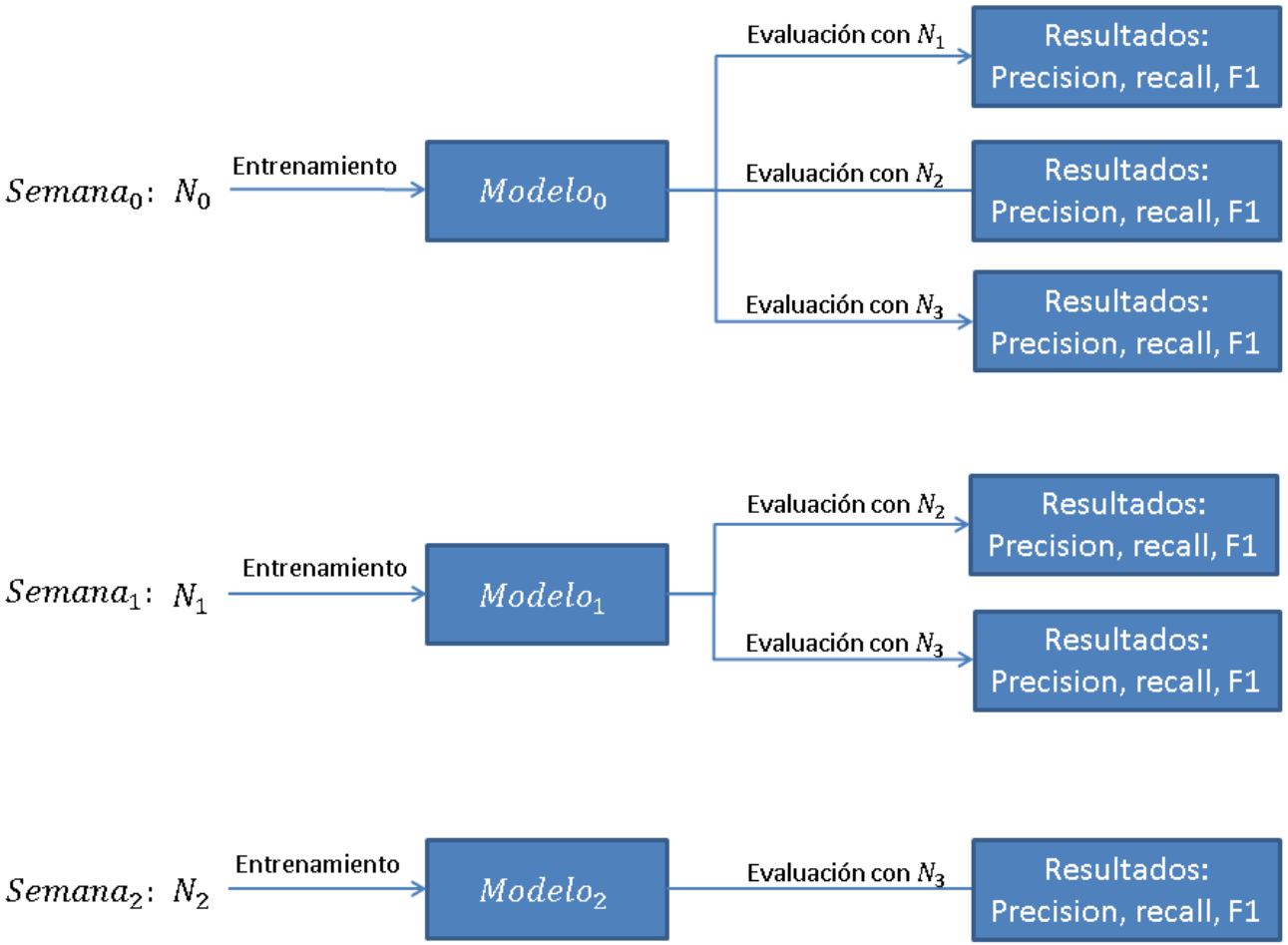
Red Neuronal convolucional

Hiperparmámetro	**Nombre del hiperparámetro**	**Valor**
dropout_keep_prob	Probabilidad de dropout	0.9
windows_size	Tamaño de la ventana	3
embed_size	Tamaño del vector de representación	300
lr	Tasa de aprendizaje	100
label_size	Cantidad de clases	0.001
max_epochs	Épocas de entrenamiento	5
num_filters	Cantidad de filtros	24
filter_size	Tamaño de los filtros	1, 2, 3

Cuadro 3: Hiperparámetros del modelo

Procesos de evaluación

Comparación contra reconocimiento de entidades de Stanford (CRF).



Entrenamiento

Dataset:

- Incorporación de noticias: 10300 noticias.

Entrenamiento

Dataset:

- Incorporación de noticias: 10300 noticias.
- Wikipedia en español.

Entrenamiento

Dataset:

- Incorporación de noticias: 10300 noticias.
- Wikipedia en español.
- Conll2002.

Entrenamiento

Dataset:

- Incorporación de noticias: 10300 noticias.
- Wikipedia en español.
- Conll2002.

Resultados:

- 54’710.000 frases

Entrenamiento

Dataset:

- Incorporación de noticias: 10300 noticias.
- Wikipedia en español.
- Conll2002.

Resultados:

- 54’710.000 frases
- 1 389’381.510 palabras.

Entrenamiento

Dataset:

- Incorporación de noticias: 10300 noticias.
- Wikipedia en español.
- Conll2002.

Resultados:

- 54’710.000 frases
- 1 389’381.510 palabras.
- Vocabulario:1’037.718 palabras.

Entrenamiento

Dataset:

- Incorporación de noticias: 10300 noticias.
- Wikipedia en español.
- Conll2002.

Resultados:

- 54’710.000 frases
- 1 389’381.510 palabras.
- Vocabulario:1’037.718 palabras.
- Tiempo de entrenamiento: 2 días en MacbookPro, 16GB, 4 procesadores 😮😕

Entrenamiento

Dataset:

- Incorporación de noticias: 10300 noticias.
- Wikipedia en español.
- Conll2002.

Resultados:

- 54’710.000 frases
- 1 389’381.510 palabras.
- Vocabulario:1’037.718 palabras.
- Tiempo de entrenamiento: 2 días en MacbookPro, 16GB, 4 procesadores 😮😕
- AWS: 6 horas en c4.4xlarge, 30GB, 16 procesadores 😊

Resultados de entrenamiento - Semana 0

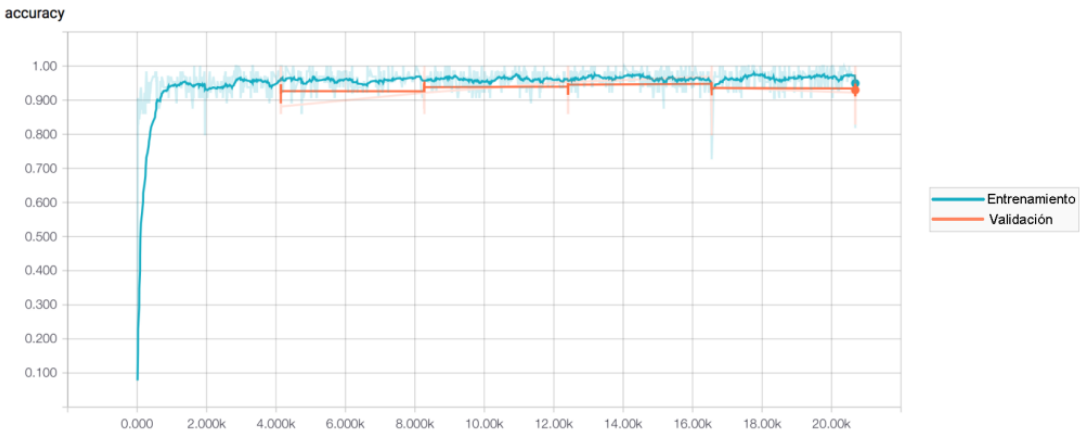


Figura 30: Precisión de entrenamiento

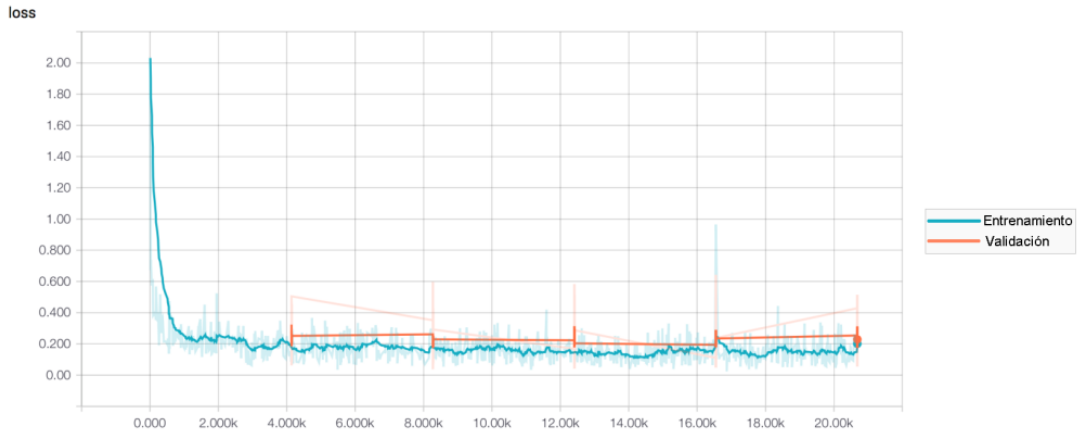


Figura 31: Función de perdida en entrenamiento

Resultados de entrenamiento - Semana 1

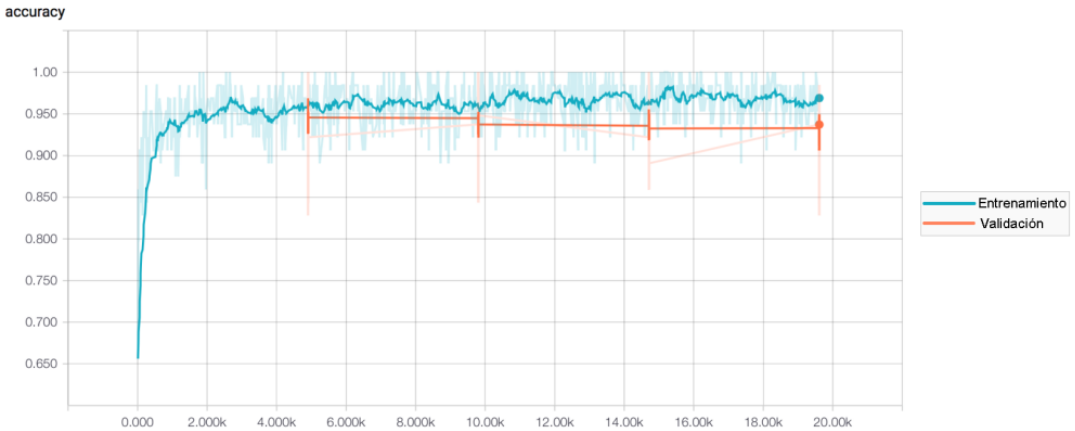


Figura 32: Precisión de entrenamiento

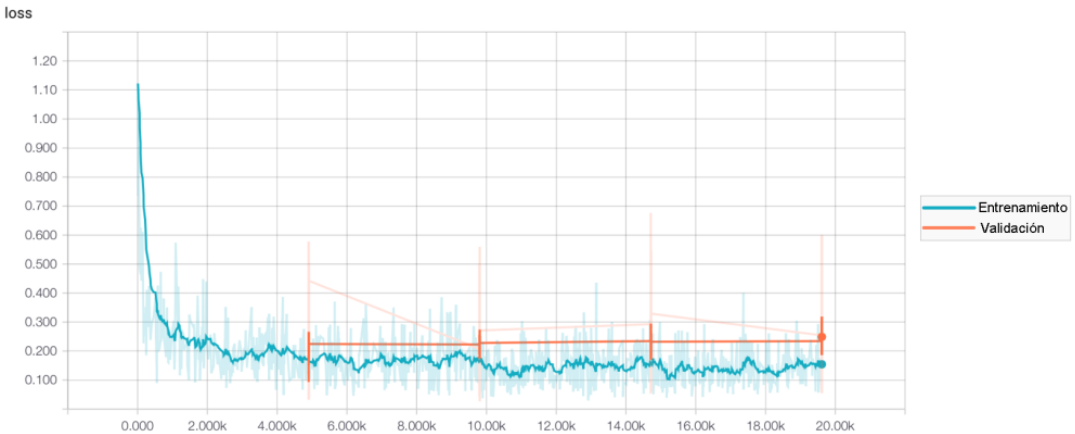


Figura 33: Función de perdida en entrenamiento

Resultados de entrenamiento - Semana 2

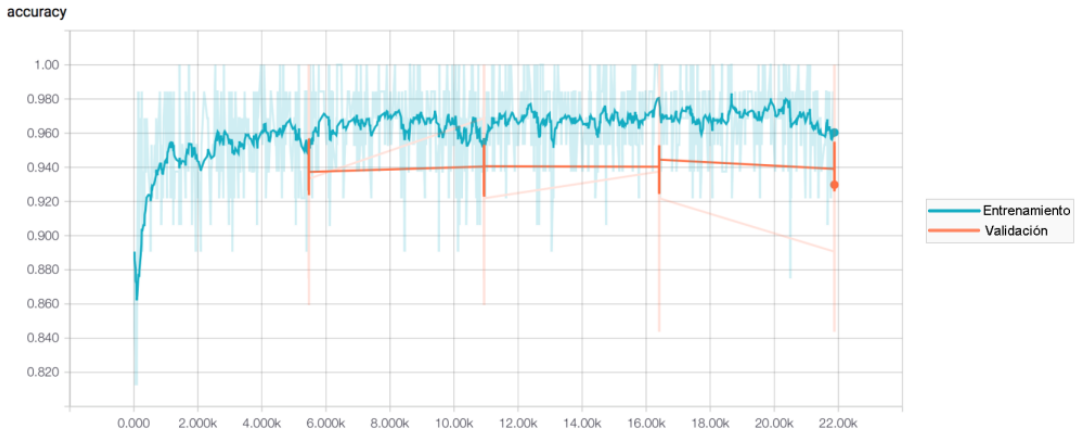


Figura 34: Precisión de entrenamiento

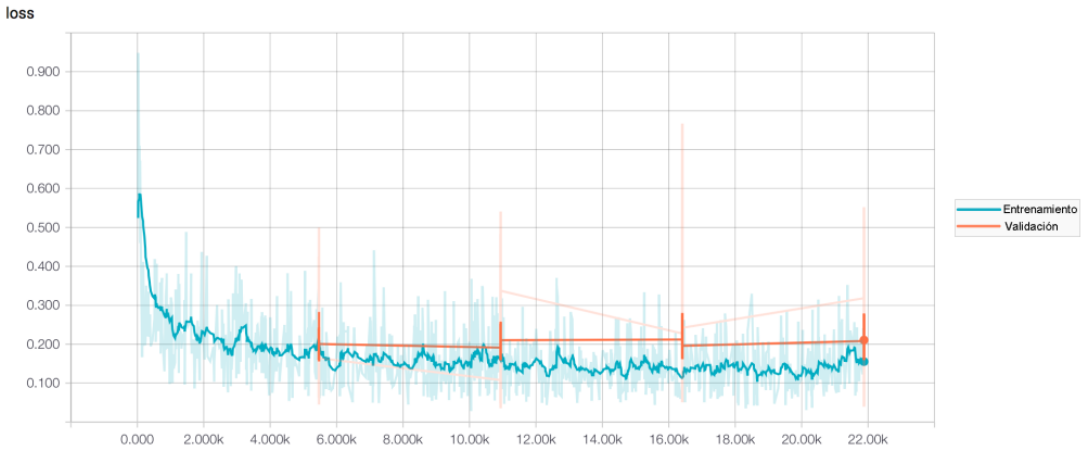


Figura 35: Función de perdida en entrenamiento

Resultados

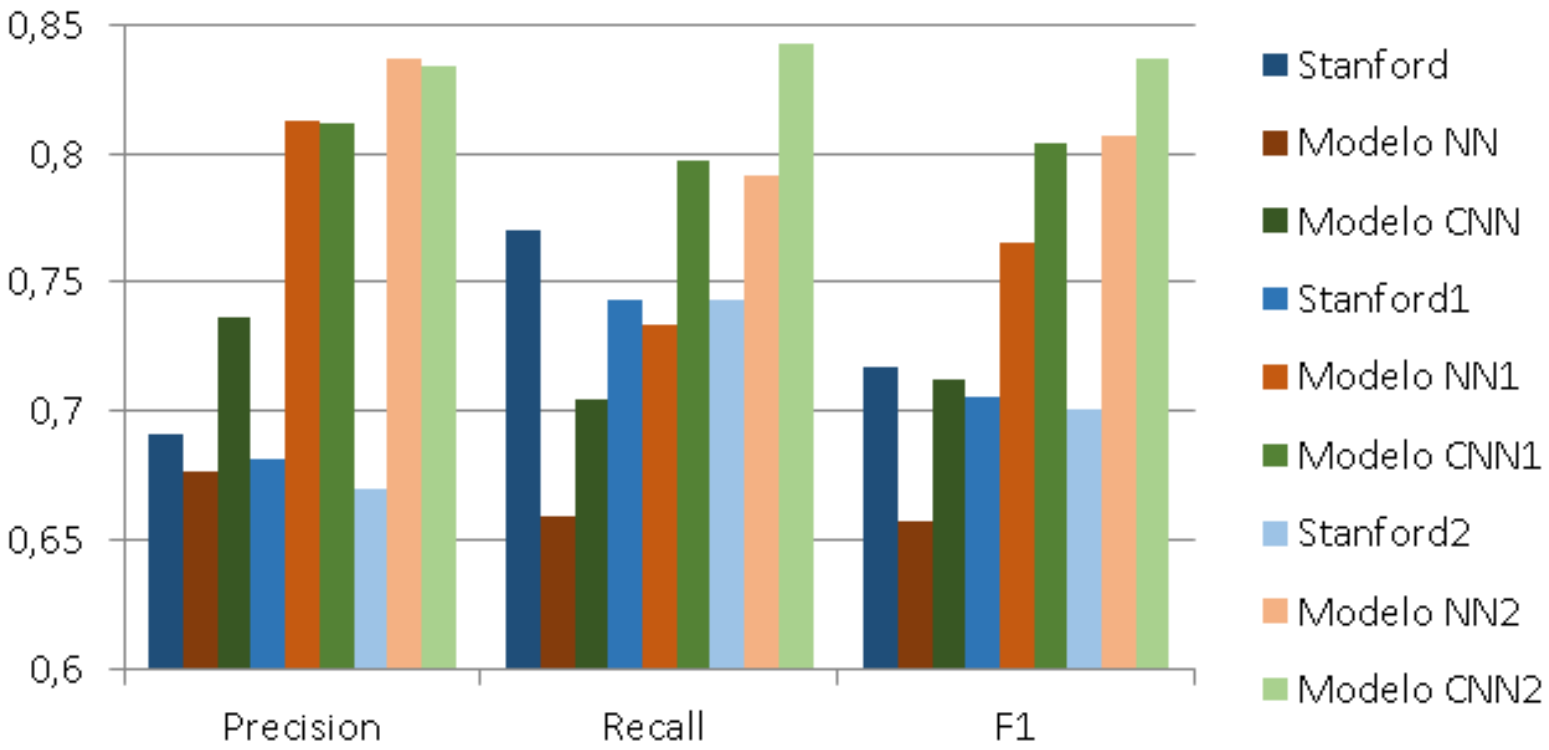


Figura 36: Medidas de desempeño de los modelos de reconocimiento de entidades

Resultados

armados ilegales y como centro de cultivos ilícitos . “En este sector ha sido constante la presencia de grupos armados . Ahí han operado paramilitares , las Farc , el Eln y bandas criminales que saben de su estratégica ubicación como corredor de movilidad” , dice el secretario de gobierno del Cauca , Amarildo Correa . La Policía señala que desde ahí se parte hacia El Plateado y luego son trochas y senderos hacia el Pacífico por los que se mueven armas , drogas y víveres para quienes viven en la ilegalidad . Otra forma de ingreso es desde el Pacífico por los caudalosos ríos que recorren la zona y que también sirven para sacar drogas . La Armada Nacional mantiene control en los ríos y esteros , lo que le ha permitido este año la incautación de 24.000 kilos de cocaína movilizados en lanchas que salen buscando alcanzar alta mar . De acuerdo con la Armada Nacional , que realiza operaciones en la Costa Pacífica , en la zona operan los frentes 60 , 30 y la columna Móvil Daniel Aldana . Allí se mueven los cabecillas de las Farc ‘Ramirito’ y ‘Pocillo’ . Este último fue a quien se le atribuyen los más sangrientos ataques contra la estación de Policía en El Mango y en la cabecera municipal de Argelia . Este grupo controla y se financia con los cultivos y laboratorios de coca . BOGOTÁ y POPAYÁN

Figura 37: Ejemplo de red neuronal convolucional - Semana 0

armados ilegales y como centro de cultivos ilícitos . “En este sector ha sido constante la presencia de grupos armados . Ahí han operado paramilitares , las Farc , el Eln y bandas criminales que saben de su estratégica ubicación como corredor de movilidad” , dice el secretario de gobierno del Cauca , Amarildo Correa . La Policía señala que desde ahí se parte hacia El Plateado y luego son trochas y senderos hacia el Pacífico por los que se mueven armas , drogas y víveres para quienes viven en la ilegalidad . Otra forma de ingreso es desde el Pacífico por los caudalosos ríos que recorren la zona y que también sirven para sacar drogas . La Armada Nacional mantiene control en los ríos y esteros , lo que le ha permitido este año la incautación de 24.000 kilos de cocaína movilizados en lanchas que salen buscando alcanzar alta mar . De acuerdo con la Armada Nacional , que realiza operaciones en la Costa Pacífica , en la zona operan los frentes 60 , 30 y la columna Móvil Daniel Aldana . Allí se mueven los cabecillas de las Farc ‘Ramirito’ y ‘Pocillo’ . Este último fue a quien se le atribuyen los más sangrientos ataques contra la estación de Policía en El Mango y en la cabecera municipal de Argelia . Este grupo controla y se financia con los cultivos y laboratorios de coca . BOGOTÁ y POPAYÁN

Figura 38: Ejemplo de red neuronal convolucional - Semana 2

Resultados

- Los modelos de deep learning tienen un comportamiento satisfactorio.

Resultados

- Los modelos de deep learning tienen un comportamiento satisfactorio.
- Los modelos pueden continuar aprendiendo.

Resultados

- Los modelos de deep learning tienen un comportamiento satisfactorio.
- Los modelos pueden continuar aprendiendo.
- Es posible reconocer nuevas entidades.

Modelos implementados

Modelos implementados

- Minhash/LSH: Datasketch, 5-gramas y 256 funciones de hash.

Modelos implementados

- Minhash/LSH: Datasketch, 5-gramas y 256 funciones de hash.
- Clustering: Promedio por parrafo, distancia coseno.

Modelos implementados

- Minhash/LSH: Datasketch, 5-gramas y 256 funciones de hash.
- Clustering: Promedio por parrafo, distancia coseno.
- LDA: Gensim, Scikit-learn.

Resultados

Modelos de temas

Por simple observación los modelos no generan resultados satisfactorios.

Resultados

Modelos de temas

Por simple observación los modelos no generan resultados satisfactorios.

"Si hay paz, las Fuerzas Militares y de Policía deben ser fortalecidas –afirma–. Recuerde que unas Fuerzas Militares fortalecidas dejan Estados soberanos y unas debilitadas dejan Estados fallidos. Además tenemos que construir la paz y para eso se requiere mantener fuerzas modernas y efectivas que puedan seguir manteniendo la seguridad y la soberanía.	http://www.eltiempo.com/politica/justicia/escandalo-corrupcion-corte-constitucional-victor-pacheco-ante-camara-de-acusaciones-/15361227	Pacheco, quien desde que estalló el escándalo ha mantenido bajo perfil y guardó silencio ante los graves señalamientos, tendrá que decir a los investigadores si el magistrado Jorge Pretelt le pidió 500 millones de pesos para fallar una tutela a favor de Fidupetrol. Aunque el magistrado Jorge Pretelt no sea procesado penalmente por el supuesto soborno para fallar la tutela a favor de Fidupetrol, juristas consideran que éticamente cometió una falta grave al reunirse con un abogado que tenía intereses en un proceso en el que Pretelt podría incidir, al menos con su voto.
	http://www.eltiempo.com/politica/partidos-politicos/elecciones-2015-candidatos-quemados-reciben-reposicion-de-votos/16413810	Además de 'pomada' para quemaduras, en el caso de las alcaldías, quienes hayan logrado al menos el 4 % de la votación total, recibirán 1.815 pesos por voto.Como una forma de garantizar la democracia, la Constitución Política establece que el Estado debe reponer parte del dinero a las campañas de los candidatos que hayan superado un umbral que establece la ley. Cada año, como lo establece la ley, el Consejo Nacional Electoral (CNE) fija nuevos valores a la reposición por voto válido obtenido por cada candidato. En el caso de los gobernadores, el CNE también hace reposición por voto válido a los candidatos si superaron el 4 % de la votación total.
	http://www.eltiempo.com/politica/justicia/los-mensajes-sobre-la-sexualidad-de-la-procuraduria/14038355	A través de varios mensajes reitera la importancia de decidir con responsabilidad y hace especial énfasis en que ante cualquier duda sobre la sentencia c-355/06, que habla de la despenalización del aborto solo en tres casos específicos, se consulte a la Procuraduría. Ese y otros, son los consejos de la Procuraduría a los jóvenes.

Resultados

Modelos de temas

Por simple observación los modelos no generan resultados satisfactorios.

"Si hay paz, las Fuerzas Militares y de Policía deben ser fortalecidas –afirma–. Recuerde que unas Fuerzas Militares fortalecidas dejan Estados soberanos y unas debilitadas dejan Estados fallidos. Además tenemos que construir la paz y para eso se requiere mantener fuerzas modernas y efectivas que puedan seguir manteniendo la seguridad y la soberanía.

Resultados

Modelos de temas

Por simple observación los modelos no generan resultados satisfactorios.

	http://www.eltiempo.com/politica/proceso-de-paz/alocucion-presidente-santos-sobre-acuerdo-de-eln/16549998	Lo que se buscará, según añadió el mandatario, es que haya mecanismos de "coordinación" entre las mesas de La Habana y de Quito (donde se realizará el diálogo con el Eln) para determinar cómo encontrar acuerdos en este punto en específico.Santos también destacó que este paso hacia la paz ratifica el compromiso de su Gobierno con acabar con más de 50 años de guerra, por lo que reconoció el paso que decidió dar la otra guerrilla que opera en el país."En cuanto al tema de terminación del conflicto armado, el objetivo es el mismo: erradicar la violencia de la política y propiciar el tránsito del Eln a la política legal, sin armas.
"Si hay paz, las Fuerzas Militares y de Policía deben ser fortalecidas –afirma–. Recuerde que unas Fuerzas Militares fortalecidas dejan Estados soberanos y unas debilitadas dejan Estados fallidos. Además tenemos que construir la paz y para eso se requiere mantener fuerzas modernas y efectivas que puedan seguir manteniendo la seguridad y la soberanía.	http://www.semana.com/nacion/articulo/gobierno-y-eln-santos-celebra-el-inicio-de-las-conversaciones-de-paz/467274	Al mediodía de este miércoles, el presidente Juan Manuel Santos se dirigió a todos los colombianos para anunciar el inicio formal del proceso de paz con la segunda guerrilla del país, el Ejército de Liberación Nacional (ELN).Santos calificó este anuncio como un nuevo paso hacia el fin del conflicto armado, el fin de las guerrillas en América Latina, y hacia una paz estable, duradera, y completa. El mandatario recordó que el ELN nació hace más de 50 años y es una organización con "su propia historia y su propia identidad", a la que se ha combatido sin descanso.
	http://www.semana.com/nacion/articulo/frank-pearl-secuestrados-que-tiene-el-eln-no-sabemos-cuantos-son/467381	Y algo que puede generar mayor preocupación, el Gobierno, que firmó un acuerdo con esta guerrilla para avanzar hacia el fin del conflicto, no conoce cuántos secuestrados tiene esta organización ilegal. Frank Pearl, jefe del equipo negociador del gobierno con el ELN, lo confesó públicamente este jueves, cuando presentó oficialmente a los otros integrantes del equipo negociador del Gobierno, Jaime Avendaño, el general (r) Eduardo Herrera y el exministro José Noé Ríos."No sabemos exactamente cuántos secuestrados tienen en su poder", dijo. La declaración tiene más trascendencia pues el miércoles, cuando le anunció al país el inicio del

Modelos implementados

- Textrank: Mihalcea[54], summa.
- Rake: Rose[69].

Aplicación de usuario final

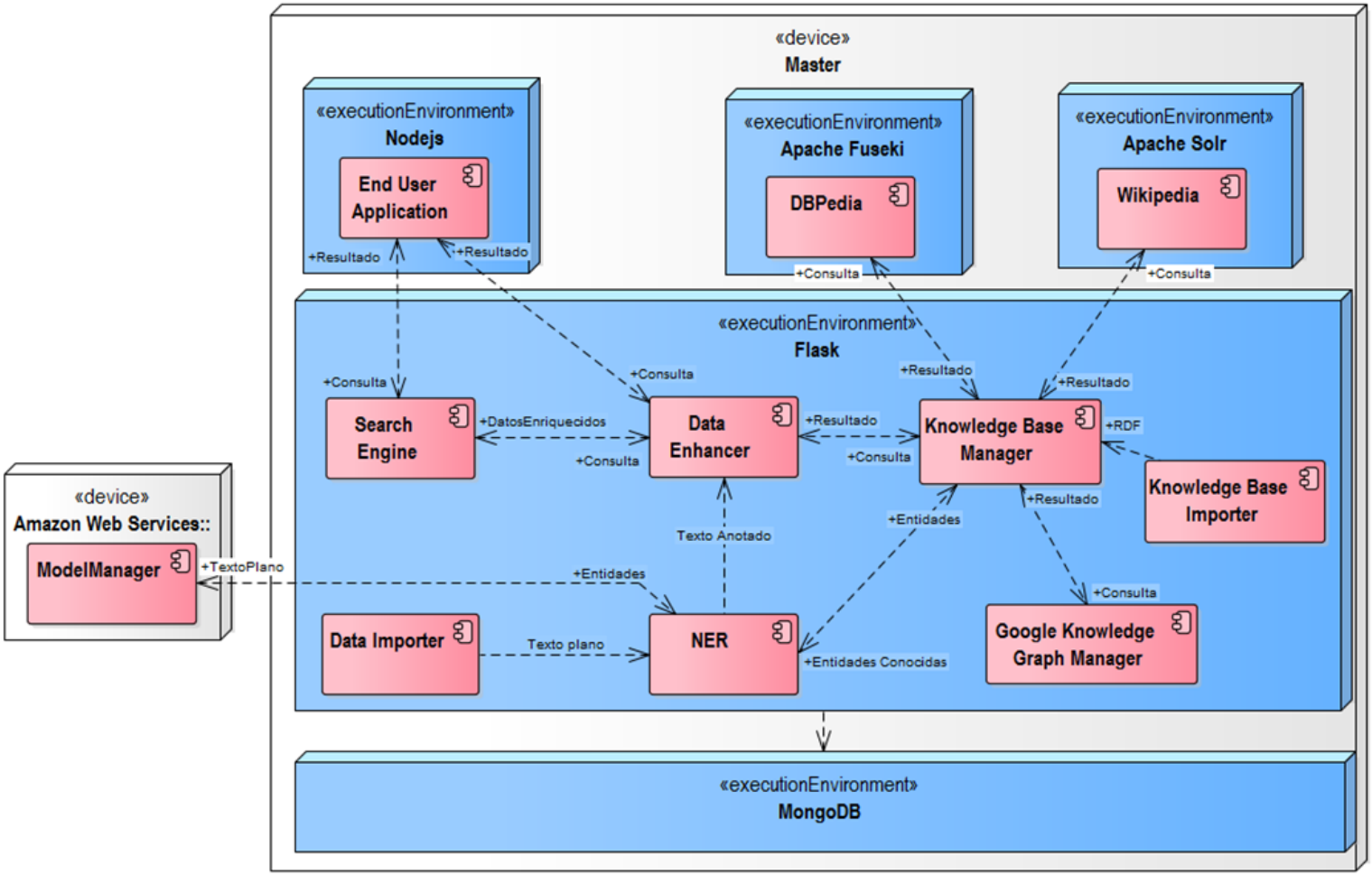


Figura 39: Despliegue de la aplicación de usuario final

Aplicación de usuario final

DEMO

Agenda

Conclusiones

- Se diseñó e implementó un sistema que permite construir contexto automáticamente en noticias de política colombiana y enriquecer el contenido por medio de bases de conocimiento y de la base de noticias recolectadas.

Conclusiones

- Se diseñó e implementó un sistema que permite construir contexto automáticamente en noticias de política colombiana y enriquecer el contenido por medio de bases de conocimiento y de la base de noticias recolectadas.
- Los modelos de reconocimiento de entidades y de temas permiten identificar relaciones en el contenido.

Conclusiones

- Se diseñó e implementó un sistema que permite construir contexto automáticamente en noticias de política colombiana y enriquecer el contenido por medio de bases de conocimiento y de la base de noticias recolectadas.
- Los modelos de reconocimiento de entidades y de temas permiten identificar relaciones en el contenido.
- Deep learning es una técnica que permite realizar análisis de contenido en español sin requerir un conocimiento detallado del dominio.

Conclusiones

- Se diseñó e implementó un sistema que permite construir contexto automáticamente en noticias de política colombiana y enriquecer el contenido por medio de bases de conocimiento y de la base de noticias recolectadas.
- Los modelos de reconocimiento de entidades y de temas permiten identificar relaciones en el contenido.
- Deep learning es una técnica que permite realizar análisis de contenido en español sin requerir un conocimiento detallado del dominio.
-

Trabajo futuro

Trabajo futuro

- Implementación de una base de conocimiento propia y especifica al dominio.

Trabajo futuro

- Implementación de una base de conocimiento propia y especifica al dominio.
  - Desambiguación y unificación de entidades.

Trabajo futuro

- Implementación de una base de conocimiento propia y especifica al dominio.
  - Desambiguación y unificación de entidades.
  - Proveer información específica de las entidades.

Trabajo futuro

- Implementación de una base de conocimiento propia y especifica al dominio.
  - Desambiguación y unificación de entidades.
  - Proveer información específica de las entidades.
- Mejorar el desempeño de los modelos.

Trabajo futuro

- Implementación de una base de conocimiento propia y especifica al dominio.
  - Desambiguación y unificación de entidades.
  - Proveer información específica de las entidades.
- Mejorar el desempeño de los modelos.
  - Mayor cantidad de datos anotados.

Trabajo futuro

- Implementación de una base de conocimiento propia y especifica al dominio.
  - Desambiguación y unificación de entidades.
  - Proveer información específica de las entidades.
- Mejorar el desempeño de los modelos.
  - Mayor cantidad de datos anotados.
  - Redes neuronales recurrentes.

Trabajo futuro

- Implementación de una base de conocimiento propia y especifica al dominio.
  - Desambiguación y unificación de entidades.
  - Proveer información específica de las entidades.
- Mejorar el desempeño de los modelos.
  - Mayor cantidad de datos anotados.
  - Redes neuronales recurrentes.
  - Deep Belief Networks.

Biliografía

[1] Ola Amayri y Nizar Bouguila. “Online news topic detection and tracking via localized feature selection”. En: **The 2013 International Joint Conference on Neural Networks (IJCNN)** (2013), págs. 1-8.

[2] Brenda Reyes Ayala y Cornelia Caragea. “Towards building a collection of web archiving research articles”. En: **Proceedings of the American Society for Information Science and Technology** 51.1 (2014), págs. 1-5.

[3] Surya Bahadur Bam y Tej Bahadur Shahi. “Named Entity Recognition for Nepali Text Using Support Vector Machines”. En: **Intelligent Information Management** Ma6rch (2014), págs. 21-29.

[4] Marco Baroni, Georgiana Dinu y German Kruszewski. “Don’t count , predict ! A systematic comparison of context-counting vs . context-predicting semantic vectors”. En: **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.** (2014), págs. 238-247.

[5] Yoshua Bengio y col. “A Neural Probabilistic Language Model”. En: **The Journal of Machine Learning Research** 3 (2003), págs. 1137-1155.

[6] Christian Bizer. “DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia”. En: **Semantic Web Journal** 6.2 (2012), págs. 167-195.

Biliografía (cont.)

[7] Christian Bizer, Tom Heath y Tim Berners-Lee. “Linked data-the story so far”. En: **International journal on Semantic Web and Information Systems** 5.3 (2009), págs. 1-22.

[8] David M Blei y John D Lafferty. “Dynamic Topic Models”. En: **International Conference on Machine Learning** (2006), págs. 113-120.

[9] David M Blei, Andrew Y Ng y Michael I Jordan. “Latent Dirichlet Allocation”. En: **Journal of Machine Learning Research** 3 (2003), págs. 993-1022.

[10] a Bordes y J Weston. “Learning structured embeddings of knowledge bases”. En: **Conference on Artificial . . .** (2011), págs. 301-306.

[11] Florin Bulgarov y Cornelia Caragea. “A Comparison of Supervised Keyphrase Extraction Models”. En: **The International World Wide Web Conference**. 2015, págs. 13-14.

[12] Cornelia Caragea, Florin Bulgarov y Rada Mihalcea. “Co-Training for Topic Classification of Scholarly Data”. En: **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Vol. 2. September. 2015, págs. 2357-2366.

Biliografía (cont.)

[13] Cornelia Caragea y col. “Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach”. En: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. 2008. 2014, págs. 1435-1446.

[14] Cornelia Caragea y col. “Document Type Classification in Online Digital Libraries”. En: **The Twenty-Eighth Annual Conference on Innovative Applications of Artificial Intelligence**. 2016.

[15] Cristian Cardellino. **Spanish Billion Words Corpus and Embeddings**. 2016.

[16] X Carreras, L Marquez y L Padró. “Named entity extraction using adaboost”. En: **Proceedings of the 6Th Conference on . . .** (2002), págs. 0-3.

[17] Soumen Chakrabarti, Martin Van Den Berg y Byron Dom. “Focused crawling: A new approach to topic-specific Web resource discovery”. En: **Computer Networks** 31.11 (1999), págs. 1623-1640.

[18] Rachel Chasin, Daryl Woodward y Jugal Kalita. “Extracting and displaying temporal entities from historical articles”. En: **National Conference on Machine Learning, Tezpur University, Assam, India** (2011), págs. 1-13.

Biliografía (cont.)

[19] Deepti Chopra y Sudha Morwal. “Named entity recognition in english language using Hidden Markov Model”. En: **Internation Journal on Computational Sciences & Appications (IJCSA)** 3.1 (2013), págs. 1-6.

[20] A Das y U Garain. “CRF-based Named Entity Recognition”. En: (2014).

[21] Sujatha Das Gollapalli y col. “Researcher homepage classification using unlabeled data”. En: **Proceedings of the 22nd . . .** (2013), págs. 471-481.

[22] S. T. Dumais y col. “Using latent semantic analysis to improve access to textual information”. En: **Proceedings of the SIGCHI conference on Human factors in computing systems - CHI ’88** (1988), págs. 281-285.

[23] Charles Elkan. “Text mining and topic models”. En: **Lecture notes** (2014).

[24] Jenny Rose Finkel, Trond Grenager y Christopher Manning. “Incorporating non-local information into information extraction systems by gibbs sampling”. En: **in Acl** 1995 (2005), págs. 363 -370.

[25] Guohong Fu y Kang-Kwong Luke. “Chinese named entity recognition using lexicalized HMMs”. En: **ACM SIGKDD Explorations Newsletter** 7.1 (2005), págs. 19-25.

Biliografía (cont.)

[26] Jianfeng Gao y col. “Modeling Interestingness with Deep Neural Networks”. En: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)** (2014), págs. 2-13.

[27] Tong Gao y col. “NewsViews: An Automated Pipeline for Creating Custom Geovisualizations for News”. En: **Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI ’14** (2014), págs. 3005-3014.

[28] Wei Gao, Peng Li y Kareem Darwish. “Joint topic modeling for event summarization across news and social media streams”. En: **Cikm** OCTOBER (2012), pág. 1173.

[29] C Lee Giles, Kurt D Bollacker y Steve Lawrence. “CiteSeer: An Automatic Citation Indexing System”. En: **ACM Conference on Digital Libraries** (1998), págs. 89-98.

[30] Sujatha Das Gollapalli y Cornelia Caragea. “Extracting Keyphrases from Research Papers Using Citation Networks”. En: **Proceedings of the 28th American Association for Artificial Intelligence** (2014), págs. 1629-1635.

[31] Sujatha Das Gollapalli y col. “Document analysis and retrieval tasks in scientific digital libraries”. En: **Communications in Computer and Information Science** 505 (2015), págs. 3-20.

Biliografía (cont.)

[32] Sujatha Das Gollapalli y col. “Improving Researcher Homepage Classification with Unlabeled Data”. En: **ACM Transactions on the Web (TWEB)** 9.4 (2015), pág. 32.

[33] Felipe González Casabianca. “Desarrollo y evaluación de una infraestructura de monitoreo y modelaje de actividades mediante flujo de datos con estrategias pervasive”. Tesis doct. Universidad de los Andes, 2015.

[34] Anja Gruenheid y col. “StoryPivot : Comparing and Contrasting Story Evolution”. En: **Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data** 1 (2015), págs. 1415-1420.

[35] Xiulan Hao y Yunfa Hu. “Topic detection and tracking oriented to BBS”. En: **2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering, CMCE 2010** 4 (2010), págs. 154-157.

[36] Zellig S. Harris. “Distributional Structure”. En: **Word** 10.2-3 (1954), págs. 146-162.

[37] Qi He y col. “Context-aware citation recommendation”. En: **Proceedings of the 19th international conference on World wide web WWW 10** (2010), pág. 421.

Biliografía (cont.)

[38] Sepp Hochreiter y col. “Gradient flow in recurrent nets: the difficulty of learning long-term dependencies”. En: **A Field Guide to Dynamical Recurrent Networks** (2001), págs. 237-243.

[39] Lei Hou y col. “NewsMiner: Multifaceted news analysis for event search”. En: **Knowledge-Based Systems** 76 (2015), págs. 17-29.

[40] Wenyi Huang y col. “RefSeer: Citation Recommendation System”. En: **Joint Conference on Digital Library 2014**. 2014, págs. 2-5.

[41] Jessica Hullman, Nicholas Diakopoulos y Eytan Adar. “Contextifier : Automatic Generation of Annotated Stock Visualizations”. En: **SIGCHI - Conference on Human Factors in Computing Systems** (2013), págs. 2707-2716.

[42] Hideki Isozaki. “Japanese named entity recognition based on a simple rule generator and decision tree learning”. En: **Proceeding ACL ’01 Proceedings of the 39th Annual Meeting on Association for Computational Linguistics** (2001), págs. 314-321.

[43] D Jannach y col. **Recommender systems: an introduction**. Vol. 40. 2011, págs. 1-335.

[44] Rie Johnson y Tong Zhang. “Effective Use of Word Order for Text Categorization with Convolutional Neural Networks”. En: **Naacl** 2011 (2015), págs. 103-112.

Biliografía (cont.)

[45] Rie Johnson y Tong Zhang. “Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding”. En: **Nips** (2015), págs. 1-9.

[46] Diederik P. Kingma y Jimmy Lei Ba. “Adam: A method for stochastic optimization”. En: **3rd International Conference for Learning Representations**. 2015, págs. 1-15.

[47] Jana Kravalov. “Czech Named Entity Corpus and SVM-based Recognizer”. En: **NEWS ’09 Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration**. 2009, págs. 194-201.

[48] Milos Krstajic y col. “Story Tracker: Incremental visual text analytics of news story development”. En: **Information Visualization** 12.3-4 (2013), págs. 308-323.

[49] Martha Ladly y col. “The CBC Newsworld Holodeck”. En: **Extended Abstracts on Human Factors in Computing Systems** (2014), págs. 363-366.

[50] Gregor Leban y col. “Event Registry – Learning About World Events From News”. En: **Www** (2014), págs. 107-110.

[51] Honglak Lee y col. “Unsupervised feature learning for audio classification using convolutional deep belief networks”. En: **Nips workshop on deep learning for speech recognition and related applications** (2009), págs. 1-9.

Biliografía (cont.)

[52] Dongning Luo y col. “EventRiver: Visually Exploring Text Collections with Temporal References”. En: **IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS** 18.1 (2012), págs. 93-105.

[53] Qiaozhu Mei y ChengXiang Zhai. “Discovering evolutionary theme patterns from text: an exploration of temporal text mining”. En: **Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining** (2005), págs. 198-207.

[54] Rada Mihalcea y Paul Tarau. “TextRank: Bringing order into texts”. En: **Proceedings of EMNLP** 85 (2004), págs. 404-411.

[55] David Mimno y col. “Optimizing Semantic Coherence in Topic Models”. En: **Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing** 2 (2011), págs. 262-272.

[56] Naji F. Mohammed y Nazlia Omar. “Arabic named entity recognition using artificial neural network”. En: **Journal of Computer Science** 8.8 (2012), págs. 1285-1293.

[57] Sudha Morwal, Nusrat Jahan y Deepti Chopra. “Named Entity Recognition using Hidden Markov Model (HMM)”. En: 1.4 (2012), págs. 15-23.

Biliografía (cont.)

[58] Kishore Neppalli y col. “MetaSeer . STEM : Towards Automating Meta-Analyses”. En: **Proceedings of the Twenty-Eighth Annual Conference on Innovative Applications of Artificial Intelligence**. 2016.

[59] David Newman y col. “Analyzing Entities and Topics in News Articles using Statistical Topic Modeling”. En: **IEEE International Conference on Intelligence and Security Informatics**. 2006.

[60] Thien Huu Nguyen y Ralph Grishman. “Relation Extraction: Perspective from Convolutional Neural Networks”. En: **Workshop on Vector Modeling for NLP** (2015), págs. 39-48.

[61] Lawrence Page y col. “The PageRank Citation Ranking: Bringing Order to the Web”. En: **World Wide Web Internet And Web Information Systems** 54.1999-66 (1998), págs. 1-17.

[62] Georgios Paliouras y col. “Learning Decision Trees for Named-Entity Recognition and Classification”. En: **ECAI Workshop on Machine Learning for Information Extraction**. 2000.

[63] Matteo Palmonari y col. “DaCENA: Serendipitous News Reading with Data Contexts”. En: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)** 7540 (2015), págs. 73-86.

Biliografía (cont.)

[64] Natalia Ponomareva y col. “Conditional Random Fields vs. Hidden Markov Models in a biomedical Named Entity Recognition task”. En: **Proc. of Int. Conf. Recent . . .** (2007).

[65] Yves Raimond y col. “Using the past to explain the present: Interlinking current affairs with archives via the semantic web”. En: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)** 8219 LNCS.PART 2 (2013), págs. 146-161.

[66] Anand Rajaraman y Jeffrey D Ullman. “Mining of Massive Datasets”. En: **Lecture Notes for Stanford CS345A Web Mining** 67 (2011), pág. 328.

[67] Lev Ratinov y Dan Roth. “Design challenges and misconceptions in named entity recognition”. En: **Proceedings of the Thirteenth Conference on Computational Natural Language Learning CoNLL 09** June (2009), pág. 147.

[68] Lev Ratinov y Joseph Turian. “Word representations : A simple and general method for semi-supervised learning”. En: **Acl** July (2010), págs. 384-394.

[69] Stuart Rose y col. “Automatic keyword extraction from individual documents”. En: **Text Mining: Applications and Theory** (2010), págs. 1-277.

Biliografía (cont.)

[70] Amarappa S y Sathyanarayana S.V. “Kannada Named Entity Recognition and Classification (NERC) Based on Multinomial Naïve Bayes (MNB) Classifier”. En: **International Journal on Natural Language Computing** 4.4 (2015), págs. 39-52.

[71] Gerard Salton y Christopher Buckley. “Term-weighting approaches in automatic text retrieval”. En: **Information Processing & Management** 24.5 (1988), págs. 1-21.

[72] Cicero Nogueira dos Santos y Maira Gatti. “Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts”. En: **Coling-2014** (2014), págs. 69-78.

[73] Sunita Sarawagi y William W. Cohen. “Semi-markov conditional random fields for information extraction”. En: **Advances in Neural Information Processing Systems 17** (2004), págs. 1185-1192.

[74] Richard Socher y col. “Reasoning With Neural Tensor Networks for Knowledge Base Completion”. En: **Proceedings of the Advances in Neural Information Processing Systems 26 (NIPS 2013)** (2013), págs. 1-10.

[75] René Speck y Axel-Cyrille Ngonga Ngomo. “Ensemble Learning for Named Entity Recognition”. En: **The Semantic Web – ISWC 2014**. 2014, págs. 519-534.

Biliografía (cont.)

[76] Daniel Sarmiento Suárez y Claudia Jiménez-Guarín. “Natural Language Processing for Linking Online News and Open Government Data Proposed Solution : SPEAk”. En: Springer International Publishing, 2014, págs. 243-252.

[77] Yaming Sun y col. “Modeling mention, context and entity with neural networks for entity disambiguation”. En: **IJCAI International Joint Conference on Artificial Intelligence** 2015-Janua.Ijcai (2015), págs. 1333-1339.

[78] Charles Sutton y Andrew McCallum. “An Introduction to Conditional Random Fields”. En: **Foundations and Trends® in Machine Learning** 42.2 (2010), págs. 105-117.

[79] Charles Sutton y Andrew McCallum. “An Introduction to Conditional Random Fields for Relational Learning”. En: ().

[80] György Szarvas, Richárd Farkas y András Kocsor. “A Multilingual Named Entity Recognition System Using Boosting and C4 . 5 Decision Tree Learning Algorithms”. En: **Discovery Science**. 2006, págs. 267-278.

[81] Koichi Takeuchi y Nigel Collier. “Use of Support Vector Machines in Extended Named Entity Recognition”. En: **COLING-02 proceedings of the 6th conference on Natural language learning - Volume 20**. 2002, págs. 1-7.

Biliografía (cont.)

[82] Jie Tang y col. “Extraction and Mining of an Academic Social Network”. En: **Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining** (2008), págs. 1193-1194.

[83] Lubaba. F. Tanisha y col. “Analyzing and Visualizing News Trends Over Time”. En: **International Conference on Industrial Engineering and Engineering Management** (2014), págs. 307-311.

[84] Vinh Tuan Thai y Siegfried Handschuh. “IVEA: Toward a Personalized Visual Interface for Exploring Text Collections”. En: **Proceedings of the 14th International Conference on Intelligent User Interfaces** (2009), págs. 479-480.

[85] Nam Khanh Tran y col. “Back to the Past : Supporting Interpretations of Forgotten Stories by Time-aware Re-Contextualization”. En: **Proceedings of the Eighth ACM International Conference on Web Search and Data Mining** (2015), págs. 339-348.

[86] L J P Van Der Maaten y G E Hinton. “Visualizing high-dimensional data using t-sne”. En: **Journal of Machine Learning Research** 9 (2008), págs. 2579-2605.

[87] Peng Wang y col. “Semantic Clustering and Convolutional Neural Network for Short Text Categorization”. En: **Proceedings ACL 2015** (2015), págs. 352-357.

Biliografía (cont.)

[88] Jian Wu y col. “CiteSeerX : AI in a Digital Library Search Engine”. En: **Proceedings of the Twenty-Sixth Annual Conference on Innovative Applications of Artificial Intelligence**. 2014, págs. 2930-2937.

[89] Jian Wu y col. “PDFMEF: A Multi-Entity Knowledge Extraction Framework for Scholarly Documents and Semantic Search”. En: **Proceedings of the 8th International Conference on Knowledge Capture** (2015), pág. 13.

[90] Yu-chieh Wu y col. “Support Vector Machines”. En: **Knowledge Discovery in Life Science Literature**. 2006, págs. 91-103.

[91] Tze-I Yang, Andrew J. Torget y Rada Mihalcea. “Topic modeling on historical newspapers”. En: **LaTeCH ’11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities** (2011), págs. 96-104.

[92] Bohai Yu, Xia Zhang y Zhengyou Xia. “News Event Detection Based Web Big Data”. En: **11th International Conference** 7996.July (2013), págs. 443-449.

[93] Shengkang Yu y col. “Tracking news article evolution by dense subgraph learning”. En: **Neurocomputing** 168 (2015), págs. 1076-1084.

Biliografía (cont.)

[94] Xiang Zhang, Junbo Zhao y Yann LeCun. “Character-level Convolutional Networks for Text Classification”. En: **Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH** (2015), págs. 3057-3061.

[95] Ye Zhang y Byron Wallace. “A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification”. En: **arXiv preprint arXiv:1510.03820** 1 (2015).

[96] GuoDong Zhou y Jian Su. “Named entity recognition using an HMM-based chunk tagger”. En: **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics ACL 02** July (2002), págs. 473-480.

Conclusiones

- Se diseñó e implementó un sistema que permite construir contexto automáticamente en noticias de política colombiana y enriquecer el contenido por medio de bases de conocimiento y de la base de noticias recolectadas.
- Los modelos de reconocimiento de entidades y de temas permiten relacionar las noticias.

Algoritmo de inferencia - Clasificador Bayesiano

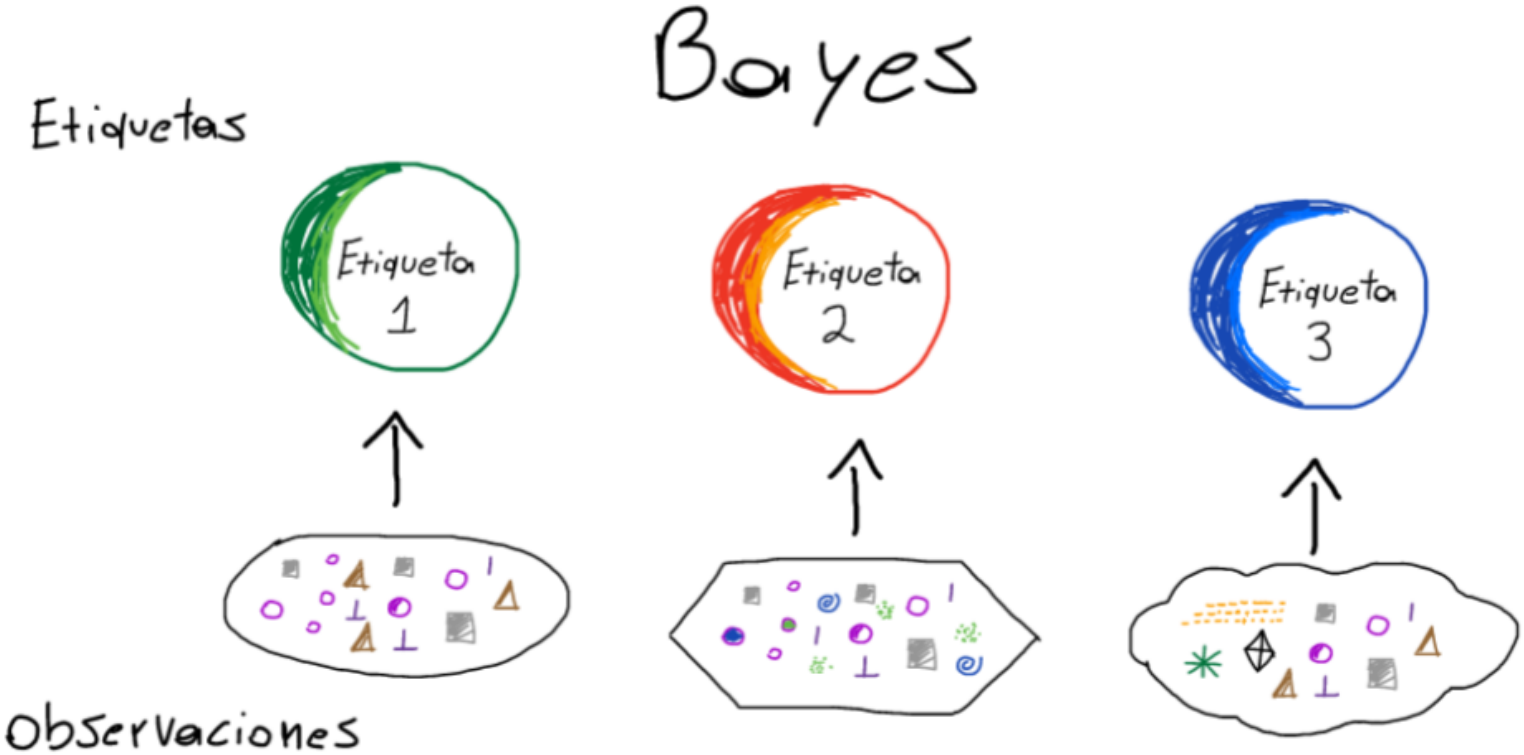


Figura 40: Naïve Bayes [54]

$$\underset{c_j \in C}{argmax}\, P(c_j) \prod_i p(x_i|c_j)$$

Algoritmo de inferencia - Modelos ocultos de Markov

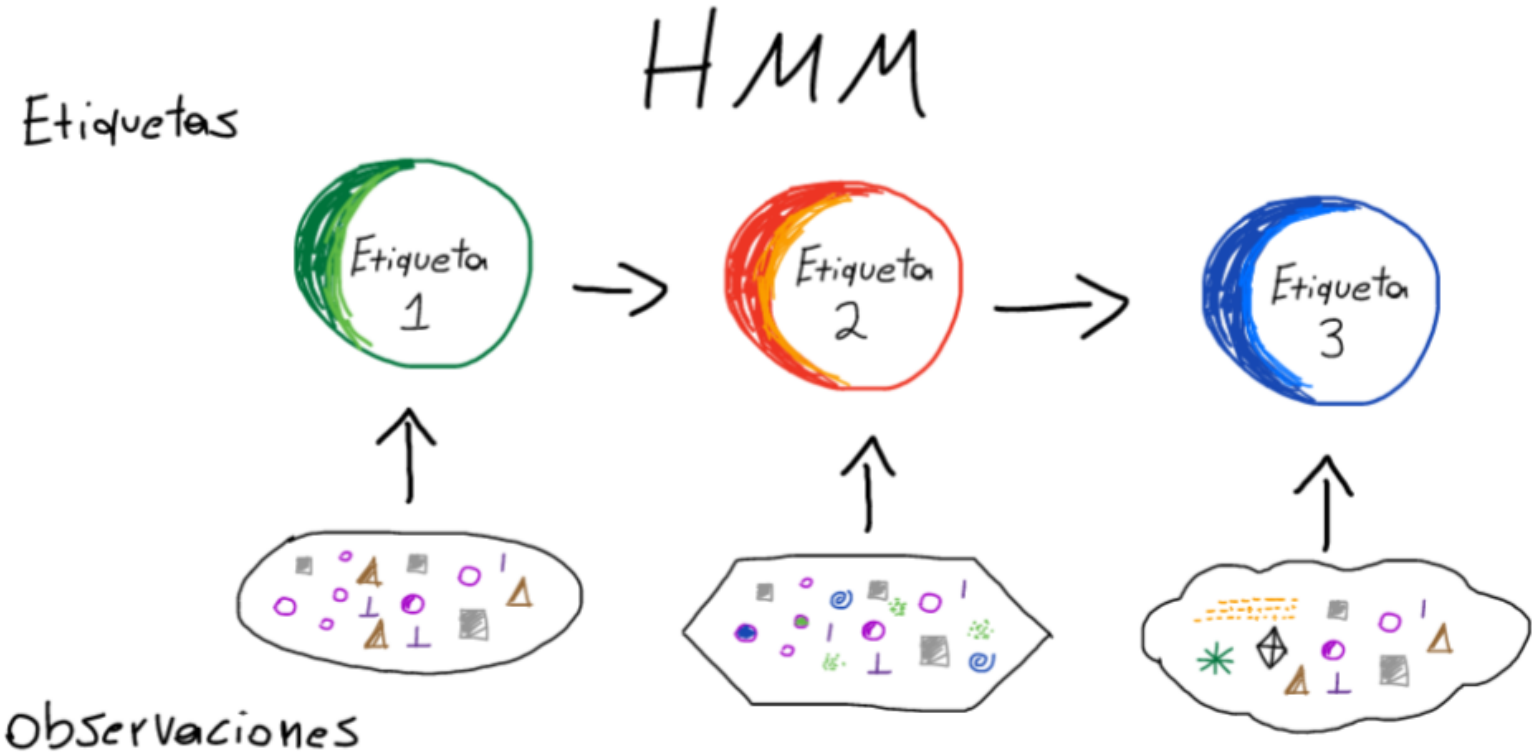


Figura 41: Modelos ocultos de Markov [54]

$$\underset{c_j \in C}{argmax}\, P(c_j|c_{j-1}) \prod_i p(x_i|c_j)$$

Algoritmo de inferencia - Conditional Random Fields

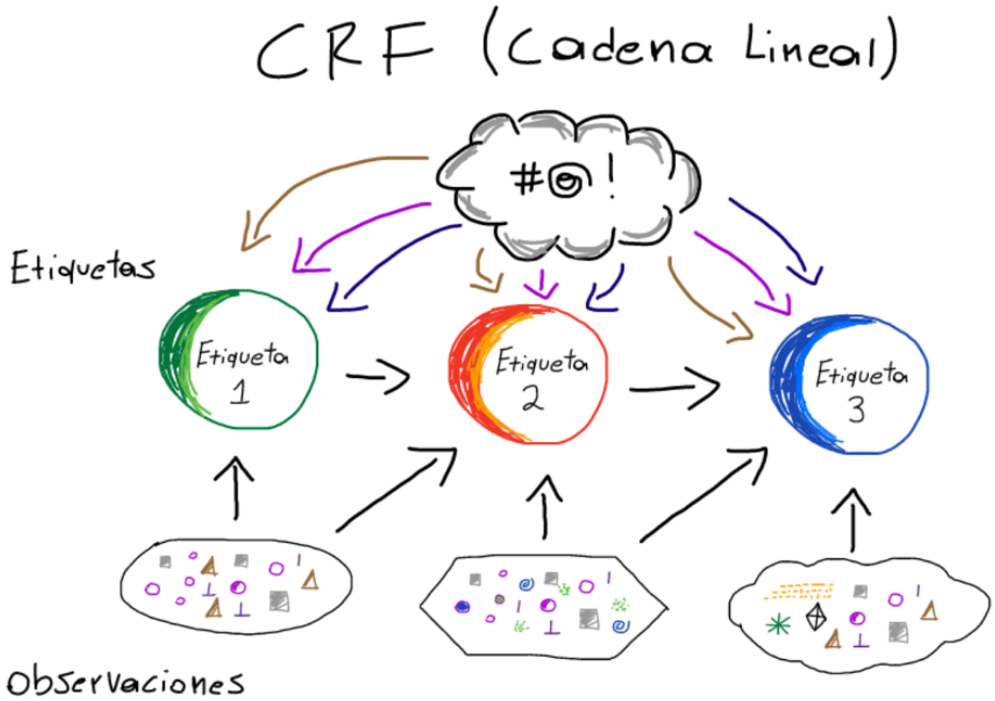


Figura 42: Conditional Random Fields [54]

$$score(l|s) = \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i, l_{i-1})$$

Algoritmo de inferencia - Arboles de decisión

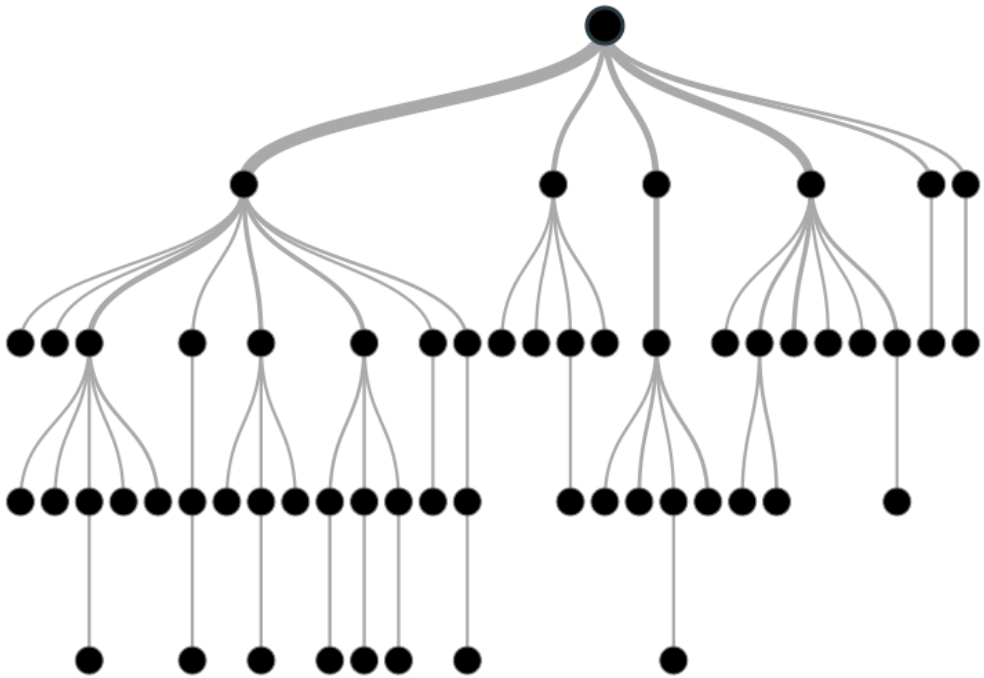


Figura 43: Arbol de decisión

$$X = \bigcap_{k=1}^{K} x^k$$

Algoritmo de inferencia - Máquina de soporte vectorial

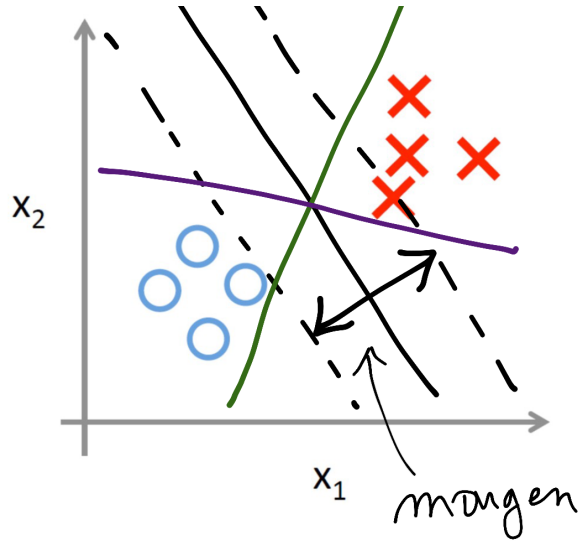


Figura 44: Máquina de soporte vectorial

$$\underset{w}{argmin}\, C \sum_{i=1}^{n} [y^{(i)} costo_1(w^T x^{(i)}) + (1 - y^{(i)}) costo_0(w^T x^{(i)})]$$