# Notas del desarrollo para la documentación del proceso de creación del módulo de detección de entidades

## Recopilación de datasets

Noticias, tweets (de la zona de argentina; violentos; que contenían la palabra mujer), frases comunes de google, etc.

Los tweets los extrajimos utilizando la API de twitter, y posteriormente hicimos un trabajo de limpieza eliminando: menciones, links, hashtags y tweets realizados por bots o en inglés.

## Entrenamiento 1: Frases comunes

Etiqueté aproximadamente 2000 frases de este dataset. De esta exploración y etiquetado pude observar que muchas frases eran muy similares entre sí, por lo que su etiquetado manual pudo haber sido realizado de forma más eficaz usando augmentations. Además en esta primera ronda de etiquetado consideré como entidades PERS palabras como ‘amigo, madre, padre, tío’, luego de charlar con nuestro director concluimos que sería mejor solo etiquetar aquellas entidades que son explícita y específicamente lo que se está etiquetando, para lograr mejores resultados (PERS solo nombres propios, MEDICIÓN solo número seguido por unidad, MISC todas aquellas entidades que podrían encuadrar en alguna de las entidades definidas pero no está muy claro).

Corriendo con un batch =1000, frases de entrenamiento=1800…(configuración base autocompletada de spacy, este modelo corriendo 5 loops de entrenamiento logró un **fscore de 54** (best fue en el 3 loop, y en los últimos dos bajo). Se podría correr nuevamente con un batch un poco menor (no menos de 800, porque si se deben calcular muchos gradientes al aumentar el número de batches se dificulta el backpropagation)

Al probar este primer modelo con una frase sencilla "Juan Martinez nació en Hawaii el 10 de Marzo de 1961." El modelo identificó

Juan Martinez PERSONA

Hawaii LUGAR

Marzo PERSONA

Y cuando cambié el nombre de Juan Martinez a Barack Obama, no detectó a la persona.

La fecha no fue detectada ya que en el dataset la mayoría de los números están escritos (no por digitos)

## Entrenamiento 1: Tweets de Argentina

Se etiquetaron aproximadamente 2500 tweets. Las frases eran muy diferentes entre sí, tanto en composición, acciones, referencias y estructura. Consideré como entidad PERSONA a sustantivos comunes como “hijo, abuela, yo, él”. También tuve varias ambiguedades al clasificar las entidades FECHA/HORA/TIEMPO, donde varias veces intercambié el criterio para etiquetarlo. Las mayores ambiguedades se dieron cuando clasifiqué a las entidades MISC, ya que en ese caso el criterio inicial era muy amplio. Terminé clasificando una gran cantidad de sustantivos comunes que no cumplían ninguno de los criterios para que sean una entidad de las otras clasificaciones como “vida, libro, país”.

Corriendo con un batch =1000, frases de entrenamiento=1800, con 14 loops de entrenamiento se logró un **fscore de 27** (en el quinto loop). Se puede correr con un batch menor.

Al probar este primer modelo con una frase sencilla "Juan Martinez nació en Hawaii el 10 de Marzo de 1961." El modelo identificó:

10 de Marzo HORA

1961 TIEMPO

## Entrenamiento 2: Frases comunes

Después de definir una forma más clara de considerar las entidades, seleccione 200 frases del conjunto anterior para poder etiquetarlas usando este nuevo enfoque (un 10% de la cantidad de frases anterior) para hacer una primera prueba, y observar los resultados. Si estos son favorables incorporaría más frases a etiquetar manualmente o intentaremos aplicar un método de ‘active learning’ para el etiquetado (si fuera posible)

(Luego del etiquetado agregue 32 frases más, para completar algunas entidades que eran poco representadas en el conjunto, por lo que el tamaño real del dataset reducido resultó de 232 frases)

El bajo número de ejemplos afecta el **f score**, que en el caso anterior rondaba casi el 60 ahora llega a **31**

Al probar este primer modelo con una frase sencilla "Juan Martinez nació en Hawaii el 10 de Marzo de 1961." El modelo identificó:

Juan MISC

Hawaii LUGAR

10 de Marzo FECHA

1961 DINERO

## Entrenamiento 2: Tweets de Argentina

Se etiquetaron alrededor de 200 tweets seleccionados. Se restringieron de manera más estricta los criterios para cada entidad: para PERSONA por ejemplo solo se toman nombres propios, para LUGAR solo el sustantivo que indica un lugar específico, MEDIDA solo un cantidad de cosas cuantificable, y para MISC solo entidades que pueden cumplir con cualquiera de las otros criterios pero de manera ambigua sin el contexto que tenemos.

El primer resultado tenía un bajo número de ejemplos, por lo que el **F score** bajó con respecto a los 2500 tweets, y el mejor resultado fue de 20. Sin embargo el descenso no fue tan pronunciado como con las frases comunes, donde el número bajó de 60 a 31. En cambio con los tweets bajó de 27 a **20**, lo que puede indicar que la mayor cohesión del etiquetado compensó en parte el descenso abrupto de ejemplos.

Al probar este primer modelo con una frase sencilla "Juan Martinez nació en Hawaii el 10 de Marzo de 1961." El modelo identificó:

Marzo PERSONA

Se hizo una mejora al dataset, agregando 100 tweets mas, y haciendo algunos ajustes en la etiquetación que se hizo anteriormente. Por ejemplo, se removieron ejemplos en los que palabras como “dia” o “noche” eran considerados como FECHA o “semana” era considerado TIEMPO. Se vió una mejora del fscore, llegando hasta 27. Cuando se lo sometió a la prueba "Juan Martinez nacio en Hawaii el 10 de Marzo de 1961." el modelo identificó:

10 de Marzo de 1961 HORA

Además, cuando se lo sometió a la frase expandida: "Juan Martinez nacio en Hawaii el 10 de Marzo de 1961. Durante la decada del 80, empezó a aparecer en varias películas. En 1981, viajó a Argentina y fundó dos empresas. Un mes despues, en diciembre, chocó con su auto en la playa. Llamó a su amigo Luis para decirle que perdió la memoria. Cuando Luis le preguntó como recordaba su nombre, Juan se dio cuenta de que estaba bien. Le pagó 2000 pesos por ayudarlo." El modelo identificó:

10 de Marzo de 1961 HORA

80 MISC

Argentina LUGAR

dos empresas TIEMPO

playa LUGAR

## Entrenamiento 2: Combinación de tweets y frases

Tomando los dos datasets (al rededor de 400 ejemplos) con un batch =150 se llego a un modelo que logró un **fscore de 29**.

Debido a estos resultados concluimos que la naturaleza de los tweets, hacen que estos no sean los mejores para poder entrenar un sistema de detección de entidades.

Al probar este primer modelo con una frase sencilla "Juan Martinez nació en Hawaii el 10 de Marzo de 1961." El modelo identificó:

Juan Martinez PERSONA

Hawaii MISC

10 de Marzo HORA

1961 FECHA

También se hizo la prueba con la frase expandida: “Juan Martinez nacio en Hawaii el 10 de Marzo de 1961. Durante la decada del 80, empezó a aparecer en varias películas. En 1981, viajó a Argentina y fundó dos empresas. Un mes despues, en diciembre, chocó con su auto en la playa. Llamó a su amigo Luis para decirle que perdió la memoria. Cuando Luis le preguntó como recordaba su nombre, Juan se dio cuenta de que estaba bien. Le pagó 2000 pesos por ayudarlo." El modelo identificó:

Juan Martinez PERSONA

Hawaii MISC

10 de Marzo HORA

1961 FECHA

Durante PERSONA

80, empezó a aparecer en varias películas. En 1981 FECHA

Argentina PERSONA

diciembre LUGAR

playa LUGAR

Luis PERSONA

Juan MISC

2000 pesos DINERO

Para la misma frase de prueba expandida, se probó usando el modelo que combina las 200 frases comunes con la versión expandida de 300 tweets. El modelo identificó:

Juan Martinez PERSONA

Hawaii PERSONA

10 de Marzo de 1961 FECHA

Durante PERSONA

80 MISC

1981 LUGAR

Argentina PERSONA

playa LUGAR

Luis PERSONA

Luis PERSONA

Juan PERSONA

2000 pesos DINERO

## Entrenamiento 2 con vectores: Combinación de tweets v2 y frases

## Para el mismo dataset compuesto por las frases comunes y los 300 tweets expandidos, se probó el entrenamiento con el uso de vectores. El pipeline utilizado fue el de “es\_core\_news\_lg”. La velocidad de entrenamiento bajó notoriamente, pero el resultado fue mucho mejor para el mismo dataset.

## El **fscore máximo obtenido fue de 52**. Se probó con la frase de prueba expandida: "Juan Martinez nacio en Hawaii el 10 de Marzo de 1961. Durante la decada del 80, empezó a aparecer en varias películas. En 1981, viajó a Argentina y fundó dos empresas. Un mes despues, en diciembre, chocó con su auto en la playa. Llamó a su amigo Luis para decirle que perdió la memoria. Cuando Luis le preguntó como recordaba su nombre, Juan se dio cuenta de que estaba bien. Le pagó 2000 pesos por ayudarlo.". El modelo identificó:

Juan Martinez PERSONA

Hawaii LUGAR

10 de Marzo de 1961 FECHA

80 TIEMPO

Argentina MISC

dos empresas MEDIDA

mes TIEMPO

playa LUGAR

Luis PERSONA

Luis PERSONA

Juan PERSONA

pagó 2000 pesos DINERO

## Entrenamiento 3: Expansión de frases comunes.

Debido a los problemas mencionados del dataset de frases comunes decidí tomar el dataset de 200 frases, expandirlo con: +450 frases generadas usando Chat GPT 3, y más frases del dataset original (pero esta vez etiquetadas con el criterio más riguroso)

Se llego a un dataset de alrededor de 650 frases etiquetadas.

-En un primer lugar se entrenó el modelo con el pipeline de entrenamiento vacía, con un batch size de 200, este modelo llegó a un **fscore de 74.4**

## Se probó con la frase de prueba expandida: "Juan Martinez nacio en Hawaii el 10 de Marzo de 1961. Durante la decada del 80, empezó a aparecer en varias películas. En 1981, viajó a Argentina y fundó dos empresas. Un mes despues, en diciembre, chocó con su auto en la playa. Llamó a su amigo Luis para decirle que perdió la memoria. Cuando Luis le preguntó como recordaba su nombre, Juan se dio cuenta de que estaba bien. Le pagó 2000 pesos por ayudarlo.". El modelo identificó:

Juan Martinez PERSONA

Hawaii LUGAR

10 de Marzo FECHA

1981 LUGAR

Argentina LUGAR

Luis PERSONA

Luis PERSONA

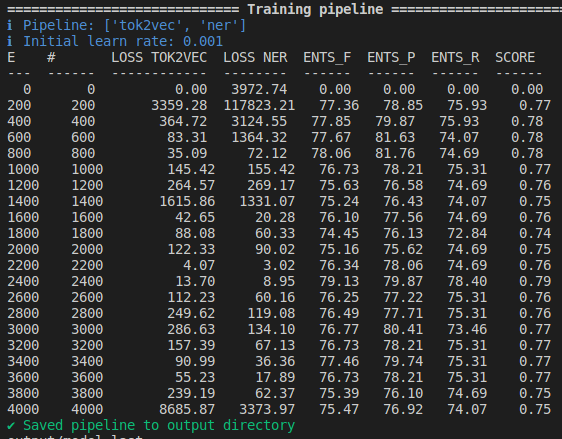
recordaba FECHA

Juan PERSONA

2000 pesos DINERO

## Entrenamiento 3b: Expansión de frases comunes con configuración alterada para mayor precisión

Se entrenó al modelo con el mismo dataset que el anterior, pero modificando unos pocos atributos clave en el config.cfg y agregando un pipeline preentrenado es\_core\_news\_md (no se uso el es\_core\_news\_lg porque era muy pesado). El entrenamiento se volvió mucho más lento, pero el **fscore** aumentó considerablemente: en el mejor resultado llegó a **79**.



Se probó con una frase de prueba expandida:

*Juan Martinez nacio en Hawaii el 10 de Marzo de 1961. Durante la decada del 80 , empezó a aparecer en varias películas. En 1981, viajó a Argentina y fundó dos empresas. Un mes despues, en diciembre, chocó con su auto en la playa. Llamó a su amigo Luis para decirle que perdió la memoria. Cuando Luis le preguntó como recordaba su nombre, Juan se dio cuenta de que estaba bien. Le pagó 2000 pesos por ayudarlo. Luis viajó dos noches despues, en el teatro municipal, para ir a un recital de musica de su grupo favorito, Los Rockeros. Durante el recital, a las 0 AM, se prendieron 12 luminarias para celebrar la llegada de Año Nuevo. Luis y sus amigos laura y marcos prendieron tres velas para acompañar. Se fueron a sus casas, alrededor de las tres y media de la mañana. Tardaron 25 minutos en llegar.*

El modelo identificó:  
Juan Martinez PERSONA

Hawaii MISC

10 de Marzo FECHA

Argentina LUGAR

dos empresas MEDIDA

diciembre FECHA

playa LUGAR

Luis PERSONA

Luis PERSONA

Juan PERSONA

2000 pesos DINERO

Luis PERSONA

dos noches MEDIDA

municipal LUGAR

0 AM HORA

12 luminarias MEDIDA

Año Nuevo MISC

Luis PERSONA

tres velas MEDIDA

tres y media HORA

25 minutos TIEMPO

## Entrenamiento 4: Expansión de frases comunes + 300 tweets.

-Se entrenó el modelo con el pipeline de entrenamiento vacía, con un batch size de 200, este modelo llegó a un **fscore de 56.5**

## Se probó con la frase de prueba expandida: "**Juan Martinez** nacio en **Hawaii** el **10 de Marzo de 1961**. Durante la decada del 80, empezó a aparecer en varias películas. En **1981**, viajó a **Argentina** y fundó **dos empresas**. **Un mes** despues, en diciembre, chocó con su auto en la playa. Llamó a su amigo **Luis** para decirle que perdió la memoria. Cuando **Luis** le preguntó como recordaba su nombre, **Juan** se dio cuenta de que estaba bien. Le pagó **2000 pesos** por ayudarlo.". El modelo identificó:

Juan Martinez PERSONA

10 de Marzo FECHA

Argentina LUGAR

diciembre LUGAR

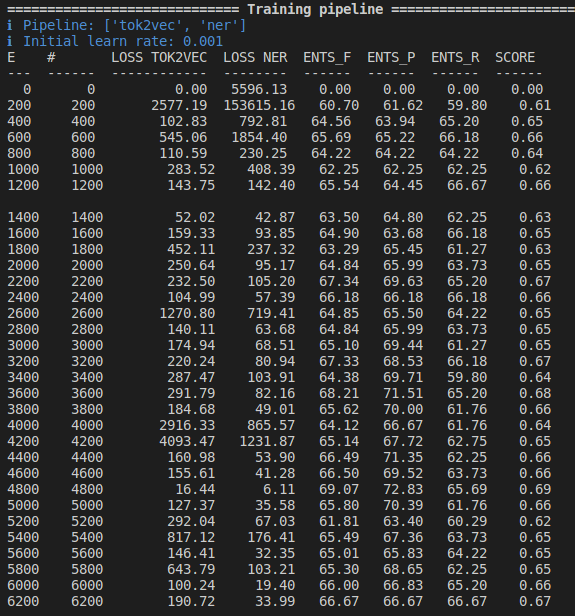
playa LUGAR

Juan PERSONA

2000 pesos DINERO

## Entrenamiento 4b: Expansión de frases comunes + 300 tweets con configuración alterada para mayor precisión

Se entrenó al modelo con el mismo dataset que el anterior (expansión de frases comunes + 300 tweets, pero modificando unos pocos atributos clave en el config.cfg y agregando un pipeline preentrenado es\_core\_news\_md (no se uso el es\_core\_news\_lg porque era muy pesado). El entrenamiento se volvió muchísimo más lento. Después de varias y hasta 6000 iteraciones, **fscore** aumentó notablemente: en el mejor resultado llegó a **69.**



Se probó con una frase de prueba expandida:

*Juan Martinez nacio en Hawaii el 10 de Marzo de 1961. Durante la decada del 80 , empezó a aparecer en varias películas. En 1981, viajó a Argentina y fundó dos empresas. Un mes despues, en diciembre, chocó con su auto en la playa. Llamó a su amigo Luis para decirle que perdió la memoria. Cuando Luis le preguntó como recordaba su nombre, Juan se dio cuenta de que estaba bien. Le pagó 2000 pesos por ayudarlo. Luis viajó dos noches despues, en el teatro municipal, para ir a un recital de musica de su grupo favorito, Los Rockeros. Durante el recital, a las 0 AM, se prendieron 12 luminarias para celebrar la llegada de Año Nuevo. Luis y sus amigos laura y marcos prendieron tres velas para acompañar. Se fueron a sus casas, alrededor de las tres y media de la mañana. Tardaron 25 minutos en llegar.*

Juan Martinez PERSONA

Hawaii el LUGAR

10 de Marzo de 1961 FECHA

Argentina LUGAR

dos empresas MEDIDA

playa LUGAR

Luis PERSONA

Luis PERSONA

Juan PERSONA

2000 pesos DINERO

Luis PERSONA

dos noches TIEMPO

teatro municipal LUGAR

0 AM HORA

12 luminarias TIEMPO

Año Nuevo MISC

Luis PERSONA

tres velas MEDIDA

tres y media HORA

25 minutos TIEMPO

## Entrenamiento 4.3 Con transformers:

ChatGPT + 200 frases + tweets + noticias sin transformers llego a 61 de fscore

ChatGPT + 200 frases + noticias sin transformers llego a 67 de fscore

Con los entrenamientos anteriores, el crecimiento del dataset y los cambios en la configuración para mejorar el score tuvieron un efecto muy importante en el tiempo que se necesitaba para entrenar el modelo.

Explorando la performance durante el entrenamiento, se observó que siempre se usaba un solo thread del procesador y una cantidad mínima de RAM, por lo que había un embotellamiento debido a la configuración de Spacy.

La solución para poder habilitar el uso de más recursos es el uso de los transformers. Mediante unos cambios en la configuración, se pudo usar casi todos los cores al mismo tiempo y el uso de la RAM subió a varios GB.

Además, se observó que los score de los modelos (frases + chatgpt + noticias) y (frases + chatgpt + tweets + noticias) se acercaron mucho, 77 y 79 respectivamente.

