Introducción al Procesamiento de Lenguaje Natural - Informe de Laboratorio

Parte 1

Esta parte del laboratorio tiene por objetivo experimentar con modelos de lenguaje para la generación de textos (modelos text-to-text) y a la vez, poner en práctica conceptos y herramientas vistos en el curso, como vectorización n-gramas y mínima distancia de edición. En particular se prueban los modelos Maximum Likelihood Estimator (MLE) disponible en la librería scikit-learn y multilingual-e5-large a través de hugging face. Para esta primera parte se utilizará el texto "Cuentos de Amor de Locura y de Muerte" de Horacio Quiroga. A continuación se presentan los pasos utilizados para la extracción del texto.

Extracción de los textos

El texto original contiene metadata que no es de interés para los experimentos, en Snippet_1 se puede ver como se extrae y pre-procesa el texto. Para su extracción se utiliza la función re.search de Python. Adicionalmente, se eliminan espacios en blanco al inicio y fin del string, con la función strip. Por último, se reemplazan los saltos de línea seguidos de una palabra en minúscula por un espacio en blanco (" "), usando el método re.sub.

Experimentando con n-gramas y pre-procesamientos

Por razones de tiempo, se acotaron los experimentos en base a diferentes pre-procesamientos a siete pipelines diferentes: n-gramas de 2 a 5 palabras, pasar a minúsculas, reemplazar saltos de línea y símbolos de puntuación. En la sección de Anexos, en la Tabla 1 se encuentra la lista completa de pre-procesamientos así como una breve descripción de cada uno. Para la generación, se utilizan los textos `"las vacas"`, `"el cielo"` y adicionalmente se introduce el texto `"Esteban Podeley"` para analizar resultados utilizando un texto poco frecuente.
A modo ilustrativo, a continuación se presentan los resultados obtenidos para el texto `"las vacas"`. En la sección Anexos, en Tabla 2 se encuentran los resultados obtenidos para los tres textos y todos los pre-procesamientos.

idx	pipeline	prompt	output
1	pipeline_1 n=4	las vacas	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con
2	pipeline_2 n=3	las vacas	dormitaban al sol ya caliente , rumiando . Pero en el patio y Alfonso la llamó en silencio . Pasábanse horas sin oir el angustioso
3	pipeline_3 n=2	las vacas	dormitaban al fin se había reforzado su corazón siempre en la pipa y el perro había retirado . A media hora en tierra , del
4	pipeline_4 n=5	las vacas	dormitaban al sol ya caliente , rumiando . Pero cuando los pobres caballos pasaron por el camino , ellas abrieron los ojos

			despreciativas : --
5	pipeline_5 n=4, lowercase	las vacas	estaban inmóviles , mirando fijamente el verde paraíso inalcanzable . -- ¿por qué ? ¿qué le pasa ? -- nada , sino que está bien
6	pipeline_6 n=4, lowercase, replace \n “ “	las vacas	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con
7	pipeline_7 n=4, lowercase, rep(\n, “ “) rep(punc_sym)	las vacas	dormitaban al sol ya caliente rumiando Pero cuando los pobres caballos pasaron por el camino ellas abrieron los ojos despreciativas Son los caballos Los alambres

Los pipelines 1, 2, 4 y 7, producen sentencias parcialmente correctas gramaticalmente. Sin embargo, otros pipelines, como el 2 y 3, generan sentencias menos coherentes y con mayores errores gramaticales. Esto sugiere que reducir el valor de N en los n-gramas tiene un impacto negativo en la correctitud sintáctica de las sentencias. Por lo tanto, usar un N mayor (como N=4 o N=5) parece mejorar la generación de sentencias correctas. En cuanto a los otros pre-procesamientos, pasar a minúsculas no parece afectar mayormente en la generación de sentencias, e incluso puede generar resultados diferentes, posiblemente por la unificación de palabras similares. En cuanto a remover saltos de línea adicionales, no tiene un impacto significativo por sí solo, y eliminar los símbolos de puntuación tampoco produce mejoras, ya que solo genera oraciones más largas sin mejorar sustancialmente los resultados, algo que también se puede lograr aumentando el parámetro de longitud.

Mínima distancia de edición

Los experimentos de mínima distancia de edición producen 63 resultados (7 pipelines x 3 vecinos x 3 textos de ejemplo), por lo tanto se selecciona el pipeline 6 y el texto generado a partir del prompt `"las vacas"`, para desarrollar algunas observaciones. Todos los resultados para este prompt se pueden encontrar en la Tabla 3, mientras que los 63 resultados en el notebook. **Texto generado:** *“dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con”.*

idx	neigh_idx	sentencia
15	1	Pero cuando está conmigo, entonces no aparta los ojos de ellos.
16	2	Arrizabalaga y la señora se reían, volviéndose a menudo, y la joven no apartaba casi sus ojos de Nébel.
17	3	El caballo, por mayor intimidad de trato, es sensiblemente más afecto al hombre que la vaca.

En la tabla se puede observar que dentro del top 3 de vectores más cercanos, se encuentra la sentencia 1 que es muy similar a la segunda oración del texto generado. En particular, analizando todos los resultados obtenidos (ver todos los resultados en el notebook), resulta evidente que para N mayores como n=4 o n=5 se generan oraciones con un alto nivel de similitud a sus vecinos más cercanos. A su vez, convertir a minúsculas o remover saltos de línea refuerza este resultado. Por otro lado, para N menores (bigramas y trigramas), los textos generados tienen un bajo nivel de coincidencia (en palabras) con los vecinos más

cercanos.
Por otro lado, es evidente que la estrategia de n-gramas con N grandes tiende a sobre-ajustar el modelo a los datos de entrenamiento. Observar como con un prompt poco frecuente, `"Esteban Podeley"`, se obtiene un texto bastante diferente a sus 3 vecinos más cercanos. **Texto generado:** *“, peones de obraje , volvían a Posadas en el _Silex_ , con quince compañeros . Podeley , cuya fiebre anterior había tenido honrado y”.*

idx	neigh_idx	sentencia
57	1	Durante el viaje había sido un excelente compañero, admirando por su cuenta y riesgo, y hablando poco.
58	2	Reverberaba ahora delante de ellos un pequeño páramo de greda que ni siquiera se había intentado arar.
59	3	Me desperté, y volví a soñar: el tal salón de baile estaba frecuentado por los muertos diarios de una epidemia.

Sentence Transformers

Mediante los embeddings de sentence transformers y utilizando la distancia euclídea (default) para Nearest Neighbors, se obtienen resultados más interesantes. Por ejemplo, repitiendo el análisis realizado con la distancia de edición, mediante el pipeline 6, se obtienen los siguientes resultados. **Texto generado:** *“dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con”.*

idx	neigh_idx	sentencia
15	1	Detrás de él, las vacas dormitaban al sol ya caliente, rumiando.
16	2	Pero cuando está conmigo, entonces no aparta los ojos de ellos.
17	3	Y juro que fueron fuertes las dos horas que pasamos mi mujer y yo, con la luz prendida hasta que amaneció, ella acostada, yo sentado en la cama, vigilando sin cesar la arpillera flotante.

Se puede observar que la primera parte del texto generado, coincide en gran medida con el primer vecino. Además, la segunda parte del texto generado, coincide con el segundo vecino. Todos los resultados para este texto se pueden ver en la Tabla 4.
Por otro lado, para el ejemplo de la oración con el prompt `"Esteban Podeley",` se obtienen oraciones más cercanas al texto generado. Para ver los 63 resultados obtenidos, ver el notebook con el código.

idx	neigh_idx	sentencia
57	1	#LOS MENSÚ# Cayetano Maidana y Esteban Podeley, peones de obraje, volvían a Posadas en el _Silex_, con quince compañeros.
58	2	El _Silex_ volvió a Posadas, llevando con él al mensú empapado aún en pesadillas nocturnas.
59	3	Podeley, cuya fiebre anterior había tenido honrado y periódico ritmo, no presagió nada bueno para él de esa galopada de accesos casi sin intermitencia.

Distancia edición vs. sentence transformers

A modo de comparación, ambos enfoques son bien diferentes y permiten obtener resultados muy distintos:

- La distancia de edición mide la similitud de caracteres, funcionando bien para sentencias idénticas o muy similares, pero falla si difieren en palabras o en el orden de las mismas.
- La distancia euclídea en embeddings de sentence-transformers captura la semántica, permitiendo comparar sentencias con variaciones en longitud, palabras y orden, encontrando vecinos cercanos semánticamente.

Observaciones sobre n-gramas

De los experimentos realizados, se infieren las siguientes observaciones sobre la estrategia de n-gramas para vectorización de textos:

- Utilizar valores de *n* grandes (como *n* = 4) en modelos de *n*-gramas permite capturar mejor el contexto y generar oraciones más correctas gramaticalmente.
- Sin embargo, estos modelos suelen sobre-ajustarse al conjunto de entrenamiento, generando textos muy similares al original y repitiendo fragmentos largos cuanto mayor sea el valor de *n*.
- En conclusión, los modelos basados en *n*-gramas no generalizan bien, ya que sus representaciones son esparsas y capturan poco contexto, a diferencia de los embeddings pre-entrenados que son más compactos y complejos en información contextual.

Parte 2

Esta parte del laboratorio, se centra en el entrenamiento de modelos de clasificación de texto, utilizando oraciones de diferentes autores.

Análisis del corpus

En la siguiente tabla se muestran la cantidad de sentencias para cada autor (clase) en cada uno de los conjuntos de datos. De la misma, se deduce que el dataset presenta un fuerte desbalance, siendo la clase dominante Quiroga. En particular en Train, las proporciones son: 54% Quiroga, 38% Becquer y 7.37% Martin Fierro. Esto seguramente traiga asociado dificultades adicionales al momento de entrenar modelos de clasificación.

Corpus	**Quiroga**	**Becquer**	**Martin Fierro**	**Total**
Train	2469	1727	334	4530
Dev	529	370	72	971
Test	530	370	71	971

Entrenando modelos con datos de dev

A continuación se listan los resultados obtenidos para diferentes combinaciones de pre-procesamientos, técnicas de vectorización y modelos. El detalle de los parámetros utilizados en cada uno de los modelos, así como lo que se espera probar con cada uno de ellos, se puede encontrar en la Tabla 5, en la sección de Anexos.

id	Repr. Text.	Modelo	Precision	Recall	F1-score
0	-	Random Choice	0.32	0.32	0.28
1	BoW	SVM	0.53	0.39	0.33
2	TF-IDF	SVM	0.48	0.49	0.48
3	BoW	MLP Classifier	0.75	0.46	0.46
4	TF-IDF	MLP Classifier	0.86	0.60	0.60
5	TF-IDF + SMOTE	MLP Classifier	0.73	0.68	0.70
6	**Sentence Transformers + SMOTE**	**SVM**	**0.86**	**0.85**	**0.85**
7	Sentence Transformers + SMOTE	MLP Classifier	0.81	0.86	0.83

Para cada modelo no solo importa los valores de Precision, Recall y F1-score, sino también analizar en detalle cómo es la performance en cada clase. A priori, salvo que exista una determinación particular sobre ponderar la performance de una clase sobre otra (por ejemplo en el contexto de fraud-detection ponderar más los falsos positivos que falsos negativos), interesa que la performance sea más o menos equitativa entre clases.

A modo de ejemplo, BoW + SVM, tiende a sobre-ajustar sobre los textos de la clase *Quiroga*, generando un sesgo de predicción en esta clase. Notar que no se capturó ningún ejemplo de la clase *Martin Fierro*.

```
F-Score macro: 33.49
              precision    recall  f1-score   support

      quiroga       0.58      1.00      0.73       529
      becquer       1.00      0.16      0.27       370


martin_fierro       0.00      0.00      0.00        72

     accuracy                           0.60       971
    macro avg       0.53      0.39      0.33       971
 weighted avg       0.70      0.60      0.50       971
```

Este problema se ataca de forma efectiva, mediante oversampling con SMOTE. Como se puede ver en la tabla anterior en el experimento nro 6, aumenta tanto la Precision como el

Recall macro. A su vez, cuando se observan los resultados por clase, se puede ver que el modelo logra generalizar todas las clases sin desarrollar sesgo en predicción.

```
F-Score macro: 85.48
               precision    recall  f1-score   support

      quiroga       0.86      0.93      0.89       529
      becquer       0.90      0.80      0.85       370
martin_fierro       0.83      0.82      0.83        72

     accuracy                           0.87       971
    macro avg       0.86      0.85      0.85       971
 weighted avg       0.87      0.87      0.87       971
```

Evaluación de modelos en Test

A continuación se muestran los resultados de evaluación sobre el dataset de test, para los tres modelos con los que se obtuvieron mejores resultados utilizando la partición de dev. En Classification Report 1 se pueden ver a su vez por clase para el modelo nro 6.

id	Repr. Text.	Modelo	Precision	Recall	F1-score
5	TF-IDF + SMOTE	MLP Classifier	0.8	0.75	0.77
6	**Sentence Transformers + SMOTE**	**SVM**	**0.87**	**0.87**	**0.87**
7	Sentence Transformers + SMOTE	MLP Classifier	0.82	0.86	0.84

Conclusiones

- Los resultados evidencian la necesidad de abordar el desbalance en el dataset para lograr un clasificador aceptable en todas las clases. Oversampling mediante SMOTE**,** parece ser una estrategia más que eficaz para resolver este problema.
- Aunque sentence-transformers es la técnica de vectorización más efectiva, con TF-IDF también se logran resultados razonables. A priori, si no existe ninguna restricción que impida el uso de transformers, parece más que razonable explorar otros modelos pre-entrenados para comparar dentro del universo de embeddings pre-entrenados. Por otro lado, BoW por sí mismo, no permite obtener resultados aceptables en este problema.
- SVM en combinación con Sentence-Transformers arroja los mejores resultados, incluso mejor que MLP Classifier, aunque parezca contraintuitivo. Esto se explica por la complejidad extra que tiene el MLP Classifier, para obtener buenos resultados. En particular es necesario ajustar los distintos parámetros como capas ocultas, funciones de activación, optimizadores, entre otros, para alcanzar una arquitectura óptima. En este trabajo se trabajó con valores de referencia.
- La performance de **Sentence-Transformers + SVM** es consistente entre los datasets de desarrollo y test, lo cual sugiere que el modelo generaliza bien sin sobre-ajustarse a los datos de entrenamiento.

Anexos

Snippet_1

Ejemplo de código utilizado para extracción de texto quiroga, así como pre-procesamiento inicial, indicado en el notebook (por más detalles referirse al notebook complementario a este informe).

```
# 1. Extraemos texto útil
quiroga_text_extract_pattern = r"#Cuentos de Amor de Locura y de
Muerte#(.*)FIN\n"  #
match = re.search(quiroga_text_extract_pattern, quiroga_raw, re.DOTALL)

# 2. Eliminamos espacios en blanco al inicio y final
quiroga_text = match.group(1).strip()

# 3. Removemos saltos de línea con palabra en minúscula a la derecha
breakline_lowercase_pattern = r"\n([a-z])"
quiroga = re.sub(breakline_lowercase_pattern, r" \1", quiroga_text)
```

Tabla 1

Resumen de los diferentes pipelines de procesamientos utilizados para experimentación.

Nombre	**Parámetros**	**Objetivo**
pipeline_1	n=4	Probar el efecto de n grande
pipeline_2	n=3	Probar el efecto reducir n
pipeline_3	n=2	Probar el efecto reducir n
pipeline_4	n=5	Probar el efecto n más grande
pipeline_5	n=4, lowercase()	Para un n fijo, agregar convertir a minúsculas
pipeline_6	n=4, lowercase(), replace(“\n”, “ ”)	Para un n fijo, agregar eliminar todos los saltos
pipeline_7	n=4, lowercase(), replace(“\n”, “ ”), replace(punc_symbols, “ “)	Para un n fijo, además eliminar símbolos de puntuación y caracteres especiales

Tabla 2

Ejemplos de textos generados por el modelo de n-gramas y MLE para diferentes pre-procesamientos.

idx	pipeline	prompt	output
0	pipeline_1	las vacas	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con
1	pipeline_2	las vacas	dormitaban al sol ya caliente , rumiando . Pero en el patio y Alfonso la llamó en silencio . Pasábanse horas sin oir el angustioso
2	pipeline_3	las vacas	dormitaban al fin se había reforzado su corazón siempre en la pipa y el perro había retirado . A media hora en tierra , del
3	pipeline_4	las vacas	dormitaban al sol ya caliente , rumiando . Pero cuando los pobres caballos pasaron por el camino , ellas abrieron los ojos despreciativas : --
4	pipeline_5	las vacas	estaban inmóviles , mirando fijamente el verde paraíso inalcanzable . -- ¿por qué ? ¿qué le pasa ? -- nada , sino que está bien
5	pipeline_6	las vacas	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con
6	pipeline_7	las vacas	dormitaban al sol ya caliente rumiando Pero cuando los pobres caballos pasaron por el camino ellas abrieron los ojos despreciativas Son los caballos Los alambres
7	pipeline_1	el cielo	constantemente encapotado y lluvioso , provocáronle verdaderas alucinaciones de perros que entraban al trote por la portera . Había un motivo real para este temor
8	pipeline_2	el cielo	constantemente encapotado y lluvioso , provocáronle verdaderas alucinaciones de perros que no vaya yo jamás a explicarme qué combinaciones de visitas , casamientos y garden
9	pipeline_3	el cielo	de que está todo el piso . Alrededor del menguante . Pero yo con sorpresa . Pero éste : la loma , y bajo el
10	pipeline_4	el cielo	constantemente encapotado y lluvioso , provocáronle verdaderas alucinaciones de perros que entraban al trote por la portera . Había un motivo real para este temor
11	pipeline_5	el cielo	constantemente encapotado y lluvioso , provocáronle verdaderas alucinaciones de perros que entraban al trote por la portera . había un motivo real para este temor
12	pipeline_6	el cielo	constantemente encapotado y lluvioso , provocáronle verdaderas alucinaciones de perros que entraban al trote por la portera . Había un motivo real para este temor
13	pipeline_7	el cielo	fijo en sequía con chubascos de cinco minutos se descomponía por fin en las lagartijas Aún en noviembre cuando tenía ya en jaque a todas
14	pipeline_1	Esteban Podeley	, peones de obraje , volvían a Posadas en el _Silex_ , con quince compañeros . Podeley , cuya fiebre anterior había tenido honrado y

15	pipeline_2	Esteban Podeley	, más ansiosa aún . Recurrió entonces a un hombre discreto . Véase : Fuí a lo que es patrimonio específico de los corazones inferiores
16	pipeline_3	Esteban Podeley	bajaron tambaleantes de todo el piso . Alrededor del menguante . Pero yo con sorpresa . Pero éste : la loma , y bajo el
17	pipeline_4	Esteban Podeley	, peones de obraje , volvían a Posadas en el _Silex_ , con quince compañeros . Podeley , labrador de madera , tornaba a los
18	pipeline_5	Esteban Podeley	esperaba una lluvia , y salté de costado , con las rodillas recogidas hasta el pecho . ¿qué sería ? y la respiración también ...
19	pipeline_6	Esteban Podeley	, peones de obraje , volvían a Posadas en el _Silex_ , con quince compañeros . Podeley , cuya fiebre anterior había tenido honrado y
20	pipeline_7	Esteban Podeley	peones de obraje volvían a Posadas en el _Silex_ con quince compañeros Podeley labrador de madera tornaba a los nueve meses la contrata concluída y

Tabla 3

idx	prompt	model_name	output_text	idx	neighborh
0	las vacas	pipeline_1	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con	1	Pero cuando está conmigo, entonces no aparta los ojos de ellos.
1	las vacas	pipeline_1	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con	2	Arrizabalaga y la señora se reían, volviéndose a menudo, y la joven no apartaba casi sus ojos de Nébel.
2	las vacas	pipeline_1	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con	3	El caballo, por mayor intimidad de trato, es sensiblemente más afecto al hombre que la vaca.
3	las vacas	pipeline_2	dormitaban al sol ya caliente , rumiando . Pero en el patio y Alfonso la llamó en silencio . Pasábanse horas sin oir el angustioso	1	Celia, mi tía mayor, que había concluído de dormir la siesta, cruzó el patio y Alfonso la llamó en silencio con la mano.
4	las vacas	pipeline_2	dormitaban al sol ya caliente , rumiando . Pero en el patio y Alfonso la llamó en silencio . Pasábanse horas sin oir el angustioso	2	Al bajar el sol volvieron, pero Berta quiso saludar un momento a sus vecinas de enfrente.
5	las vacas	pipeline_2	dormitaban al sol ya caliente , rumiando . Pero en el patio y	3	Volvió a su cobertizo, y en el camino sintió un ligero cosquilleo en la espalda.

			Alfonso la llamó en silencio . Pasábanse horas sin oir el angustioso		
6	las vacas	pipeline_3	dormitaban al fin se había reforzado su corazón siempre en la pipa y el perro había retirado . A media hora en tierra , del	1	Como las fieras amaestradas, los perros conocen el menor indicio de borrachera en su amo.
7	las vacas	pipeline_3	dormitaban al fin se había reforzado su corazón siempre en la pipa y el perro había retirado . A media hora en tierra , del	2	Por lo demás, se alternaban con su hija para ir a ver a la enferma.
8	las vacas	pipeline_3	dormitaban al fin se había reforzado su corazón siempre en la pipa y el perro había retirado . A media hora en tierra , del	3	Benincasa se observaba muy de cerca en los pies la placa lívida de la mordedura.
9	las vacas	pipeline_4	dormitaban al sol ya caliente , rumiando . Pero cuando los pobres caballos pasaron por el camino , ellas abrieron los ojos despreciativas : --	1	Pero cuando los pobres caballos pasaron por el camino, ellas abrieron los ojos despreciativas: --Son los caballos.
10	las vacas	pipeline_4	dormitaban al sol ya caliente , rumiando . Pero cuando los pobres caballos pasaron por el camino , ellas abrieron los ojos despreciativas : --	2	Tarde ya, cuando el sol acababa de entrarse, los dos caballos se acordaron del maíz y emprendieron el regreso.
11	las vacas	pipeline_4	dormitaban al sol ya caliente , rumiando . Pero cuando los pobres caballos pasaron por el camino , ellas abrieron los ojos despreciativas : --	3	Luis María, por su parte, se permite pasarle la mano por la barbilla cuando entra y ella está sentada de espaldas.
12	las vacas	pipeline_5	estaban inmóviles , mirando fijamente el verde paraíso inalcanzable . -- ¿por qué ? ¿qué le pasa ? -- nada , sino que está bien	1	Las vacas estaban inmóviles, mirando fijamente el verde paraíso inalcanzable.
13	las vacas	pipeline_5	estaban inmóviles , mirando fijamente el verde paraíso inalcanzable . -- ¿por qué ? ¿qué le pasa ? --	2	Ayestarain tornó a mirarme fijamente, pero esta vez creí notar un vago, vaguísimo dejo de amargura.

			nada , sino que está bien		
14	las vacas	pipeline_5	estaban inmóviles , mirando fijamente el verde paraíso inalcanzable . -- ¿por qué ? ¿qué le pasa ? -- nada , sino que está bien	3	Esta vez Vezzera me miró fijamente a los ojos: --¿Por qué no quieres ir?
15	las vacas	pipeline_6	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con	1	Pero cuando está conmigo, entonces no aparta los ojos de ellos.
16	las vacas	pipeline_6	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con	2	Arrizabalaga y la señora se reían, volviéndose a menudo, y la joven no apartaba casi sus ojos de Nébel.
17	las vacas	pipeline_6	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con	3	El caballo, por mayor intimidad de trato, es sensiblemente más afecto al hombre que la vaca.
18	las vacas	pipeline_7	dormitaban al sol ya caliente rumiando Pero cuando los pobres caballos pasaron por el camino ellas abrieron los ojos despreciativas Son los caballos Los alambres	1	Pero cuando los pobres caballos pasaron por el camino, ellas abrieron los ojos despreciativas: --Son los caballos.
19	las vacas	pipeline_7	dormitaban al sol ya caliente rumiando Pero cuando los pobres caballos pasaron por el camino ellas abrieron los ojos despreciativas Son los caballos Los alambres	2	Porque, naturalmente, cuanto más intensos eran los raptos de amor a su marido e hija, más irritable era su humor con los monstruos.
20	las vacas	pipeline_7	dormitaban al sol ya caliente rumiando Pero cuando los pobres caballos pasaron por el camino ellas abrieron los ojos despreciativas Son los caballos Los alambres	3	Caminando, comiendo, curioseando, el alazán y el malacara cruzaron la capuera hasta que un alambrado los detuvo.

Tabla 4

idx	prompt	model_name	output_text	idx	neighborh
0	las vacas	pipeline_1	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con	1	Detrás de él, las vacas dormitaban al sol ya caliente, rumiando.
1	las vacas	pipeline_1	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con	2	Pero cuando está conmigo, entonces no aparta los ojos de ellos.
2	las vacas	pipeline_1	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con	3	Y juro que fueron fuertes las dos horas que pasamos mi mujer y yo, con la luz prendida hasta que amaneció, ella acostada, yo sentado en la cama, vigilando sin cesar la arpillera flotante.
3	las vacas	pipeline_2	dormitaban al sol ya caliente , rumiando . Pero en el patio y Alfonso la llamó en silencio . Pasábanse horas sin oir el angustioso	1	Celia, mi tía mayor, que había concluído de dormir la siesta, cruzó el patio y Alfonso la llamó en silencio con la mano.
4	las vacas	pipeline_2	dormitaban al sol ya caliente , rumiando . Pero en el patio y Alfonso la llamó en silencio . Pasábanse horas sin oir el angustioso	2	Detrás de él, las vacas dormitaban al sol ya caliente, rumiando.
5	las vacas	pipeline_2	dormitaban al sol ya caliente , rumiando . Pero en el patio y Alfonso la llamó en silencio . Pasábanse horas sin oir el angustioso	3	Pasábanse horas sin oir el menor ruido.
6	las vacas	pipeline_3	dormitaban al fin se había reforzado su corazón siempre en la pipa y el perro había retirado . A media hora en tierra , del	1	Los peones que por a o b llegaban a la siesta, admiraron siempre la obstinación del perro, resoplando en cuevitas bajo un sol de fuego, si bien la admiración de aquellos no pasaba del cuadro de caza.
7	las vacas	pipeline_3	dormitaban al fin se había reforzado su corazón siempre en la pipa y el perro había retirado . A media hora en tierra , del	2	Los perros, entonces, sintieron más el próximo cambio de dueño, y solos, al pie de la casa dormida, comenzaron a llorar.
8	las vacas	pipeline_3	dormitaban al fin se había reforzado su corazón siempre en la pipa y el perro había retirado . A media hora en tierra , del	3	Cinco cigarrillos dejaron su tabaco adentro; y sentándonos entonces con las rodillas altas, encendí la pipa y aspiré.
9	las vacas	pipeline_4	dormitaban al sol ya caliente , rumiando . Pero cuando los pobres caballos pasaron por el camino , ellas	1	Pero cuando los pobres caballos pasaron por el camino, ellas abrieron los ojos despreciativas: --Son los caballos.

			abrieron los ojos despreciativas : --		
10	las vacas	pipeline_4	dormitaban al sol ya caliente , rumiando . Pero cuando los pobres caballos pasaron por el camino , ellas abrieron los ojos despreciativas : --	2	Detrás de él, las vacas dormitaban al sol ya caliente, rumiando.
11	las vacas	pipeline_4	dormitaban al sol ya caliente , rumiando . Pero cuando los pobres caballos pasaron por el camino , ellas abrieron los ojos despreciativas : --	3	El viento, muy frío, cristalizaba aún más la claridad de la mañana de oro, y los caballos, que sentían de frente el sol, casi horizontal todavía, entrecerraban los ojos al dichoso deslumbramiento.
12	las vacas	pipeline_5	estaban inmóviles , mirando fijamente el verde paraíso inalcanzable . -- ¿por qué ? ¿qué le pasa ? -- nada , sino que está bien	1	Las vacas estaban inmóviles, mirando fijamente el verde paraíso inalcanzable.
13	las vacas	pipeline_5	estaban inmóviles , mirando fijamente el verde paraíso inalcanzable . -- ¿por qué ? ¿qué le pasa ? -- nada , sino que está bien	2	Se miraron fijamente, insistentemente, aislados del mundo en aquella recta paralela de alma a alma que los mantenía inmóviles.
14	las vacas	pipeline_5	estaban inmóviles , mirando fijamente el verde paraíso inalcanzable . -- ¿por qué ? ¿qué le pasa ? -- nada , sino que está bien	3	Quedó inmóvil, toda ojos.
15	las vacas	pipeline_6	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con	1	Detrás de él, las vacas dormitaban al sol ya caliente, rumiando.
16	las vacas	pipeline_6	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con	2	Pero cuando está conmigo, entonces no aparta los ojos de ellos.
17	las vacas	pipeline_6	dormitaban al sol ya caliente , rumiando . Pero cuando está conmigo , entonces no aparta los ojos de mi mujer y yo , con	3	Y juro que fueron fuertes las dos horas que pasamos mi mujer y yo, con la luz prendida hasta que amaneció, ella acostada, yo sentado en la cama, vigilando sin cesar la arpillera flotante.
18	las vacas	pipeline_7	dormitaban al sol ya caliente rumiando Pero cuando los pobres caballos pasaron por el camino ellas abrieron los ojos despreciativas Son los caballos Los alambres	1	Pero cuando los pobres caballos pasaron por el camino, ellas abrieron los ojos despreciativas: --Son los caballos.
19	las vacas	pipeline_7	dormitaban al sol ya caliente rumiando Pero cuando los pobres caballos pasaron por el camino ellas abrieron los	2	Los dos caballos, vueltos ya a su pacífica condición de animales a que un solo hilo contiene, se sintieron ingenuamente

			ojos despreciativas Son los caballos Los alambres		deslumbrados por aquel héroe capaz de afrontar el alambre de púa, la cosa más terrible que puede hallar el deseo de pasar adelante.
20	las vacas	pipeline_7	dormitaban al sol ya caliente rumiando Pero cuando los pobres caballos pasaron por el camino ellas abrieron los ojos despreciativas Son los caballos Los alambres	3	Después de trasponer la loma, los caballos vieron de pronto a las vacas detenidas en el camino, y el recuerdo de la tarde anterior excitó sus orejas y su paso: querían ver cómo era el nuevo alambrado.

Tabla 5

Parámetros por modelo, técnica de vectorización y/o tipo de preprocesamiento (estos se combinan para realizar experimentos).

Representación	**Parámetros**	**Comentarios**
BoW	ngram_range=(2,4), strip_accents= 'unicode'	Como se observa en la Parte 1 de este documento, se espera que esta representación no sea muy buena capturando la semántica de las oraciones y las palabras en su contexto. Sin embargo puede ser un buen modelo de referencia como baseline.
TF-IDF	use_idf=True	TF-IDF es un mecanismo de vectorización complementario a BoW que permite ponderar mejor palabras poco frecuentes en el texto. Si bien no resuelve el problema de capturar contextos, en algunos escenarios como clasificación de texto, permite obtener resultados muy interesantes.
Sentence Transformers	model=intfloat/multilingual-e5-large	Como se observa en la Parte 1, los embeddings de este modelo capturan de buena forma la semántica de las oraciones para los fragmentos de texto de Quiroga analizados. Es natural entonces pensar, que estos vectores podrían ayudar a obtener buenos resultados con un clasificador de texto.
Pre-procesamiento	**Parámetros**	
SMOTE	sampling_strategy='auto', k_neighbors=5	Con SMOTE se pretende resamplear el dataset de entrenamiento, permitiendo balancear los ejemplos de entrenamiento por clase y así entrenar modelos más balanceados. En particular, se espera que esta técnica permita obtener mejores resultados para los textos de la clase Martin Fierro, en donde se aprecia para el modelo

		baseline resultados muy malos, por ser la clase menos sub-representada.
Tokenizer	Se usa PunktSentenceTokenizer	Tokenizador por defecto utilizado por NLTK, como parte del pre-procesamiento de sentencias provisto por el equipo docente.
Modelo	**Parámetros**	
SVM		Modelo sencillo pero efectivo, rápido de entrenar, que en combinación tanto con técnicas de vectorización simples como BoW / TF-IDF, como técnicas más complejas como sentence-transformers, nos permiten obtener una buena idea de la performance que se puede alcanzar en este problema, sin entrar en el uso de técnicas complejas y avanzadas.
MLP Classifier	ngram_range=(2,4), hidden_layer_sizes=(50,), ctivation='relu', solver='adam', alpha=0.0001, learning_rate_init=0.001, max_iter=50, momentum=0.9, nesterovs_momentum=True, validation_fraction=0.1	Algoritmo de clasificación de scikit-learn para redes neuronales (Multi-Layer Perceptron). Interesa probar el potencial de este modelo vs. SVM, sin entrar en un trabajo profundo de optimización de la arquitectura del modelo como la cantidad de hidden_layers, funciones de activación, entre otros.

Classification Report 1

A continuación se muestran los resultados del classification report en Test para el mejor modelo.

```
F-Score macro: 86.6
              precision    recall  f1-score   support

      quiroga       0.86      0.93      0.89       530
      becquer       0.88      0.77      0.82       370
martin_fierro       0.86      0.90      0.88        71

     accuracy                           0.87       971
    macro avg       0.87      0.87      0.87       971
 weighted avg       0.87      0.87      0.87       971
```