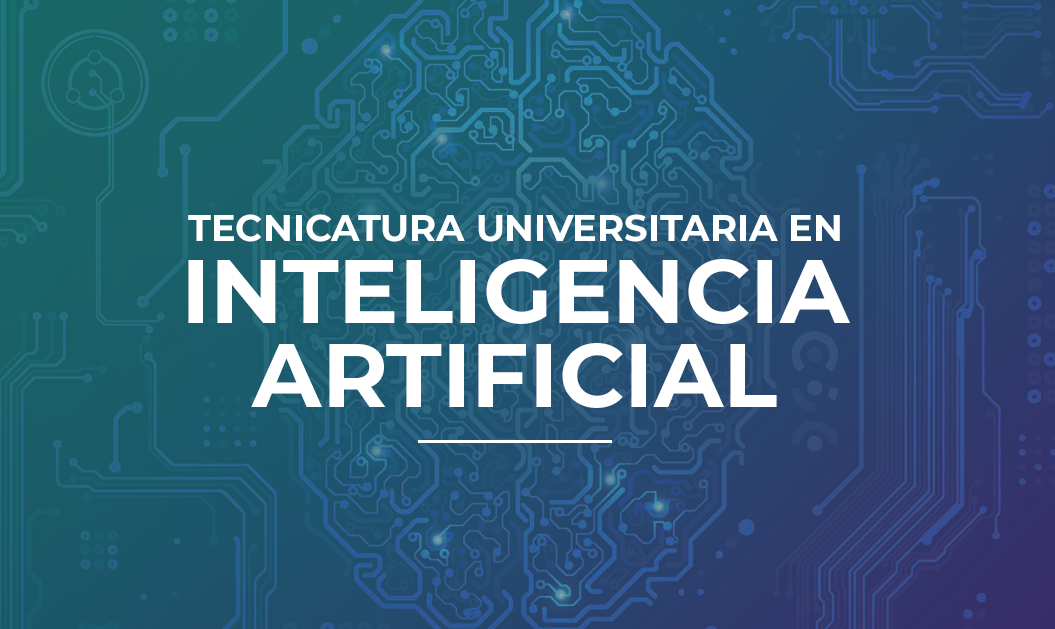
| F:\OBRAS HIDRÁULICAS\UniversidadNacionaldelRosario_Argentina_logo.jpg | F:\OBRAS HIDRÁULICAS\images.png |
| --- | --- |

**UNIVERSIDAD NACIONAL DE ROSARIO**

**Facultad de Ciencias Exactas, Ingeniería y Agrimensura**



**TECNICATURA UNIVERSITARIA EN INTELIGENCIA ARTIFICIAL**

**Asignatura:**

**NLP**

**TRABAJO PRÁCTICO 1- PARTE 2**

**TEMA: Pradera**

**PROFESORES:**

jpmanson@gmail.com

alan.geary.b@gmail.com

constantinoferrucci@gmail.com

dolores.sollberger@gmail.com

**INTEGRANTES:**

Santiago Giaveno

**FECHA:** 21.05.25

**ÍNDICE:**

**TEMA 2**

**1 | INTRODUCCIÓN 2**

**2 | DESARROLLO Y RESULTADOS 3**

**2.1 Segmentación de texto**

**2.2 Vectorización**

**2.3 Técnicas de distancias**

**2.4 Detección de idiomas 4**

**2.5 Análisis de sentimientos**

**2.6 Métodos de clasificación**

**3 | CONCLUSIONES 5**

**Introducción**

El objetivo es utilizar el repositorio de Pradera que contiene información, estadísticas y relaciones sobre el juego para aplicar diferentes herramientas aprendidas como lo son la segmentación, vectorización, similitud de texto, extracción de sustantivos y categorización de las mismas, detección de idiomas, análisis de sentimientos con modelos pre-entrenados y entrenamiento de un modelo de clasificación, con el fin de tener un modelo en donde pueda predecir la categoría a través de la consulta ingresada.

**DESARROLLO Y RESULTADOS**

**Segmentación de texto**

Probe con varios métodos de segmentación y fui comparando cada uno de ellos:

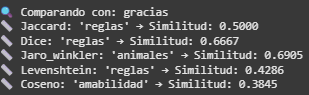
* Spacy: En cuanto a tiempo computacional era rápido, en cuanto a segmentación me dejaba muchos fragmentos con pocas palabras o una palabra o solamente letras, lo que luego dificultaba a la hora de hacer similitud de texto y demás.
* Stanza: Dividía bien en fragmentos, no dejando palabras separadas y letras solas. Pero el problema es que el tiempo computacional era bastante, con un texto extenso iba a tardar mucho.
* Pysbd: Finalmente este es el método que me queda para segmentar el texto(ej1). Es un balance entre los dos anteriores métodos, dejaba fragmentos con menos palabras pero eran menos que Spacy, por lo tanto aplicaba un filtro y el tiempo computacional no era tanto como Stanza

**Vectorización**

* Doc2Vec: Dependiendo del contexto y ejercicio, fui usando diferentes modelos para vectorizar. Para el ejercicio 1 que tenía que hacer análisis de similitud de frases con los fragmentos, necesitaba el contexto, entonces me hizo mejor trabajo Doc2Vec, aunque su tiempo computacional es más alto.
* TF-IDF: Aunque no captura contexto, es el más rápido y simple. Lo uso para hacer diferentes comparativas de modelos de clasificación en el ejercicio 6
* Spacy: Para el ejercicio 3, uso Spacy, que en sí es más rápido que Doc2vec, es más básico, ya que usa modelos pre entrenados y no se entrena sobre el corpus. Por lo tanto en el caso que tenia que comparar solamentes palabras me pareció que Spacy era más rápido y entendía lo básico del contexto
* Bert: Para el ejercicio 5, ya que había que hacer análisis de sentimiento y el modelo tenía que entender perfectamente el contexto, use Bert que es el que mejor comprensión semántica tiene. Tiene un tiempo computacional lento pero vale la pena para estos casos. También lo use para crear el modelo de clasificación de categorías en el ejercicio 6.

**Técnicas de distancias**

Fui comparando todas las técnicas aprendidas en cada ejercicio que lo requería. Técnicas como Jaccard, Dice, Jaro-Winkler, Levenshtein me daban un score bajo cuando se trataba de comparar frases o fragmentos, y uno alto cuando se trataba de palabras, tiene sentido ya que estas solo evalúan similitud de forma ortográfica (letras comunes o transformaciones). Pero esto no nos servía ya que queríamos tener contexto de las frases o palabras para poder compararlas, la similitud coseno fue la mejor para cada caso ya que capta relaciones conceptuales y no solo textuales.



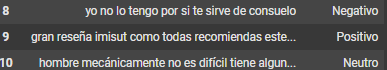
Comparación de ‘gracias’ con diferentes sustantivos que encontró cada técnica más parecida.

**Detección de idiomas**

Para detectar el idioma de cada archivo decidí usar langdetect, ya que es una herramienta simple y rápida para este tipo de problemas en donde están bien determinado de qué idioma se trata en cada archivo y no hay una variedad dentro de ellos. Decidí excluir los archivos que contenían enlaces ya que no se podía determinar un idioma a un url, al menos que se entre a cada uno de ellos y se verifique.

**Analisis de sentimientos**

Para hacer un análisis de los sentimientos de las reseñas dadas por las personas, decidí usar un modelo de clasificación de bert pre entrenado multilingual. En donde me devolvía números(1 al 5) y un score de confianza(entre 0 y 1) para cada comentario. Luego filtrando por 1-2 comentarios negativos, 3 neutros y 4-5 positivos obtuve lo siguiente:



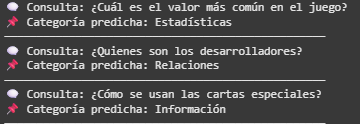
**Modelo de clasificación**

Para entrenar un modelo de clasificación en donde pueda predecir la categoría a través de la consulta ingresada lo primero que hice fue crear un csv con más de 300 consultas cada una con su categorización(información-relación-estadística) con ayuda de un chatbot.

Luego fui creando diferentes modelos de clasificación simples, optimizando sus parámetros con GridSearch para tener un modelo óptimo y vectorizando con TD-IDF para una mayor rapidez y simpleza. Probé con LogisticRegression,BayesIngenuos y RandomForest. Lo que me dio buenas métricas, para finalmente comparar con un modelo más complejo para ver si valía la pena un mayor costo y tiempo computacional. Lo que hice fue vectorizar con Bert y luego usar RandomForest, lo que me terminó mejorando bastante el modelo, superando a los modelos bases.



Por último comprobé que las predicciones sean correctas con el último modelo mencionado que es con el que me quedo.



**CONCLUSIÓN**

Como conclusión de esta parte del trabajo práctico, puedo decir que no siempre un modelo o herramienta más complejo va a ser el mejor en todos los casos. Siempre hay que ir probando y comparando para justificar el mayor uso de costo computacional, lo mejor es encontrar un balance entre precisión y rendimiento. Como también cada herramienta es útil para cada caso, no siempre una técnica nos va a servir para cada ejercicio, hay que ver el contexto y objetivos. Hay que saber cuándo invertir en modelos más sofisticados.

Vinculo Proyecto: [Tp1\_P2\_individualNLP.ipynb](https://colab.research.google.com/drive/1e6kPTeuAB_nvn6S6QWbXgXy2EgUrsjjO?usp=sharing)