LoreNexus: Descifrando la procedencia de nombres en mundos virtuales, donde confluyen universos fantásticos, ficticios, históricos y culturales.

Proponente: Aingeru García Blas

1. Descripción

En entornos de videojuegos online, los usuarios suelen crear nombres personalizados para sus personajes inspirados en universos ficticios (e.g: `Star Wars`, `El Señor de los Anillos...`), videojuegos (e.g: `World of Warcraft`, `Final Fantasy`), así como en nombres propios de culturas y etnias globales, incluyendo nombres de origen **árabe, asiático, sudamericano, nórdico, entre otros**. Además, algunas personas optan por nombres históricos o figuras controvertidas. Estos nombres modificados reflejan patrones de derivación propios de cada universo o cultura.

La tarea consistiría en desarrollar un sistema de clasificación de nombres derivados que sea capaz de identificar el origen o inspiración (universo, cultura etc.) de un nombre introducido, incluso si este ha sido alterado o adaptado como es común en videojuegos y/o mundos virtuales en internet. La idea es entrenar el sistema para detectar patrones, prefijos, sufijos, y combinaciones de carácteres característicos de cada contexto o universo, lo cual (espero) permitirá clasificar nombres derivados aún cuando presentan variaciones. Para ello, se recolectarán datos de diferentes fuentes para ser fusionados en un dataset que contendrá nombres etiquetados de diferentes universos ficticios, culturas globales, contextos históricos etc.

No obstante, el uso de nombres derivados en mundos virtuales tiene inherente el hecho de que algunos jugadores tienden a utilizar nombres con connotaciones ofensivas (e.g: racistas, homófobas etc.) con lo que también se considerará en el proceso mediante datos oportunos. Así como un posible tratamiento para nombres compuestos, que también son muy comunes.

2. Objetivos

Z1: Preparación y "curación" del dataset

Se recopilarán datos de múltiples fuentes, como datasets generales de nombres o en algunos casos, APIs más especializadas/temáticas (e.g, una API de Star Wars). Cada conjunto de datos será etiquetado según su universo de origen (por ejemplo, `Star Wars`, `Harry Potter`, etc.).

Oier sugirió también el uso de **NER**, lo que considero muy buena idea y utilizaré para mapear información extra como nombres que no existan ya, localizaciones etc, con objeto de complementar el dataset base, con libros en formato pdf como fuente.

Z2: Modelado y exploración de diferentes enfoques.

- **Tokenización a nivel de carácteres**: A sugerencia de Oier también, probaré una tokenización basada en carácteres, lo cual debería de permitir al modelo captar patrones específicos en los nombres (como sufijos `-ian`, `-mir`, `-el`) que reflejan las variaciones lingüísticas de cada universo y además, variaciones de los nombres, lo cual creo que me será de gran utilidad para capturar la esencia en las modificaciones que puedan realizar los usuarios, (e.g: `Araghornn` en lugar de `Aragorn`).

- **Exploración de modelos**: Se evaluarán modelos de PLN preentrenados y técnicas de **RNN-LSTM**. Tras investigar un poco, he visto modelos **transformers** como **CharacterBERT** [1] [2] [3] que podría ajustar al problema mediante fine-tuning para capturar patrones a nivel de caracteres. También consideraré **RoBERTa** [4], ya que dispone de una funcionalidad que considero podría resultar útil: "*Byte-level BPE vocabulary: Uses BPE with bytes as a subunit instead of characters, accommodating Unicode characters.*"

- **Data augmentation**: Con objeto de mejorar el modelo y "corregir" posibles desbalances entre la distribución de los datos (algunos universos tendrán muchísimas más instancias que otros), implementaré técnicas de aumento de datos, como duplicación/adición/substracción de letras, cambios fonéticos menores y la inclusión de prefijos y sufijos que puedan ser representativos de cada universo/cultura, para intentar ayudar al modelo a generalizar mejor sobre nombres con variaciones.

- Busco ser capaz de responder: ¿Cuáles son las fortalezas y debilidades de cada modelo al clasificar nombres con variaciones derivadas de cada universo o cultura?

Z3:

Aplicación a gran escala: El modelo optimizado se aplicará a una base de datos de alrededor de 1 millón de nombres (tras filtrado) de un MMORPG, recogidos a lo largo de 12 años (proyecto propio) y con usuarios de todo el mundo. Esta base de datos **no formará parte de la validación o entrenamiento**, sino que se utilizará para analizar patrones y categorías de nombres en la exposición final del proyecto. Estos nombres, sin usarse en el entrenamiento o validación, se utilizarán para análisis estadístico.

Análisis demográfico y cultural: También sugerido por Oier, a partir de los metadatos (por ejemplo, IPs de los usuarios), se generarán estadísticas para observar tendencias demográficas sobre los universos ficticios y categorías de nombres más populares por región.

- **Me gustaría poder responder:** ¿Qué observaciones se pueden hacer en cuanto a las tendencias sobre los universos ficticios y nombres culturales más populares por jugadores de todo el mundo y cómo varían en función de la región de origen?

3. Material

- Dataset de nombres etiquetados por universo o cultura de origen, históricos, insultos, palabras racistas…etc
- Modelos preentrenados y ajustados (**CharacterBERT**, **RoBERTa**) y librerías como **Transformers** de Huggin Face.
- Herramientas de tokenización y utilidades varias (como **spaCy**).

Utilizaré Google Colab para aprovechar el uso del entorno y runtime basado en GPUs, pero en general, dispondré de tanto el notebook como los datasets o scripts/utilidades que desarrolle para pre-procesamiento, daga augmentation etc en un repositorio alojado en mi cuenta de Github: geru-scotland, al cual daré acceso al tutor o tutora que se me asigne, como se indica en las especificaciones del proyecto.

4. Referencias

[1] https://github.com/helboukkouri/character-bert
[2] https://huggingface.co/helboukkouri/character-bert
[3] https://arxiv.org/abs/2010.10392
[4] https://huggingface.co/docs/transformers/model_doc/roberta