Procesamiento del Lenguaje Natural

Julián **Arce** - 60509

Sebastián **Itokazu** - 60392

Gian Luca **Pecile** - 59235

Valentín **Ratti** - 60031

Índice

Resúmen

En el contexto de la reciente huelga de guionistas SAG-AFTRA[1], se ha desencadenado un conflicto entre guionistas, actores y estudios de cine debido, entre una de las razones, el uso generalizado de la inteligencia artificial en la creación de guiones.
Uno de los principales puntos de negociación es la regulación ante el abuso de la IA para reemplazar a escritores y actores, así como la apropiación no consentida de obras preexistentes con el propósito de generar nuevo contenido.
Este trabajo tiene como objetivo abordar las implicaciones de esta problemática y proponer una posible solución: la detección de autoría en fragmentos de guiones. Se investigará si es posible mediante herramientas de procesamiento de lenguaje natural identificar la esencia de un guionista en un fragmento de guión como una forma de abordar las preocupaciones de plagio y apropiación indebida de obras.
Para lograr esto se trabajará con un corpus de obras de guionistas seleccionados por su estilo particular que puede ser identificado en las películas que adaptan los guiones.

Corpus

El corpus seleccionado es, en su composición genética proveniente de una selección de 3000 guiones encontrado en kaggle[2]. Este a su vez está separado en tres secciones que incluyen: movie_characters; movie_metadata y screenplay_data. El principal enfoque del análisis se basa en el último. A su vez, en el mismo se tienen los guiones raw o en plano como como .txt, con anotaciones manuales, con un proceso de lemmatization aplicado, con anotaciones para ser procesado por BERT y por último en formato .json separado por anotaciones por regla.
Por la cantidad de guiones y datos que, en base al trabajo que se desea realizar, son innecesarios se optó por recortar manualmente el dataset por uno más conciso y coherente con el análisis que se desea hacer del estilo de escritura de ciertos guionistas. En particular se seleccionaron guiones escritos por: Richard Linklater, Paul Thomas Anderson, Woody Allen, Charlie Kaufman y Todd Solondz. Estos fueron curados en particular porque se conoce previamente de sus películas y distintivo estilo de escritura el cual se puede notar entre cinéfilos y se cree que va a enriquecer el análisis. Esto se podría generalizar para más guionistas para así cumplir nuestro objetivo principal que es discernir si un determinado guión podría haber sido creado con IA.

Análisis

Exploración

A continuación se muestra la representación de cloud of words de todos los diálogos presentes en los guiones de los guionistas seleccionados en el siguiente orden: Aaron

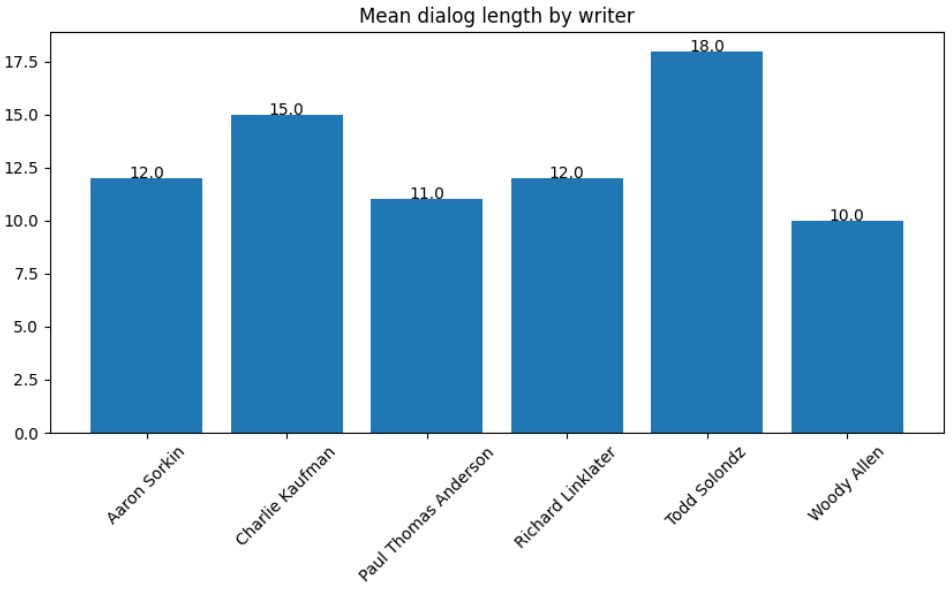
[1] Writers strike: Why A.I. is such a hot-button issue in Hollywood's labor battle with SAG-AFTRA, Fortune
[2] dataset, última vez revisado: 01/10/23.

Sorkin, Charlie Kaufman, Paul Thomas Anderson, Richard Linklater, Todd Solondz y Woody Allen:

De aquí se puede ver el enfoque que le dan ciertos autores a la mención de personajes como es el caso de Charlie Kaufman y Richard Linklater por sobre el uso de palabras descriptivas como es el caso de Woody Allen.

No se hace stemming de los datos ingresados para el cloud of words ya que el largo de los guiones se considera que no lo amerita para mejorar su performance.

Luego se decidió analizar el promedio de largo de diálogo por guionista y los resultados son los siguientes:



Esto se hizo para sacar una tendencia sobre los escritores más verborrágicos en sus diálogos.

Propuesta

Con el fin de detectar la autoría de un guión utilizando técnicas de Procesamiento de Lenguaje Natural (NLP), se considerarán las siguientes técnicas de análisis.

- **Modelo de Frecuencia de Palabras**: Cálculo la frecuencia de palabras y frases clave en los guiones de diferentes autores. Se puede utilizar TF-IDF (Term Frequency-Inverse Document Frequency) para dar más peso a las palabras y frases distintivas.
- **BERT (Bidirectional Encoder Representations from Transformers)**: Fine-tune de un modelo BERT ya pre-entrenado en guiones de autores conocidos donde luego puede ser utilizado con el fin de clasificar guiones desconocidos y definir su autoría.

En principio se planea realizar la prueba para este corpus planteado en su sección correspondiente y al verse exitoso esto, puede ser escalado agrandando la selección de guionistas al igual que sus obras.

Anexo

- Repositorio de github de trabajo: JuArce/nlp-tp (github.com).