Tecnicatura Universitaria en Inteligencia Artificial
Facultad de Ciencias Exactas, Ingeniería y Agrimensura
Universidad Nacional de Rosario

Procesamiento del Lenguaje Natural

Trabajo Práctico 1 - Parte 2

Alumno:
Pedro Casado

Resumen

Este informe presenta el desarrollo de los ejercicios correspondientes a la Parte 2 del Trabajo Práctico 1 de la materia Procesamiento de Lenguaje Natural.

Se aborda el uso de embeddings, técnicas de similitud semántica, POS, NER, análisis de sentimientos, categorización de preguntas, etc.

Introducción

El presente trabajo tiene como objetivo aplicar los conceptos y herramientas abordados en las primeras tres unidades de la materia TUIA - Procesamiento de Lenguaje Natural, a través de un conjunto de ejercicios prácticos. Para ello, se trabaja con información textual extraída de diversas páginas web públicas relacionadas con el juego de mesa Sagrada. Este informe constituye la segunda parte de un trabajo más amplio, cuya primera etapa consistió en la recolección de datos mediante técnicas de web scraping.

En esta instancia, se abordan tareas fundamentales del procesamiento de lenguaje natural (NLP), como la vectorización de texto, la comparación semántica entre fragmentos, el análisis de entidades y sentimientos, la detección de idioma y la categorización automática de preguntas. Cada uno de estos desafíos se resuelve utilizando datos reales, lo cual permite poner a prueba tanto la comprensión conceptual como la capacidad de implementación técnica.

A lo largo del informe se describen los distintos ejercicios desarrollados, las decisiones tomadas durante la implementación, los problemas encontrados y las soluciones adoptadas. Finalmente, se presenta una reflexión general sobre los aprendizajes obtenidos durante la realización del trabajo.

Metodología

Fuente de los datos

La información utilizada en este trabajo proviene de una etapa previa de recolección mediante técnicas de *web scraping*, aplicadas a distintos sitios web públicos relacionados con el juego de mesa **Sagrada**. Se recopilaron reseñas de usuarios, transcripciones de partidas (playthroughs), análisis y opiniones de terceros, reglamentos, y contenido de sitios como Wikipedia y BoardGameGeek. Los textos extraídos fueron almacenados, preprocesados y utilizados como insumo para las distintas tareas de análisis planteadas en esta segunda etapa del trabajo.

Descripción de los métodos y técnicas utilizadas

A lo largo del trabajo se implementaron diversas técnicas de procesamiento de lenguaje natural (PLN), agrupadas según los ejercicios solicitados:

- **Segmentación de texto:**
  Se utilizaron herramientas como `stanza` y `langchain` para segmentar el contenido en fragmentos manejables. Se aplicaron diferentes estrategias según la estructura del texto (e.g., segmentación semántica o basada en caracteres).

- **Vectorización y representación semántica:**
  Se emplearon modelos preentrenados de *Sentence Transformers* (SBERT) para convertir los fragmentos textuales en vectores numéricos (*embeddings*) de alta calidad.

- **Cálculo de similitud semántica:**
  Se utilizaron métricas como la **distancia coseno** y la **distancia de Levenshtein** para medir la similitud entre fragmentos de texto y consultas específicas.

- **Etiquetado gramatical y extracción de entidades:**
  A través de `spaCy`, se aplicaron técnicas de *Part-of-Speech Tagging* (POS) y *Named Entity Recognition* (NER) para identificar sustantivos y

entidades nombradas relevantes dentro de los fragmentos.

- **Detección de idioma:**
  Se utilizó la librería `langdetect` para identificar el idioma predominante de cada fragmento textual, especialmente útil dada la diversidad lingüística del corpus.

- **Análisis de sentimientos:**
  Se aplicaron modelos preentrenados basados en *Transformers* (como `distilBERT`) para analizar el sentimiento (positivo o negativo) de reseñas en distintos idiomas.

- **Clasificación de preguntas:**
  Se construyó un conjunto de 500 preguntas clasificadas en tres categorías: **estadísticas**, **información** y **relaciones**. Las preguntas se vectorizaron y luego se entrenó un modelo secuencial con `TensorFlow`, compuesto por tres capas densas y una salida *softmax*, evaluado mediante *accuracy*.

Herramientas y tecnologías empleadas

- **Software y entorno:**
  Visual Studio Code, Python 3.11

- **Librerías y frameworks:**
  `pandas`, `numpy`, `sklearn`, `spaCy`, `langdetect`, `matplotlib`, `tensorflow`, `transformers`, `keras`, `sentence-transformers`

- **Repositorio del proyecto:**
  Todo el código fuente y los datos procesados están disponibles en el siguiente repositorio de GitHub:
  https://github.com/CasadoPedro/TP1_NLP_P2

Desarrollo e implementación

Ejercicio 2:

- Se extrajo una reseña en español y se realizó limpieza del texto.
- Se segmentó usando `stanza`, debido a su soporte para el idioma español y su capacidad para manejar textos no estructurados.
- Se obtuvieron *embeddings* utilizando un modelo de `sentence_transformers`.
- Se ingresaron consultas (*queries*) y se calculó la similitud semántica con los fragmentos.
- Se utilizó **distancia coseno** para identificar las frases más relevantes.

Ejercicio 3:

- Se combinaron dos transcripciones de reseñas en inglés en un único texto extenso.
- Se segmentó el texto utilizando `RecursiveCharacterTextSplitter`, ya que el texto transcripto no contenía puntuación.
- Se realizó etiquetado gramatical (POS) y extracción de entidades (NER) con `spaCy`.
- Se filtraron fragmentos por sustantivos y se usó **distancia de Levenshtein** para buscar coincidencias con la *query*.
- Se compararon distintas métricas para seleccionar la más adecuada.

Ejercicio 4:

- Se recorrieron todos los archivos del corpus textual.
- Se identificó el idioma de cada archivo con `langdetect`.

- Se almacenaron los resultados en un `DataFrame`.

Ejercicio 5:

- Se analizaron reseñas obtenidas desde BoardGameGeek.
- Se aplicó un modelo multilingüe de análisis de sentimientos.
- Se construyó un sistema de búsqueda semántica con filtrado por sentimiento.

Ejercicio 6:

- Se generó un conjunto de 500 preguntas clasificadas manualmente.
- Las preguntas fueron etiquetadas como *estadísticas*, *información* o *relaciones*.
- Se vectorizaron utilizando el modelo SBERT con mejor rendimiento en español.
- Se entrenó un modelo secuencial de red neuronal con `TensorFlow`, utilizando 3 capas densas y una capa de salida `softmax`.
- El rendimiento se evaluó en un conjunto de validación mediante la métrica de *accuracy*.

Conclusiones

Durante la realización de este trabajo práctico surgieron diversas dificultades, principalmente relacionadas con la calidad y el formato del texto disponible. La variación de lenguaje dentro de un mismo documento, la falta de estructura, puntuación inconsistente y contenido irrelevante afectaron el desempeño de varias técnicas de NLP. Esto resalta la importancia crítica de una buena recolección y curación de datos durante la etapa de *web scraping*.

Además, se encontraron problemas técnicos al implementar algunos modelos preentrenados, debido a incompatibilidades de dependencias o limitaciones del entorno de ejecución. A pesar de ello, las herramientas actuales ofrecen soluciones muy efectivas para múltiples tareas del procesamiento de texto, permitiendo obtener resultados aceptables con esfuerzos razonables.

En resumen, este trabajo permitió aplicar de forma práctica muchos de los conceptos teóricos abordados en clase, evidenciando tanto el poder como las limitaciones del NLP cuando se trabaja con datos reales.