Universidad Nacional de Rosario

Facultad de Ciencias Exactas, Ingeniería y Agrimensura

Tecnicatura Universitaria en Inteligencia Artificial

Procesamiento del Lenguaje Natural

Trabajo Práctico Nº1

- **Leguiza, Claudia**
- **Cancio, José**
- **Rodríguez y Barros, Francisco**
- **Masciangelo, Lucia**
- **Texier, Julieta**

Resumen

Este proyecto recopiló información diversa sobre el juego de mesa **Pradera** a través de múltiples fuentes. El objetivo fue construir un conjunto de datos completo que incluyera opiniones de jugadores, detalles técnicos y conexiones relevantes entre distintos aspectos del juego, trabajando con contenido en varios idiomas.

Para la recolección de datos se utilizaron:

- Técnicas automatizadas para extraer texto.
- Procesamiento de documentos oficiales y guías.
- Transcripción de contenido audiovisual.

Introducción

Objetivo

Este trabajo tiene como objetivo extraer, procesar y analizar datos multilingües relacionados con **Pradera**. Los datos obtenidos abarcan información textual (reviews, tutoriales, manuales), estadísticas del juego y relaciones entre entidades (desarrolladores, personajes, sagas), con el fin de generar un conjunto de datos estructurados.

Metodología

Para cumplir con los requerimientos, se utilizaron las siguientes técnicas:

- Extracción de texto:
  - **Documentos:** Procesamiento de archivos PDF.
  - **Videos:** Transcripción automática de reviews y tutoriales mediante herramientas como ffmpeg , pydub y speech Recognition
  - **Foros:** Web scraping de hilos de discusión (ej: Reddit, foros oficiales) con BeautifulSoup o Selenium.
- Organización de datos:
  - Almacenamiento en formatos tabulares (CSV, Excel) para estadísticas y relaciones.
  - Las imágenes se almacenaron en formato jpg.
  - El resto de la información fue almacenada en formato txt.

Idiomas cubiertos

Los datos fueron recolectados en al menos dos idiomas:

1. **Español:** Principalmente de foros latinoamericanos y manuales locales.
2. **Inglés:** Extraídos de fuentes internacionales como Steam, Reddit o videos de YouTube.

Datos obtenidos

➢ Información (Archivos de texto):

Los archivos de texto se encuentran en la carpeta informacion.

En esta carpeta se encuentra tanto información general del juego como información relacionada al reglamento en español e inglés, mecánicas del juego, opiniones de foros, comentarios, tutoriales, etc.

Introducción:

- Descripción: El contenido es una hoja de ruta inicial para quienes buscan una descripción concisa del juego y sus especificaciones básicas. Es una presentación general del juego, destacando su temática y características principales.
- Archivo: introduccion.txt
- Fuente: https://misutmeeple.com/2021/05/resena-pradera/
- Proceso de extracción: El archivo fue obtenido mediante webscraping usando la librería BeautifulSoup.
- Archivo código fuente: Pradera_Misutmeeple.ipynb.

Mecánica del juego:

- Descripción: Contiene información sobre las mecánicas principales del juego como selección de carta, construcción de tableau cumpliendo requisitos de símbolos, y competencia por objetivos comunes. Explica cómo avanzan las partidas con reinicios estratégicos y acciones limitadas, también incluye variantes para juego en solitario.
- Archivo: Mecanica_del_juego.txt
- Fuente: https://misutmeeple.com/2021/05/resena-pradera/
- Proceso de extracción: Los dos archivos fueron obtenidos mediante webscraping usando la librería BeautifulSoup.
- Archivo código fuente: Pradera_Misutmeeple.ipynb.

Descripción general del juego:

- Descripción: El archivo contiene la descripción oficial completa del juego Meadow. Incluye las mecánicas principales como el sistema de colección de cartas y colocación de fichas, explica el objetivo del juego y las condiciones para ganar.
- Archivo: descripcion_general.txt
- Fuente: https://boardgamegeek.com/boardgame/314491/meadow
- Proceso de extracción: Para obtener esta información se utilizó Selenium WebDriver que permitió navegar automáticamente hasta la página del juego. Se extrajo el contenido de todos los párrafos de la sección de descripción. Finalmente, el texto completo se guardó en el archivo .txt.
- Archivo código fuente: Web_Scraping_dinamico_Pradera.ipynb.

Instrucciones (Español e Inglés):

- Descripción: Ambos archivos contienen un texto completo en cada idioma respectivamente con las reglas del juego, incluyendo preparación, desarrollo del juego y condiciones de victoria.

- Archivos:
  - reglamento_español.txt
  - reglamento_ingles.txt
- Fuentes: Las instrucciones del juego en el idioma español fueron extraídas del archivo “Pradera_Reglamento_(Spanish).pdf”, mientras que en el idioma inglés se extrajo del archivo “Meadow_EN_rulebook.pdf”.
- Archivo código fuente: Web_Scraping_dinamico_Pradera.ipynb.

Foros:

- Descripción: los foros de la sección “Forums” del sitio están agrupados por temática (“general”, “reviews”, “rules”, “variants”, etc). Decidimos separar los foros con mayor cantidad de opiniones en 4 archivos agrupándolos según la temática.
- Fuente: Información extraída de los foros del juego Pradera de la página BoardGameGeek | Gaming Unplugged Since 2000.
- Archivos:
  - foro_variantes.txt: Recopila variantes caseras para el juego Pradera e ideas para combinar expansiones. También propone formas de acortar partidas en 4 jugadores.
  - foro_reviews.txt: Contiene comentarios y reseñas sobre el juego destacando sus mecánicas y experiencia general. Incluye descripciones de cómo se juega, la construcción de tableros, las acciones disponibles y la estrategia, junto con opiniones personales. También menciona comparaciones con otros juegos.
  - foro_reglas.txt: Los participantes comparten interpretaciones y experiencias de juego. Contiene discusiones sobre reglas y mecánicas del juego, incluyendo preguntas y debates sobre aspectos específicos como la restricción de colocar cartas sobre ciertos animales, el orden de turnos en partidas de 3-4 jugadores, el funcionamiento de los objetivos con cartas especiales, etc.
  - foro_general.txt: El archivo contiene discusiones generales sobre el juego como consultas sobre expansiones faltantes, compatibilidad de contenido en una sola caja, límites de cartas y fichas, etc. y consejos para enseñar el juego a nuevos jugadores. También hay preguntas sobre la ubicación temática del juego. Los participantes comparten recursos útiles, como listas de cartas en Excel, y reflexiones personales sobre la experiencia de juego.
- Proceso de Extracción: Se usó web scraping dinámico mediante Selenium WebDriver para acceder y extraer el contenido, esto fue necesario por el carácter dinámico de las páginas de BGG.
- Archivo código fuente: Web_Scraping_dinamico_Pradera.ipynb.

Comentarios:

- Descripción: El archivo contiene una colección de 51 comentarios de la comunidad de jugadores donde expresan sus impresiones sobre el juego. Incluyen análisis, comparaciones con otros juegos, preguntas sobre reglas específicas, y discusiones sobre estrategias. Contiene también opiniones divididas de los jugadores.
- Archivo: comentarios.txt
- Fuente: Los comentarios fueron extraídos de la sección de respuestas del artículo "Reseña - Pradera" de la página https://misutmeeple.com (Reseña: Pradera | Misut Meeple)
- Proceso de extracción: A diferencia de las otras extracciones que requieren Selenium por tener contenido dinámico, en este caso utilizamos BeautifulSoup porque los comentarios se cargaban directamente en el HTML estático de la página. Se identificaron y capturaron los elementos específicos de la página que contenían los comentarios de los usuarios.
- Archivo código fuente: Pradera_Misutmeeple.ipynb.

Enlaces (Imágenes y enlaces temáticos):

- Descripción: Estos archivos contienen enlaces extraídos de la reseña del juego Pradera publicada en el blog https://misutmeeple.com. Contienen tanto material visual como referencias temáticas que contextualizan el juego dentro del ecosistema de juegos de mesa.
- Archivos:
  - enlaces_imagenes.txt
  - enlaces_pagina.txt
- Fuentes:
  - https://misutmeeple.com/
  - **Imágenes:** Se obtuvieron de la sección principal del artículo, donde se muestran fotografías de los componentes del juego, ilustraciones y capturas de partidas.
  - **Enlaces temáticos:** Se recopilaron de la sección de etiquetas al final de la publicación, que incluye categorías relacionadas con mecánicas de juego, diseñadores y géneros.
- Proceso de Extracción: Los dos archivos fueron obtenidos mediante webscraping usando la librería BeautifulSoup.
- Archivo código fuente: Pradera_Misutmeeple.ipynb.

Tutoriales (Español e Inglés):

- Descripción: Ambos archivos contienen un texto completo en cada idioma respectivamente con tutoriales del juego, incluyendo preparación, desarrollo del juego y condiciones de victoria.
- Archivos:
  - Pradera_Tutorial_Spanish_video.txt
  - Pradera_Tutorial_English_video.txt

- Fuentes: el tutorial del juego en idioma inglés fue extraído del archivo de video “Meadow-how to play- board game rules”, mientras que el de idioma español se extrajo del archivo “Pradera (Meadow) - Comentarios y Cómo Jugar”, ambos de la sección “Videos” de la página del juego: https://boardgamegeek.com/boardgame/314491/meadow/videos/all
- Proceso de Extracción: Los dos archivos fueron obtenidos extrayendo en primer lugar el audio de cada archivo de video (en formato .wav), luego fragmentándolos en segmentos más pequeños y finalmente extrayendo el texto de estos. Las librerías utilizadas fueron **ffmpeg, pydub y Speech_Recognition.**
- Archivo código fuente: Extraccion_texto_videosypdf.ipynb.

➢ Estadísticas:

Los datos estadísticos se guardaron en archivo tabulare (.csv) con métricas del juego. Se encuentran en la carpeta estadisticas.

- Archivos:
  - meadow_stats.csv
  - Fuente: https://boardgamegeek.com/boardgame/314491/meadow/stats
  - Descripción: Estadísticas scrapeadas de la ficha de Pradera en BGG usando la librería Selenium en Python, incluyendo ratings, rankings y datos de la comunidad.
  - Archivo código fuente: Web_Scraping_dinamico_Pradera.ipynb

```
            valor_estadistica = stats_group.find_element(By.CSS_SELECTOR, "div.outline-item-description").text
        stats[nom_estadistica] = valor_estadistica

# Crear un DataFrame de pandas
stats_df = pd.DataFrame((stats.items()), columns=['Estadistica', 'Valor'])
print(stats_df)

driver.quit()
```

```
       Estadistica     Valor
0      Avg. Rating     7.719
1   No. of Ratings    12,201
2   Std. Deviation      1.22
3           Weight  2.25 / 5
4         Comments     1,861
5             Fans     1,164
6       Page Views   983,188
7     Overall Rank       209
8    Strategy Rank       158
9      Family Rank        28
10  All Time Plays    50,530
11      This Month       106
12             Own    23,140
13     Prev. Owned     1,770
14       For Trade       234
15   Want In Trade       509
16        Wishlist     4,854
17       Has Parts         6
18      Want Parts         5
```

Estructura archivo meadow_stats.csv

➢ Relaciones:

Las relaciones se guardaron en archivos tabulares (.csv). Se encuentran en la carpeta relaciones.

- Descripción: Este apartado organiza la información de los créditos del juego *Pradera* en una red de relaciones semánticas.
- Archivo: El resultado fue almacenado como un archivo tabular relaciones_juego.csv (imagen 2) y visualizado mediante un grafo (imagen 3)
- Proceso de extracción: A partir del archivo creditos.csv, creado durante la extracción de los créditos del juego, con los datos principales (nombre del juego, diseñador, ilustrador, editorial, mecanismos, etc.) (imagen1), se construyeron tuplas de la forma (Sujeto1 - Relación - Sujeto2) que representan conexiones como "*Pradera* - Diseñador - Klemens Kalicki". Los datos fueron normalizados, transformando listas en valores individuales y agrupando relaciones internas (como entre varios ilustradores o editores). Se utilizaron las librerías **networkx** y **matplotlib** para la visualización.
- Archivo código fuente: Web_Scraping_dinamico_Pradera.ipynb

```
# La estructura de los creditos del juego es la siguiente:
# clave                  | valor
# -----------------------|-----------------------------------------------------
# nombre_primario        | Meadow
# diseñador              | Klemens Kalicki
# ilustrador             | Karolina Kijak, Katarzyna Fiebiger
# editorial              | Rebel Studio, ADC Blackfire Entertainment, ...
# desarrollador          | Klemens Kalicki
# escritor               | Agnieszka Możejko
# mecanismo              | Hand Management, Layering, Open Drafting, ...
```

imagen 1: Estructura archivo creditos.csv

```
# Guardar las relaciones en un archivo .CSV
# Nombre del archivo
nombre_archivo = 'relaciones_juego.csv'

# Acceso a la carpeta compartida en MyDrive
directorio_guardado = f'/content/drive/My Drive/PRADERA/datos/relaciones'
os.makedirs(directorio_guardado, exist_ok=True)

# Ruta completa del archivo
ruta_completa = os.path.join(directorio_guardado, nombre_archivo)

# Guardar el DataFrame directamente en esa ruta
df_relaciones.to_csv(ruta_completa, index=False, header=True)

print(f"\nRelaciones del juego guardadas en {ruta_completa}")
```

```
  SUJETO1            RELACION                    SUJETO2
0  Meadow  NOMBRE_ALTERNATIVO                       Łąka
1  Meadow  NOMBRE_ALTERNATIVO                     Livada
2  Meadow  NOMBRE_ALTERNATIVO  Meadow Im Reich der Natur
3  Meadow  NOMBRE_ALTERNATIVO                  Na louce
4  Meadow  NOMBRE_ALTERNATIVO                    Pradera

Relaciones del juego guardadas en /content/drive/My Drive/PRADERA/datos/relaciones/relaciones_juego.csv
```

imagen 2: Estructura archivo relaciones.csv

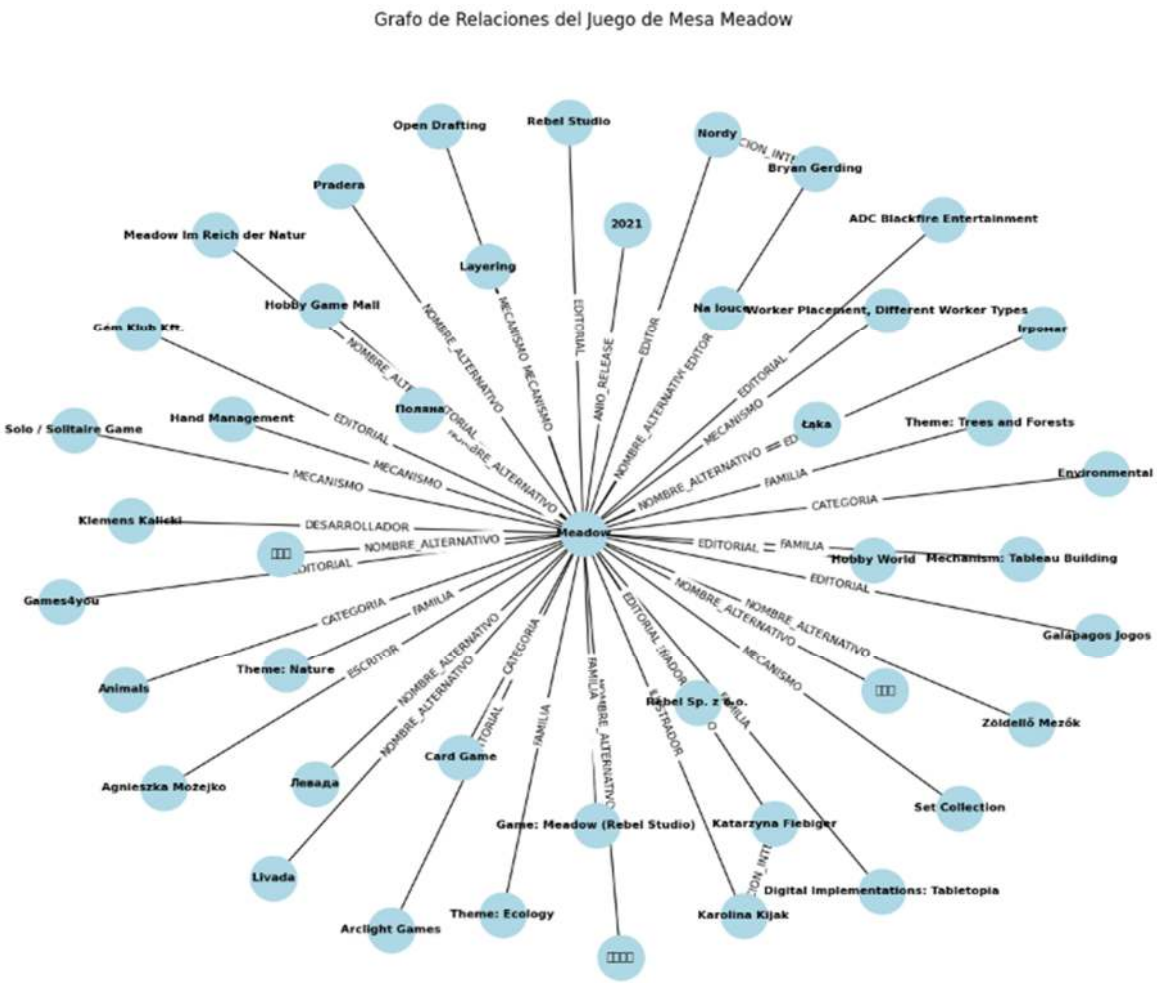

imagen 3: Grafo de relaciones del juego Pradera

➢ Análisis post-extracción

- Descripción: Este apartado resume estadísticamente la cantidad de información recolectada en los archivos producto de la extracción, referidos al juego Pradera, como ser: cantidad de archivos de texto generados, cantidad de caracteres y palabras acumuladas, máximos y mínimos en un archivo, promedio, etc. Acompañados de gráficas correspondientes.
- Archivo código fuente: Web_Scraping_dinamico_Pradera.ipynb

```
Resumen de los archivos de texto:
Cantidad total de archivos: 15
Total de caracteres acumulados: 609279
Promedio de caracteres por archivo: 40618.60
Máximo de caracteres en un archivo: 202239
Mínimo de caracteres en un archivo: 462

Total de palabras acumuladas: 109757
Promedio de palabras por archivo: 7317.13
Máximo de palabras en un archivo: 35991
Mínimo de palabras en un archivo: 18
```

Conclusión:

El proyecto logró reunir, organizar y estructurar una amplia variedad de datos sobre el juego de mesa *Pradera*, combinando fuentes multilingües y múltiples formatos (textuales, audiovisuales, gráficos, estadísticos y relacionales). A través del uso de técnicas de extracción automatizada, transcripción y análisis de contenido, se consolidó un conjunto de datos integral que permite comprender en profundidad tanto las mecánicas del juego como la percepción de los jugadores en distintas comunidades.
En particular, estos datos servirán como insumo clave para el desarrollo de un chatbot especializado, capaz de responder de forma contextualizada y precisa a consultas sobre las mecánicas, reglamento, estrategias y otros aspectos del juego *Pradera*.