CS1000 Introducción a las Ciencias de la Computación
Proyecto 1 (P1)
Pregrado
2024-1
Profesor: Ian Paul Brossard
Lab 1.02

Proyecto 1 (P1)

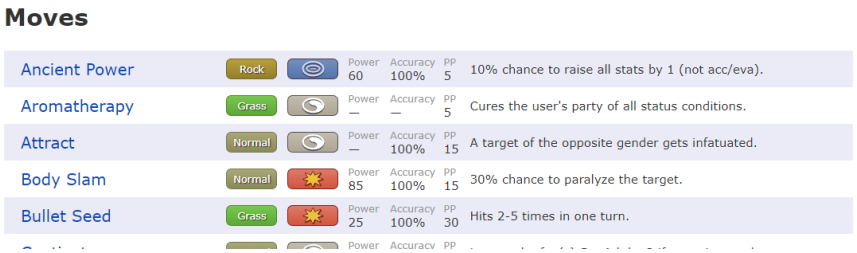
El presente trabajo cuenta con 2 preguntas. Para ambas preguntas, usted deberá trabajar con la colección de datos listada a continuación:

Dataset smogon.csv

- La colección de datos será proporcionada por el profesor. La encontrarán en el campus virtual bajo el nombre smogon.csv
- Suba este archivo .csv a Google Drive y ábralo con Google Sheets.
- Esta colección de datos corresponde a la información disponible en la página https://www.smogon.com/dex/xy/pokemon/ cada fila representa a un Pokémon.

  La columna 'Pokemon' contiene el nombre del pokémon que se está analizando, la columna 'url' contiene la página web de donde se extrajo la información textual. La columna 'texto' contiene la descripción del pokémon y la columna 'moves' contiene información acerca de todos los ataques que aprender dicho pokémon (aquí encontrará tanto el nombre del ataque como la descripción del ataque).

  Por ejemplo, en la fila 111, se encuentra el pokémon Chikorita, cuya información fue extraída del siguiente https://www.smogon.com/dex/xy/pokemon/chikorita/. Usted puede consultar dicha página web y notar que la lista de ataques empieza en la última sección, titulada «Moves», tal y como se muestra en la siguiente imagen:

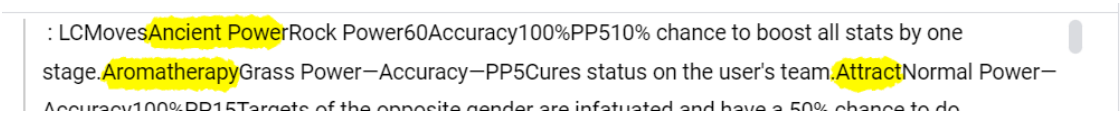


Notará que el primer movimiento que aprende Chikorita se llama Ancient Power, el segundo es Aromatheraphy, el tercero es Attract, y así sucesivamente.

Cada uno de estos ataques tiene características escritas a su costado en formato textual, por ejemplo, Ancient Power tiene la palabra «Rock», luego tiene la

palabra «Power» y el número «60», luego tiene la palabra Accuracy, y muchas otras palabras.

Toda esta información se encuentra en el archivo separado por comas, aunque se una manera desordenada y confusa.



El texto que se encuentra en esta celda no respeta las reglas sintácticas del Español, sin embargo, la computadora puede entender el conocimiento que está expresado en dicho texto.

Pregunta 1:

Para esta pregunta, usted deberá agrupar los datos usando TFIDF y PCA, y realizar una exposición sobre estos dos métodos y explicando las diferencias que encontró.

1.1. Agrupamiento mediante TF-IDF (7 ptos).

- Generar la matriz tf-idf utilizando una cantidad de n-gramas elegida por usted (unigramas, bigramas, trigramas, etc.).
- Mostrar el número total de tokens (elementos de su vocabulario) que tiene su matriz tf-idf.
- Imprimir todos los tokens (elementos de su vocabulario)
- Generar un DataFrame con la matriz tf-idf que tenga como cabeceras los elementos de su vocabulario. Imprimir dicha matriz usando la instrucción print().
- Agrupar las filas de su nuevo DataFrame, en base a sus puntuaciones tf-idf (usted puede elegir cuántos clusters desea utilizar).
- Generar un archivo de valores separado por comas (CSV) que contenga su matriz tfidf y el cluster.
- Interpretar los cluster y ponerle un nombre a cada uno. Si no puede ponerle un nombre, escriba un párrafo explicando las características principales de dicho cluster. Es posible que solo algunos clusters tengan interpretación y otros no. Por ejemplo, si usted tiene 18 clusters y solo 10 de ellos pueden ser interpretados, explique el detalle de esos 10. Si no es posible interpretar ninguno de sus clusters, entonces repita el proceso usando otro número de n_clusters.

1.2. Agrupamiento mediante PCA (7 ptos)

- Cargue en un DataFrame el archivo CSV generado en la Parte 1 del trabajo.
- Descarte la columna del cluster. Para esto, deberá referenciar el nombre de dicha columna. Imprima el DataFrame usando el comando print().

- Descarte la primera columna con el objetivo de eliminar el doble índice. Para esto, deberá referenciar el atributo columns[0] de su DataFrame. Imprima el DataFrame usando el comando print().
- Aplique Análisis de Componentes principales utilizando la cantidad de componentes que usted desee.
- Imprima el número de filas y columnas de su DataFrame original.
- Imprima el número de filas y columnas de su matriz de componentes principales.
- Generar un DataFrame nuevo con la matriz de componentes principales que tenga como cabeceras el número de componente (por ejemplo, PCA1, PCA2, PCA3, etc.). Imprimir dicha matriz usando la instrucción print().
- Agrupar las filas de su nuevo DataFrame, en base a sus puntuaciones PCA (usted puede elegir cuántos clusters desea utilizar).
- Generar un archivo de valores separado por comas (CSV) que contenga su matriz PCA y el cluster.
- Interpretar los cluster y ponerle un nombre a cada uno. Si no puede ponerle un nombre, escriba un párrafo explicando las características principales de dicho cluster. Es posible que solo algunos clusters tengan interpretación y otros no. Por ejemplo, si usted tiene 18 clusters y solo 10 de ellos pueden ser interpretados, explique el detalle de esos 10. Si no es posible interpretar nignuno de sus clusters, entonces repita el proceso usando otro número de n_clusters.

Pregunta 2 (6 ptos):

Usted implementará método para reprocesar la columna «moves» que le permita conseguir un agrupamiento que se ajuste a este dominio específico (es decir, el procesamiento que hará usted, solo servirá para agrupar los personajes de esta colección de datos, pero no tendrá utilidad para agrupar otras colecciones de datos).

Este procesamiento consiste en lo siguiente:

- Dado que el agrupamiento de personajes de esta franquicia es por tipos, usted deberá eliminar todas las palabras que no correspondan un tipo de pokémon.

- Para esto, usted deberá investigar cómo procesar toda la columna «moves» de su dataframe o de su csv, de modo que se eliminen todas las **palabras** o **partes de palabras** que no corresponden a la lista de tipos identificada por usted.

    - **Palabras:** Por ejemplo, si los únicos tipos existentes son “fire” y “water” y usted tiene la cadena “unleashes a scorching stream of water that sizzles with fire” quedaría como “water fire”.

    - **Partes de palabras:** En cambio, para el caso de palabras que contienen un tipo, se mantiene solo el tipo. Por ejemplo, la cadena “Blazing HydroburstWater projects a concentrated burst of superheated water that flashes into steam, enveloping targets in a scalding fog” quedaría como “water water”. Y la cadena “utecfirepython” quedaría como “fire”.

- Una vez que haya hecho este procesamiento, procese su nueva columna «moves» con tfidf y agrupe a los personajes. Para esta pregunta debe usar únicamente unigramas.

- Muestre los tokens.
- Indique cuántos tokens hay en total.
- Interprete los clusters obtenidos.

Entregables:

Su trabajo deberá contener los siguientes elementos:

- Un zip con el proyecto de Python (debe incluir los archivos .py y .csv usados).
- Un informe en Word en el que describa los pasos realizados, y se evidencie que está entregando todos los puntos solicitados en los párrafos anteriores. Al final de su documento debe escribir 4 conclusiones de su trabajo. En la carátula de este informe debe incluir el nombre y código de los integrantes de su grupo.
- Una presentación oral.
- Adicional: si usted usó un programa en python u otro lenguaje para preprocesar los textos antes de introducirlos en su proyecto, deberá adjuntar información probatoria de las tareas realizadas (puede ser el código fuente, un documento con capturas de pantalla y la explicación de los pasos, o un video).

Rúbrica:

Criterio	EXCELENTE	ADECUADO	MÍNIMO	INSUFICIENTE
Desarrollo de software	Diseña y elabora el software para lograr una solución adecuada al problema planteado. El software debe ser ordenado, claro y óptimo. (10 p.)	Diseña y elabora el software para lograr una solución adecuada al problema planteado. El software es solo funcionable. (6 p.)	Diseña el software para lograr una solución adecuada al problema planteado. El software no se concluye adecuadamente. (4 p.)	No logra el diseño ni la implementación correcta del software. (2 p.)
Presentación escrita	El informe contiene las secciones de Antecedentes, Fundamento Teórico, Métodos y Desarrollo y Conclusiones. Estas últimas, adecuadamente formuladas. (5 p.)	El informe contiene las secciones de Antecedentes, Fundamento Teórico, Métodos y Desarrollo, pero no pone énfasis en las conclusiones. (3 p.)	El informe contiene menos de la mitad de las secciones estipuladas, incluyendo conclusiones. (2 p.)	El informe contiene menos de la mitad de las secciones estipuladas, sin incluir conclusiones. (0 p.)
Presentación oral	El alumno presenta el proyecto en forma adecuada y responde a las preguntas del profesor en forma lógica y coherente. (5 p.)	El alumno presenta el proyecto en forma adecuada, pero no responde a todas las preguntas del profesor en forma lógica y coherente. (3 p.)	El alumno no presenta el proyecto en forma adecuada, pero responde a las preguntas del profesor en forma lógica y coherente. (2 p.)	El alumno no presenta el proyecto en forma adecuada ni responde a las preguntas del profesor en forma lógica y coherente. O no se presenta a la presentación oral. (0 p.)