samediff

Escribe tu propio cancion

Usa SameDiff para analizar los estilos de las letras de dos músicos y compón un dueto para ellos

Resumen

SameDiff compara dos o más archivos de texto y te dice qué tan similares o diferentes son. Te ayuda a ver diferencias o similitudes en las palabras usadas en cada archivo para que puedas aprender sobre el análisis cuantitativo de textos. Este folleto de actividad ayuda a los participantes a familiarizarse con el manejo de datos, comparando las letras de dos músicos e inventando una nueva canción que hubieran compuesto juntos.

Metas de aprendizaje

- Habilidad aumentada para analizar datos de textos
- Entendimiento de que la comparación de dos cosas es una forma poderosa de encontrar historias en los datos
- Consciencia de qué tipo de preguntas puedes o deberías preguntarle a los datos de texto
- Entender que el análisis algorítmico puede revelar información interesante sobre tus datos

Realizar la actividad

Resolver un problema

Analizar textos largos a mano es difícil. Una forma de entender el "corpus" de un texto es compararlo con otro, o compararlo con partes de éste. Los científicos de la computación han creado formas de ayudar, inventando varias recetas, o "algoritmos", que pueden comparar dos corpus. SameDiff corre algunos de estos algoritmos para que puedas comparar dos piezas grandes de texto.

Comparte ejemplos motivadores

Los textos largos están en todo alrededor nuestro. El día de hoy puedes descargar todos los correos de la Secretaria de Estado de EU, Hillary Clinton, cables diplomáticos de Wikileaks, o todas las novelas de Sherlock Holmes del Proyecto Gutenberg. Analizar y visualizar estos textos largos es algo muy común, de formas serias o divertidas. Muestra las piezas "Firmas de color" de Jaz Parkinson que compara los colores mencionados en diferentes libros (*http://jazparkinson.tumblr.com*), y el "Laboratorio de Investigación del Rap" de Tahir Hemphill (*http://rapresearchlab.com*).

Tiempo total

30 minutos

Tamaño del grupo

3 - 100 personas. Edades 12+.

Diseñamos esto para cerca de 30 personas, pero se podría realizar con más o con menos. Está diseñado para grupos de educación media y media superior, organizaciones de noticias, organizaciones sin fines de lucro y talleres comunitarios. No se necesita experiencia previa en el manejo de datos.

Espacio

- Un proyector conectado a una computadora
- Capacidad para separarse en grupos pequeños de 3 reunidos alrededor de una computadora
- Mesas largas o piso adecuado para poner papel sobre él, o cinta para pegarlo a las paredes, para que los participantes puedan dibujar

Materiales

- Computadoras
  1 por cada 3 participantes
- Pliegos grandes de papel
  más o menos de medio metro por 1 metro
- Lápices de colores gruesos o plumones

Realizar la actividad (continuado)

Presenta la herramienta

Abre SameDiff (*https://databasic.io/samediff*) y elige a Beyoncé y Aretha Franklin de las muestras. En la página de resultados explica que la columna de la izquierda muestra palabras únicas de Beyoncé, mientras que la columna derecha muestra palabras únicas a Aretha Franklin. Esas son las diferencias. La columna de en medio muestra las palabras que tienen en común. Llama la atención a la parte superior de la pagina de resultados, donde dice "estos dos documentos son algo similares". SameDiff usa un algoritmo llamado "similitud coseno" para dar un puntaje de similitud. La similitud coseno funciona creando una lista de palabras de Beyoncé y una lista de palabras de Aretha Franklin. Cuenta qué tan seguido aparece cada término en cada documento y luego compara qué tanto concuerdan las listas. Este es un algoritmo útil para el análisis de textos.

Da comienzo a la actividad

1. Los participantes tienen 15 minutos.
2. Los participantes trabajan en equipos de tres.
3. Cada equipo usa SameDiff para comparar las letras de dos músicos. Puesto que las colaboraciones musicales son tan populares, elige dos artistas y luego imagina cómo sonaría una canción escrita por esos dos artistas.
   https://databasic.io/samediff
4. Cada equipo escribe las letras de su canción con lápices de colores en un pedazo grande de papel.
5. Los equipos obtienen puntos adicionales si: (a) sus canciones riman, (b) se les ocurre una tonada para cantarla, o (c) la cantan en estilo karaoke frente al grupo.

Intercambio de experiencias

Toma 1 minuto para cada grupo para compartir su nueva canción. Algunas preguntas y temas que buscar y concentrarse durante la discusión:

- ¿Notaste algunos temas comunes?
- ¿Las letras resultantes son más interesantes cuando vienen de artistas cuyos trabajos son muy diferentes?
- La comparación es una forma poderosa de encontrar historias en los datos.
- ¡Trabajar con datos puede ser divertido!

A recordar

- Los algoritmos son simplemente una serie de pasos que tú (o una computadora) haces en orden para resolver un problema. Por ejemplo, cuando pierdes tus llaves corres un algoritmo para buscarlas: primero revisas tus bolsillos, la vitrina cerca de la puerta, etc.
- Una similitud coseno de 1.0 significa que son exactamente lo mismo; cero significa que son totalmente diferentes.

Términos

Algoritmo
Una serie de pasos que tú (o una computadora) haces con el fin de resolver un problema. Por ejemplo, cuando escribes un término para buscarlo en Google, el buscador corre un algoritmo para tratar de descifrar qué páginas mostrarte.

Corpus
Un conjunto de textos escritos. Por ejemplo, todas las letras de las canciones de Katy Perry. Su plural es "corpus" también.

Similtud coseno
El puntaje de similaridad coseno trata de decir qué tan similares son dos documentos, basada en el número de veces que se usan las palabras en cada uno de ellos.

DataBasic es una paquete de herramientas de uso fácil para principiantes que los introduce a conceptos del trabajo con datos. Estas herramientas sencillas hacen que sea fácil trabajar con los datos de formas divertidas, para que puedas aprender cómo encontrar grandes historias que contar.