Write Your Own Song

Use SameDiff to analyze the lyrical styles of two musicians and invent a duet for them

What is SameDiff?

SameDiff compares two or more text files and tells you how similar or different they are. It helps you see differences and similarities in the words used in each file so you can learn about quantitative analysis of text. This hands-on activity helps participants build their data literacy by comparing lyrics from two musicians and inventing a new song they would write together.

Learning Goals

• Increased ability to analyze text data.
• Understanding that comparing two things is one powerful way to find stories in data.
• Awareness of what kind of questions you can/should ask text data.
• Understanding that algorithmic analysis can reveal interesting information about your data.

Run the Activity

Solving a Problem

Analyzing large text is hard to do by hand. One way to understand a "corpus" of text is to compare it to another one, or to compare to parts of it. Computer scientists have come up with ways to help, inventing various recipes, or "algorithms", that can compare two corpora. SameDiff runs some of those algorithms for you so you can try to compare two big pieces of text to each other.

Share Inspirational Examples

Large text data is all around us. Today you can download all of Hillary Clinton's Secretary of State emails, diplomatic cables from Wikileaks, or all the Sherlock Holmes novels from the Gutenberg Project. Analyzing and visualizing these large texts is a common thing to do now, in serious or fun ways. Show Jaz Parkinson's "Color Signatures" pieces that compares the colors mentioned in different books (*http://jazparkinson.tumblr.com*), and Tahir Hemphill's "Rap Research Lab" (*http://rapresearchlab.com*).

Total time

30 to 45 Minutes

Audience

3 - 100 people. Ages 12+.
Designed for grades 6 - 12, Higher Ed classrooms, News Organizations, Non-profits, and Community Workshops. No prior experience with data is required.

Space

• A projector and computer.
• Ability to break out into small groups of 3 clustered around a computer.
• Large tables or floor, or tape to stick paper to walls so participants can draw

Supplies

• Computers
  1 for every 3 participants
• Large pieces of paper
  roughly 2 feet x 3 feet
• Thick crayons or markers

Run the Activity (continued)

Introduce the Tool

Open up SameDiff (*https://databasic.io/samediff*) and choose Beyoncé and Aretha Franklin from the samples. On the results page explain that the left column shows words unique to Beyoncé, while the right column shows words unique to Aretha Franklin. Those are their differences. The middle column shows the words they have in common. Draw their attention to the top of the results page where it says, "These two documents are sort of similar". SameDiff uses an algorithm called "cosine similarity" to give you a similarity score. Cosine similarity works by creating a list of words from Beyonce and a list of words from Aretha Franklin. It counts how often each term appears in each document and then compares how closely those two lists match each other. This is a helpful algorithm for text analysis.

Launch the Activity

1. Participants have 15 minutes.

2. Participants work in teams of three.

3. Each team uses SameDiff to compare the lyrics of two musicians. Since musical collaborations are very popular, pick two artists and then imagine what a song written by the two artists would look like.
    https://databasic.io/samediff

4. Each team writes the lyrics of their song on a big piece of paper with crayons.

5. Teams get bonus points if: (a) their song rhymes and/or (b) they come up with a tune to sing it to and/or (c) They perform it karaoke-style for the group.

Share-Back

Take 1 minute for each group to share their new song. Some questions and themes to look for and focus on during the discussion:

• Did you notice any common themes?

• Are the resulting lyrics more interesting when they come from artists whose work is very different?

• Comparison is a powerful way to find stories in data.

• Working with data can be fun!

Reminders

• We run algorithms everyday. For example, when you lose your keys you run an algorithm to search for them - first you check your pockets, the counter by the door, etc.

• A cosine similarity of 1.0 means exactly the same; zero means totally different.

Terms to Introduce

Algorithm
A set of steps you (or a computer) do in order to solve a problem.

Corpus
A collection of written texts. For example, all of the lyrics in Katy Perry songs.

Cosine Similarity
The Cosine Similarity score tries to tell you how similar two documents are based on the number of times words are used in each.

Sample Sketches