DATABASIC

Investigate

Activity #2

SameDiff

Purpose

SameDiff contrasts two or more text files and allows participants to compare the content of each file. It helps participants see differences and similarities in the words used in each file so they can learn about quantitative analysis of text. This hands-on activity helps participants build their data literacy by comparing two different datasets and creating a story.

Learning Goals

- Increased ability to analyze text data
- Develop the ability to recognize data patterns and tell a story by comparing datasets
- Develop the ability to recognize the kinds of questions that can be asked of datasets
- Understanding that algorithmic analysis can reveal interesting information about your data

Time

30 to 45 Minutes

Supplies

- Ability to break out into groups of three participants clustered around a computer
- Large tables or floor, or tape to stick paper to walls so participants can draw
- Projector
- Computers *1 for every 3 participants*
- Large pieces of paper *roughly 2 feet x 3 feet*
- Thick crayons or markers

Introductory Activity Questions

- **HOW CAN DIFFERENT WORDS BE USED TO SPIN NARRATIVES?**

- **WHY IS IT IMPORTANT TO LOOK AT DIFFERENT WAYS TEXT IS PRESENTED?**
- **HOW CAN PARTICIPANTS FORM THEIR OWN ARGUMENTS FROM COMPARING TWO DATASETS?**

Instructions

Introduce the Tool

To demonstrate the tool to participants, open up SameDiff (https://databasic.io/samediff) and paste the links to two different takes on the same news story. On the results page, explain that the right and left columns show words unique to each article. Those columns represent their differences. The middle column shows the words they have in common. There also may be a similarity score at the top of the screen. SameDiff uses an algorithm called "cosine similarity" to produce this score. This function counts how often each term appears in each document, and then compares how closely the numbers match. 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 two different sides on the same piece of news. Then, they write their own version of the story. (https://databasic.io/samediff)
4. Each team writes their version of the story on a big piece of paper.

Debrief

Take one minute for each group to share their new story. Some questions and themes to look for and focus on during the discussion might include:

- Did you notice any common themes?
- Comparison is a powerful way to find stories in data.

- Working with data does not have to be intimidating. Doing something as simple as comparing the frequency of words in documents offers a starting point for analyzing media.