# Unit 8: Analysis

# Most frequent word analysis

Lesson 55

Derek Ruths

# Overview of unit

Objectives:

- Understand the philosophy of analysis
- Learn strategies for approaching data analysis
- Know major areas of data analysis and what questions they answer

1. What is analysis?
2. What is a result?
3. Most frequent word analysis
4. Statistical analysis

# Lesson overview

## Objectives

- See an example of choosing/designing a valid method
- Understand tf-idf

## Outline

- The problem from HW3
- Possible solutions
- tf-idf

# The problem from HW3

- Each pony's most frequently used words?

# Possible approaches?

# TF-IDF

## term frequency inverse document frequency

A term's overall score should be impacted by how common it is in general.

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

idf is the penalty term

t is term

d is a document

n is the number of documents

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

0 is as common as you can get

# Lesson wrap-up

## Takeaways

- Important to think through the validity of a method
- Often it involves picking the right method
- tf-idf is very useful

## Up next

- Statistical analysis