# TEXT WITH R


# data/ted_talks_2021/

# INTRODUCTION

## EXPLORE THE DATA SET

Q1: What columns and how many rows exists in the data set?

Q2: What is one row in the data set?

Q3: What is the timespan of talks in the data?

Q4: How many talks are there per year?

Q5: Which columns contain strings?

# GROUP WORK PART 1

# SEARCH AND TRANSFORM TEXT

Q1: How many times is "education" mentioned in the title of all talks?

Q2: What is the most common first name among all speakers?

Q3: What is the most often applied tag among all talks?

Q4: How many events were hosted in New York?

Q5: Which talk tagged "technology" had the most views?

Q6: What is the top ten of first words across all talks?

Q7: What is the distribution of the description length in words?

# GROUP WORK PART 2

# TOKENIZE TEXT

Q1: Apply the five steps to tokenize the talk transcripts into atomic words!

1. Filter the data to include talks from 2010 onwards

2. Apply transformations to clean and normalize the transcripts

3. Split the transcripts into words (or tokens)

4. Remove common english stop words

5. Add part of speech tags and keep only verbs in the final result

# GROUP WORK PART 3

# RULE-BASED TEXT CLASSIFICATION

Q1: Perform a deductive topic classification to identify talks about AI!

1. Create a theory-driven dictionary for the topic AI

2. Apply the dictionary to the tokenized transcripts

3. Decide on a metric and aggregate keyword matches

4. Assign a class to each talk based on the metric

5. Review the result and refine your dictionary