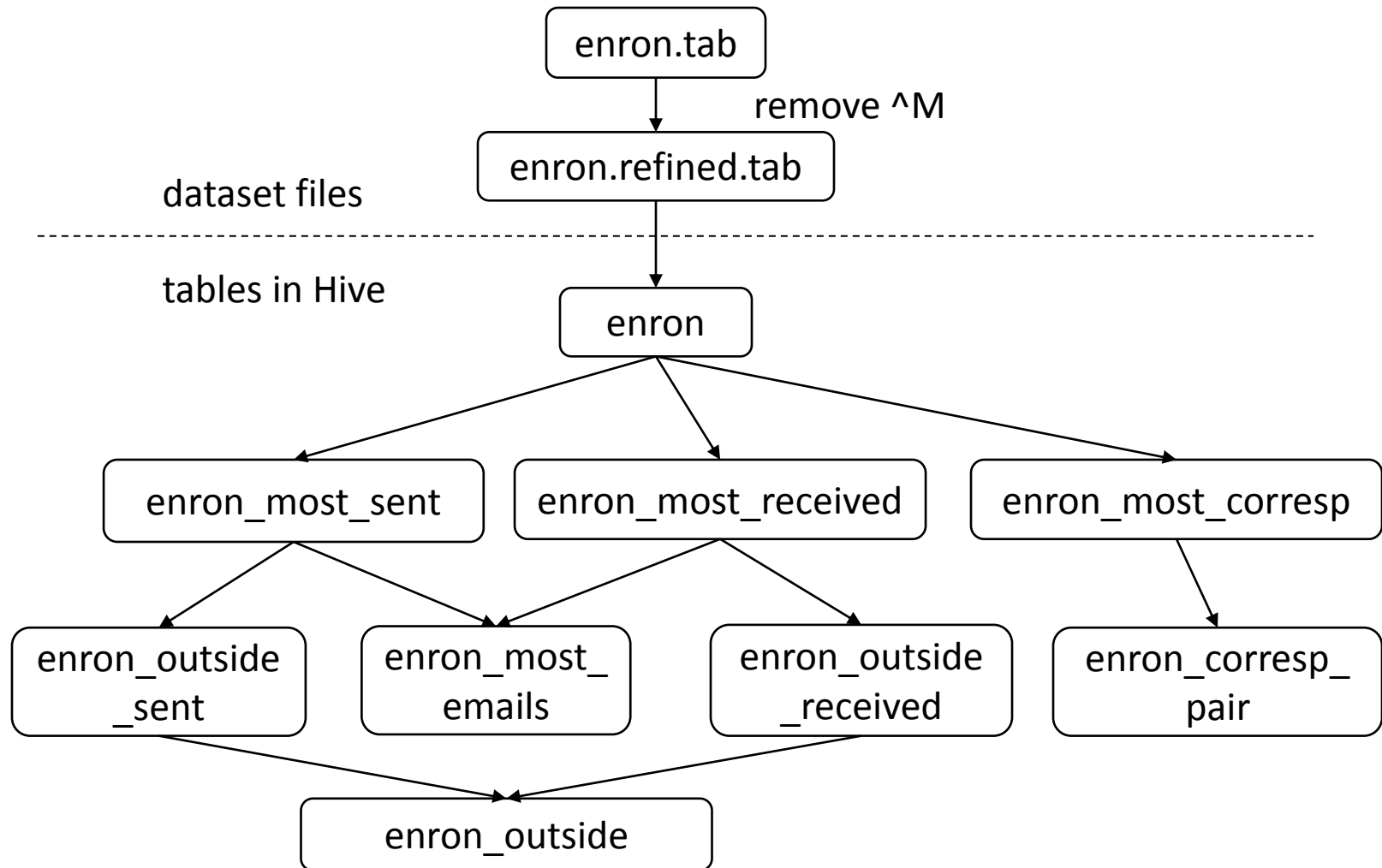


# CIS 6930 Project 2

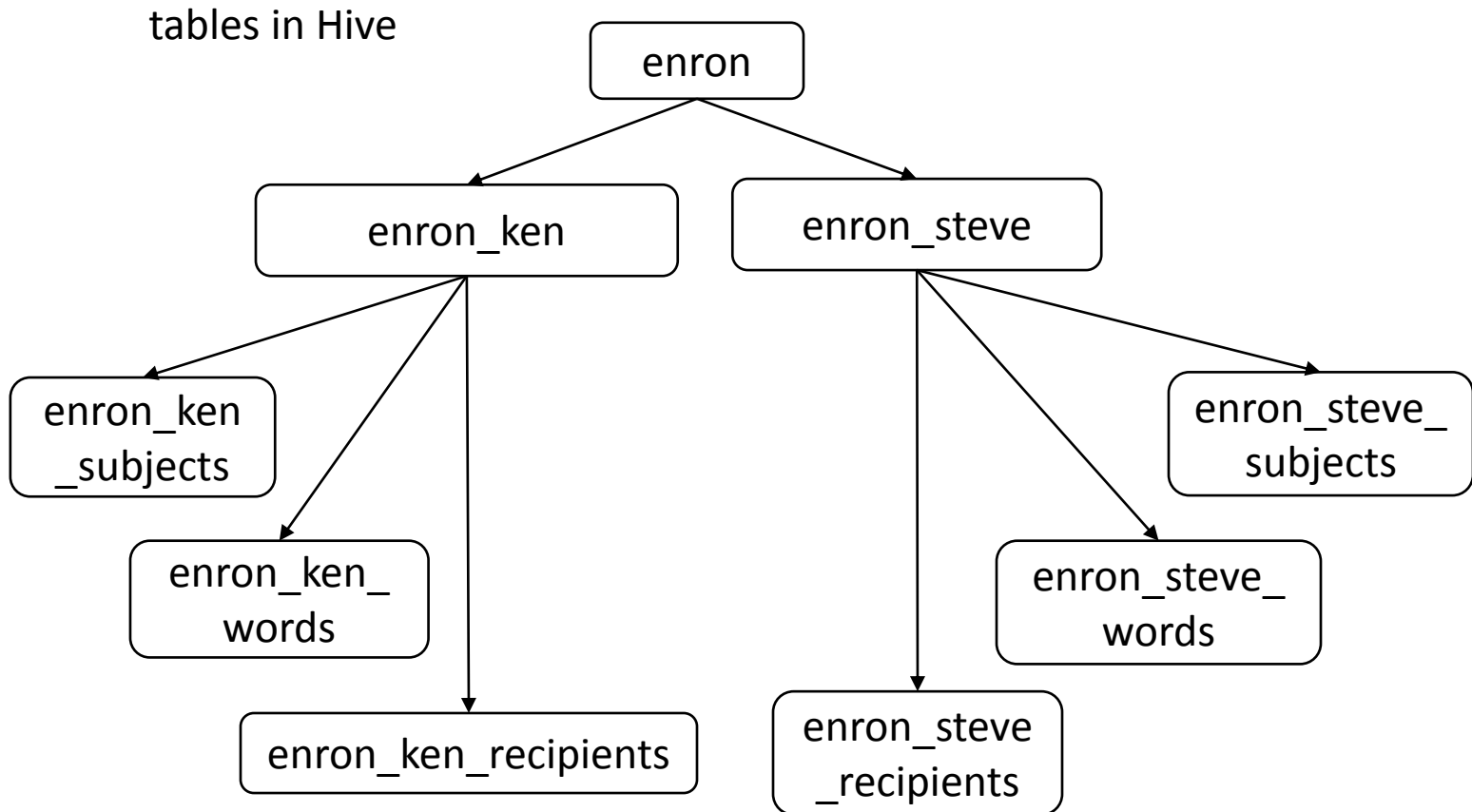
Tae Seung Kang

Yifei Sun

# Enron data processing pipeline (1)

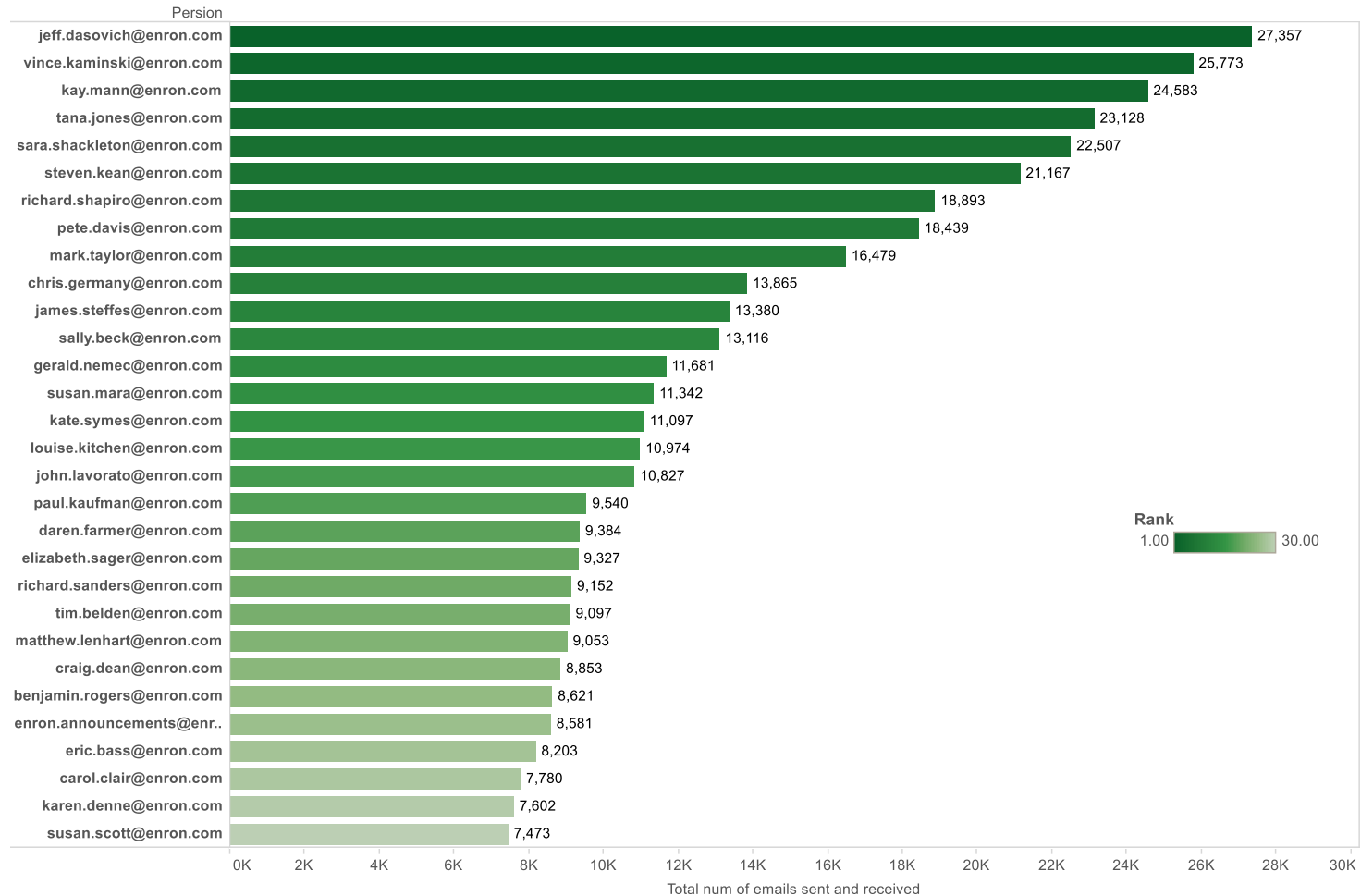


# Enron data processing pipeline (2)



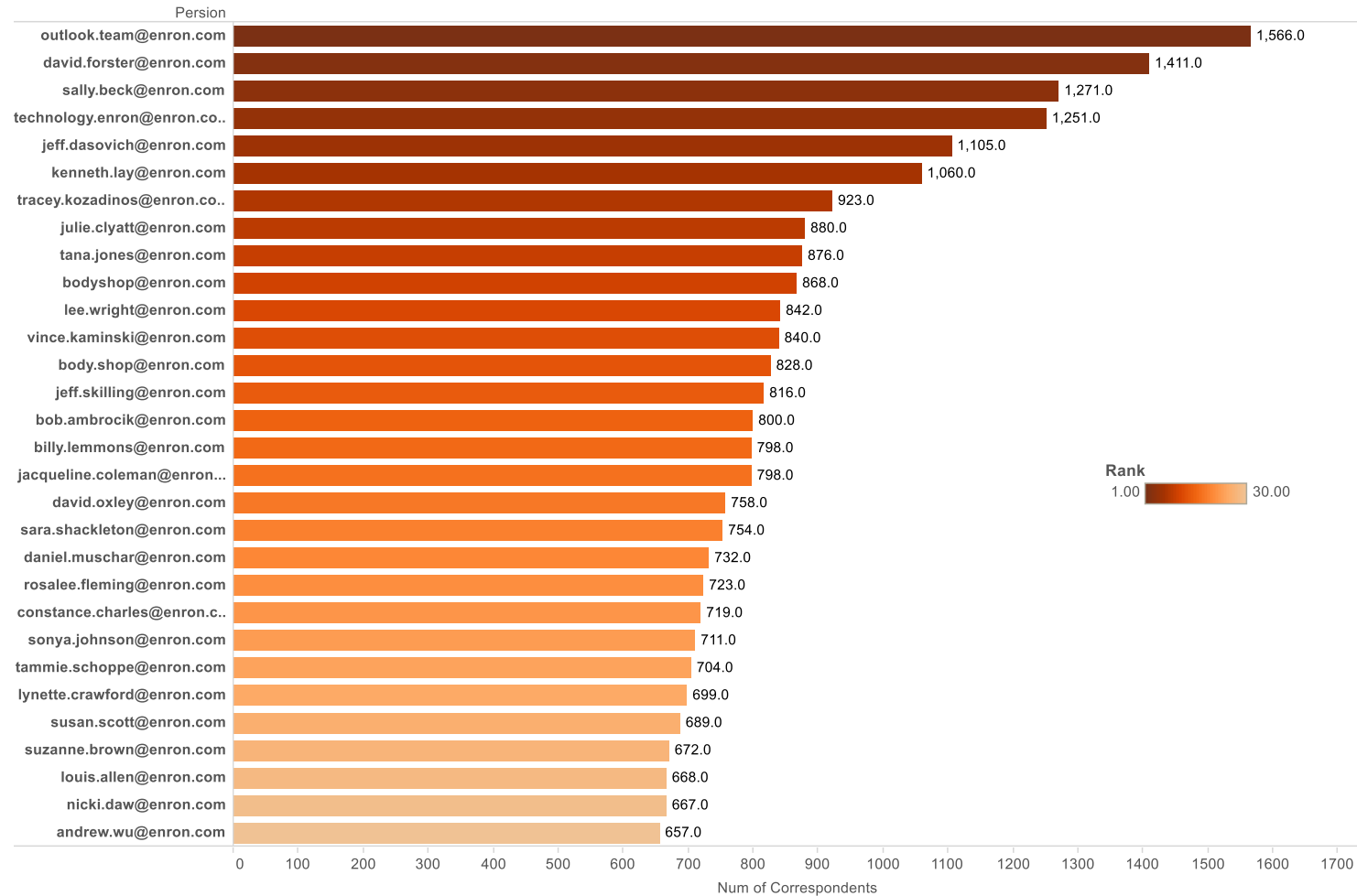
# Enron results (1)

Top 30 persions with most emails sent and received



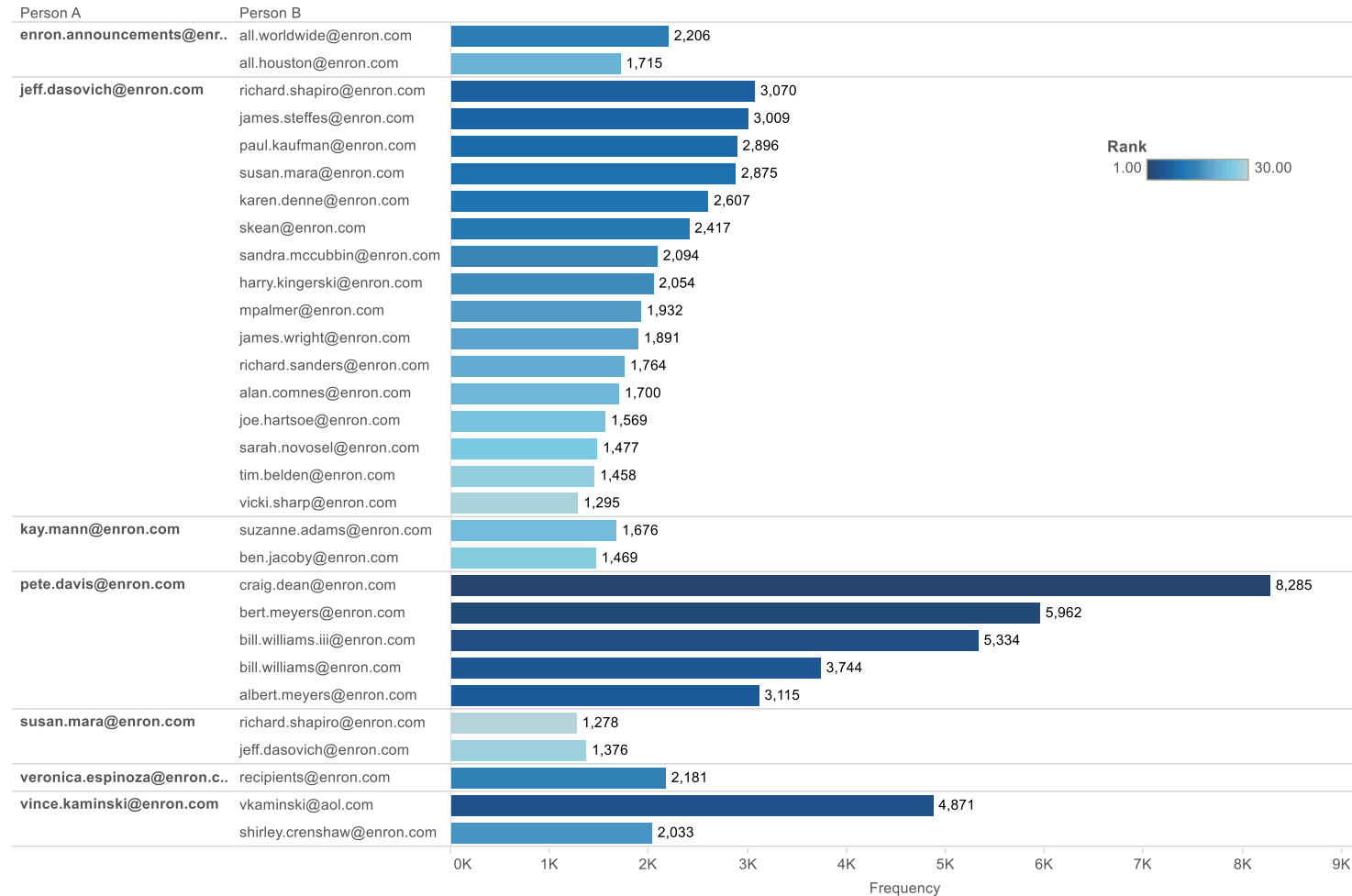
# Enron results (2)

Top 30 persions with most correspondences



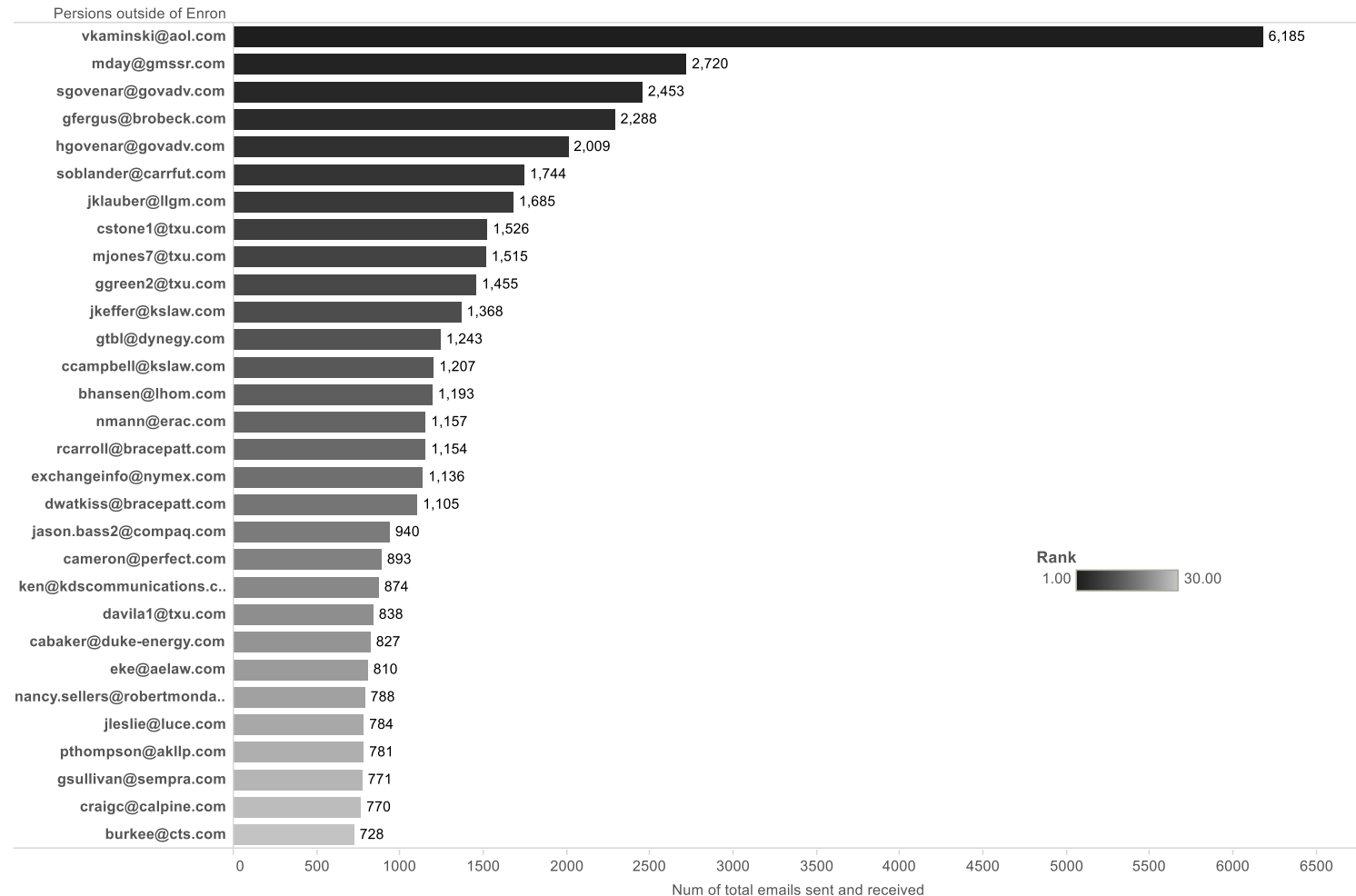
# Enron results (3)

## Persion-pairs with top 30 correspondences



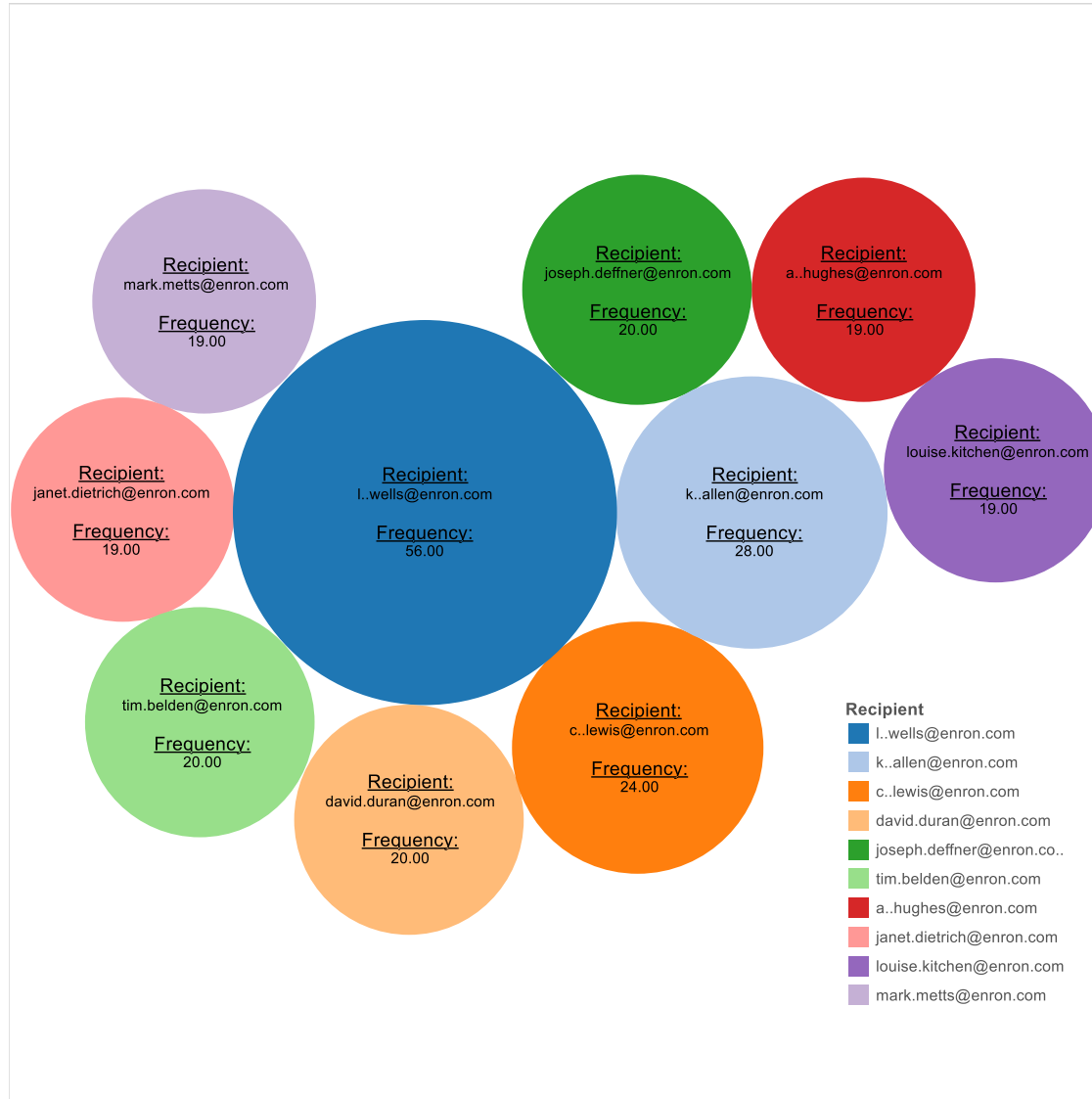
# Enron results (4)

Top 30 persons outside of Enron with most emails sent and received



# Enron results (5)

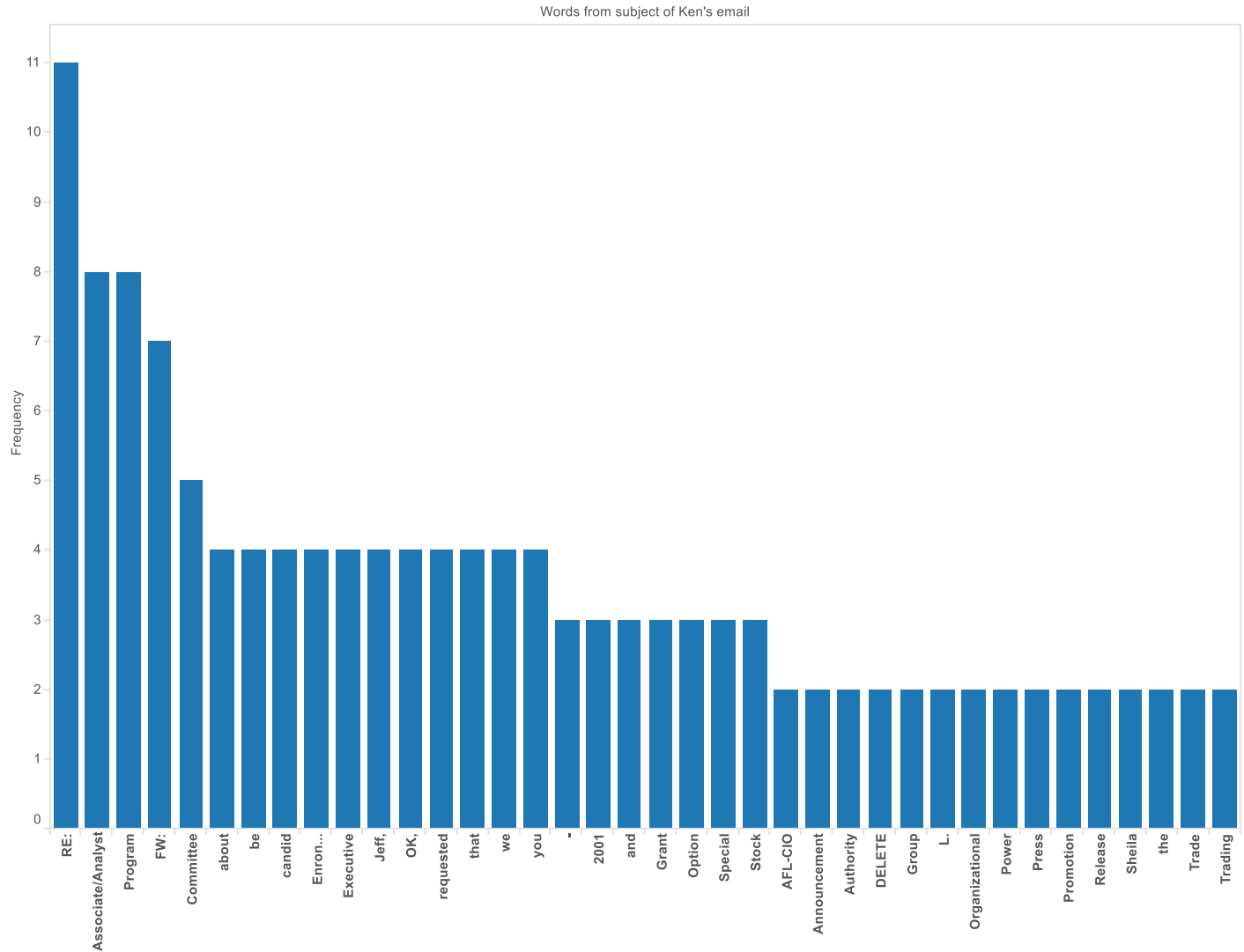
Top 10 recipients of Ken Lay





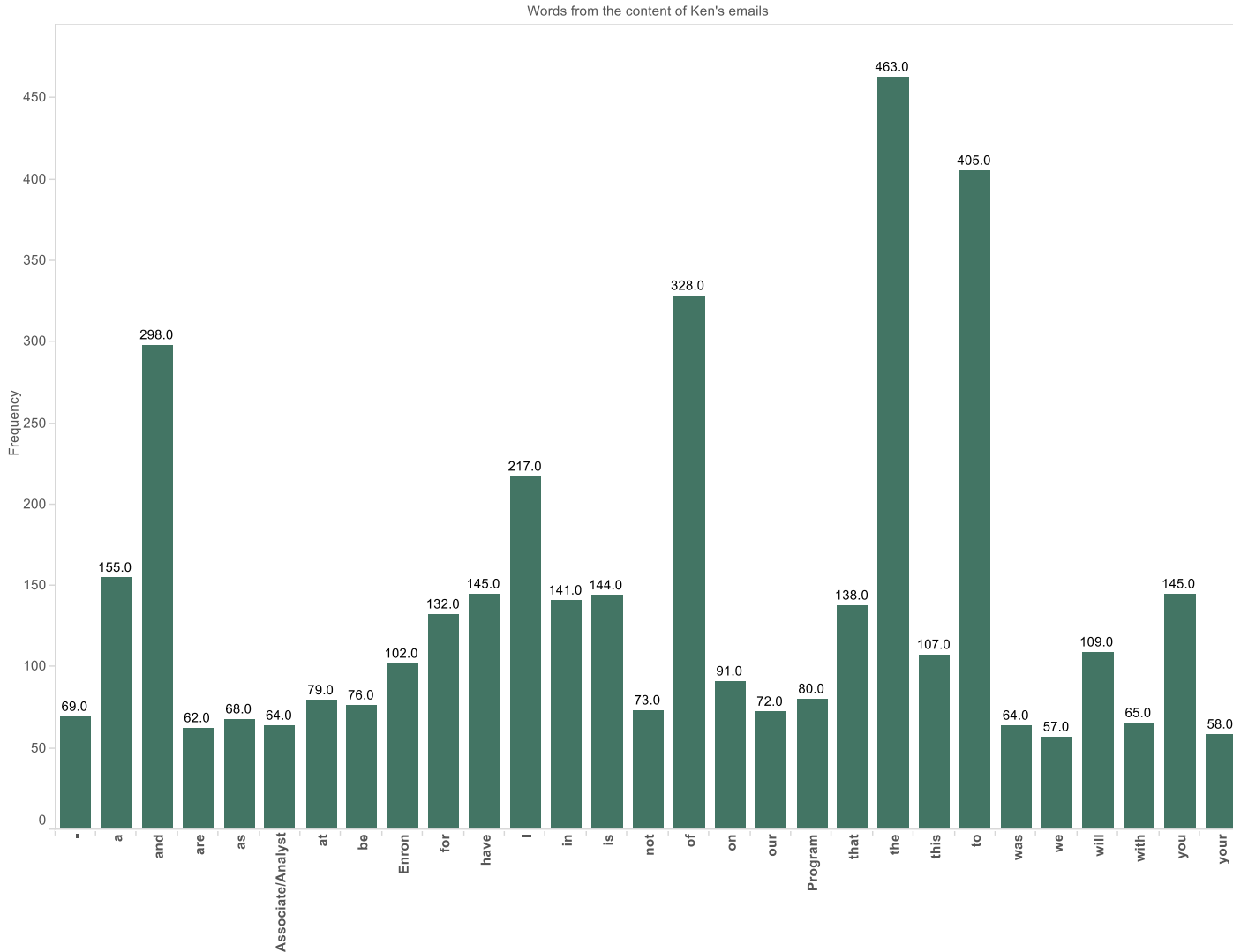
# Enron results (6)

Words from the subject of Ken's emails



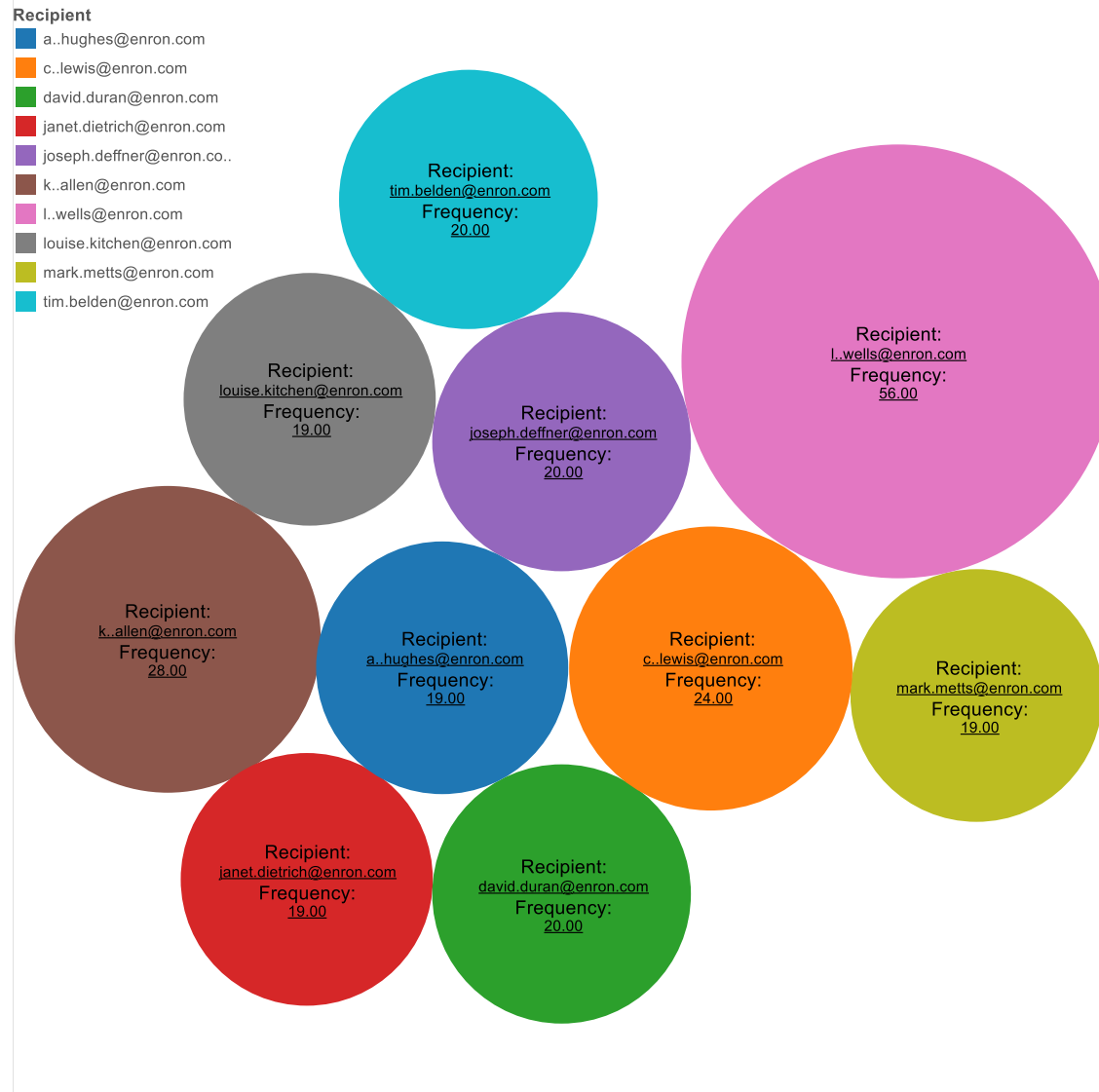
# Enron results (7)

Top 30 most frequently used words from the content of Ken's emails



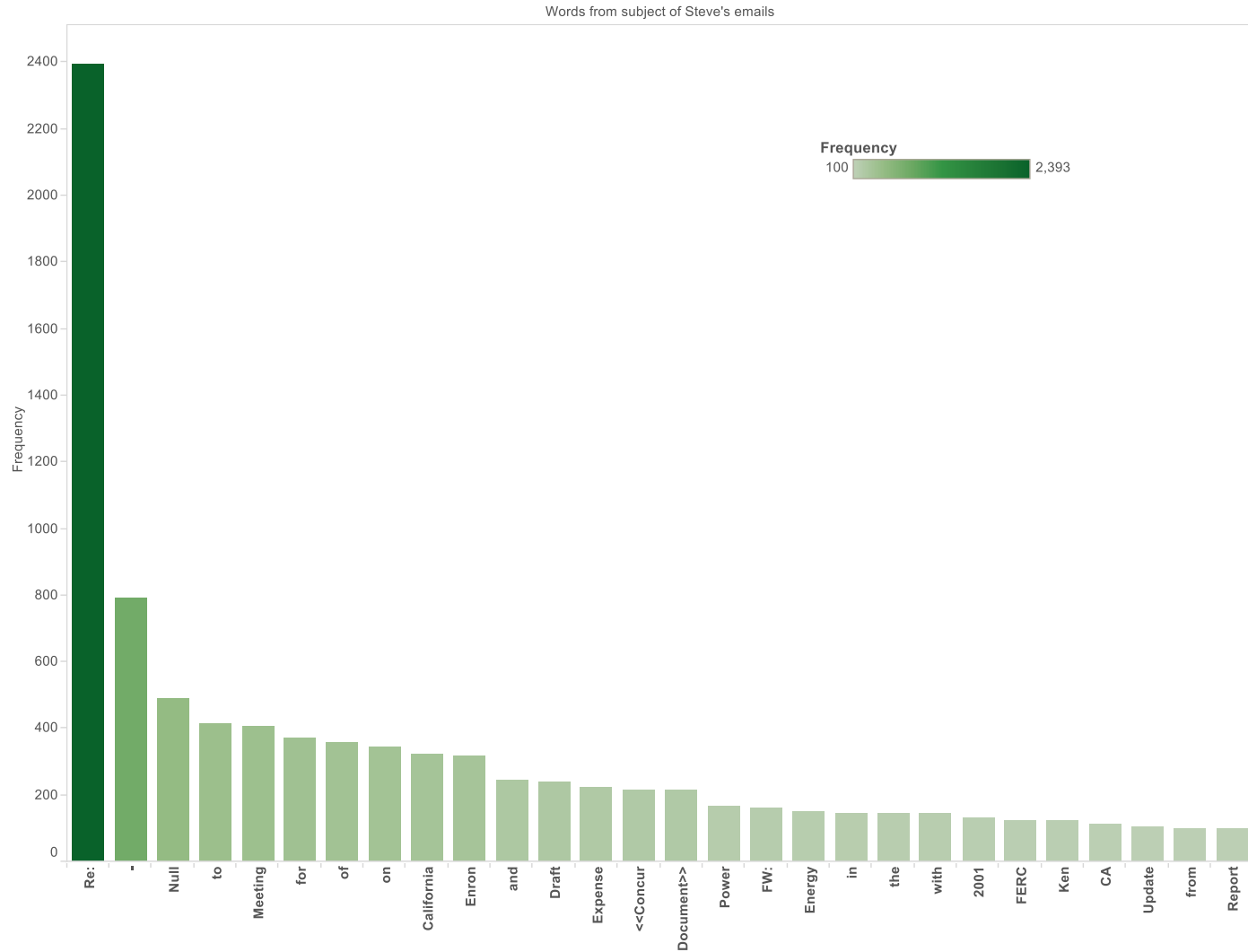
# Enron results (8)

Top 10 recipients of Steve's emails



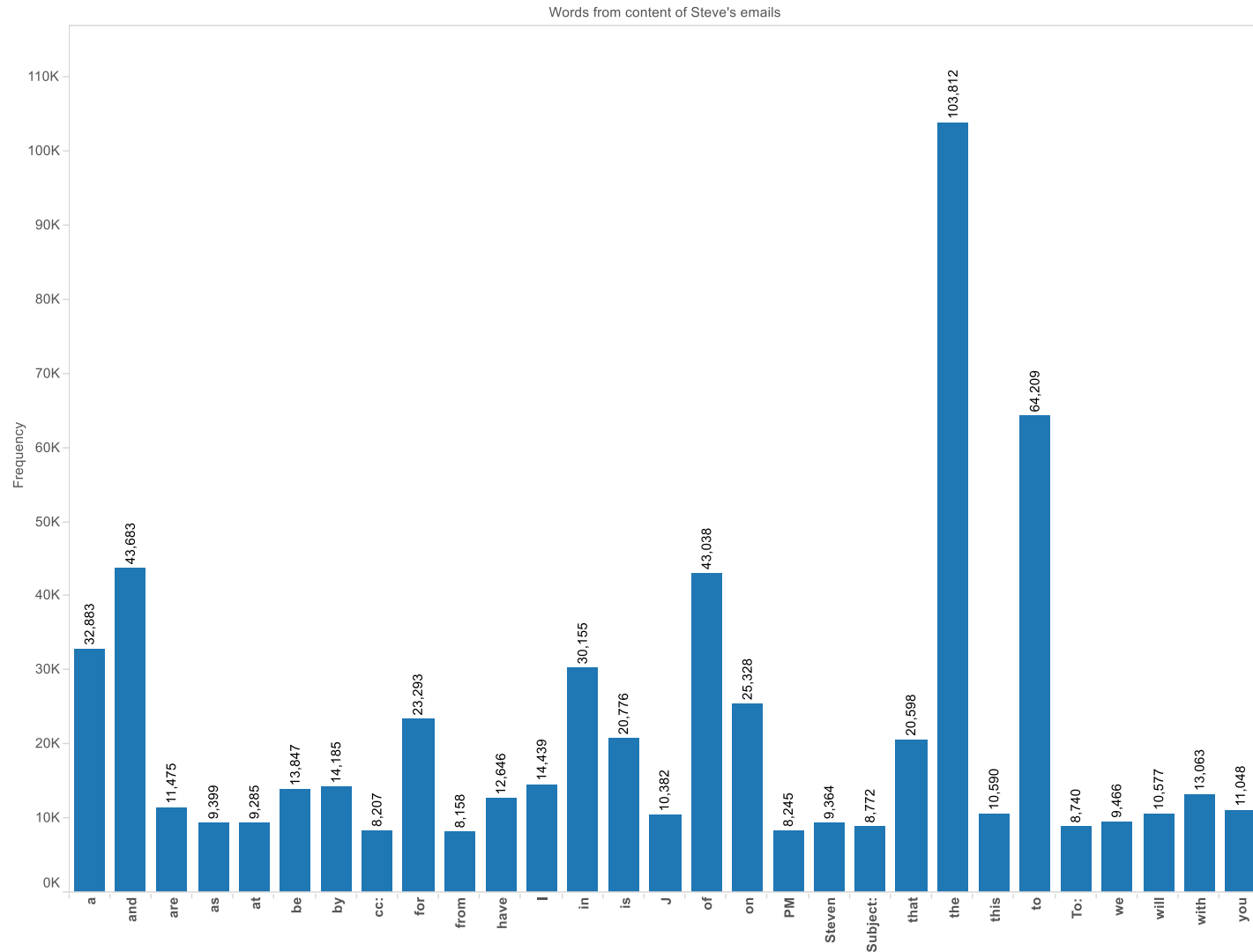
# Enron results (9)

Words from the subject of Steve's emails

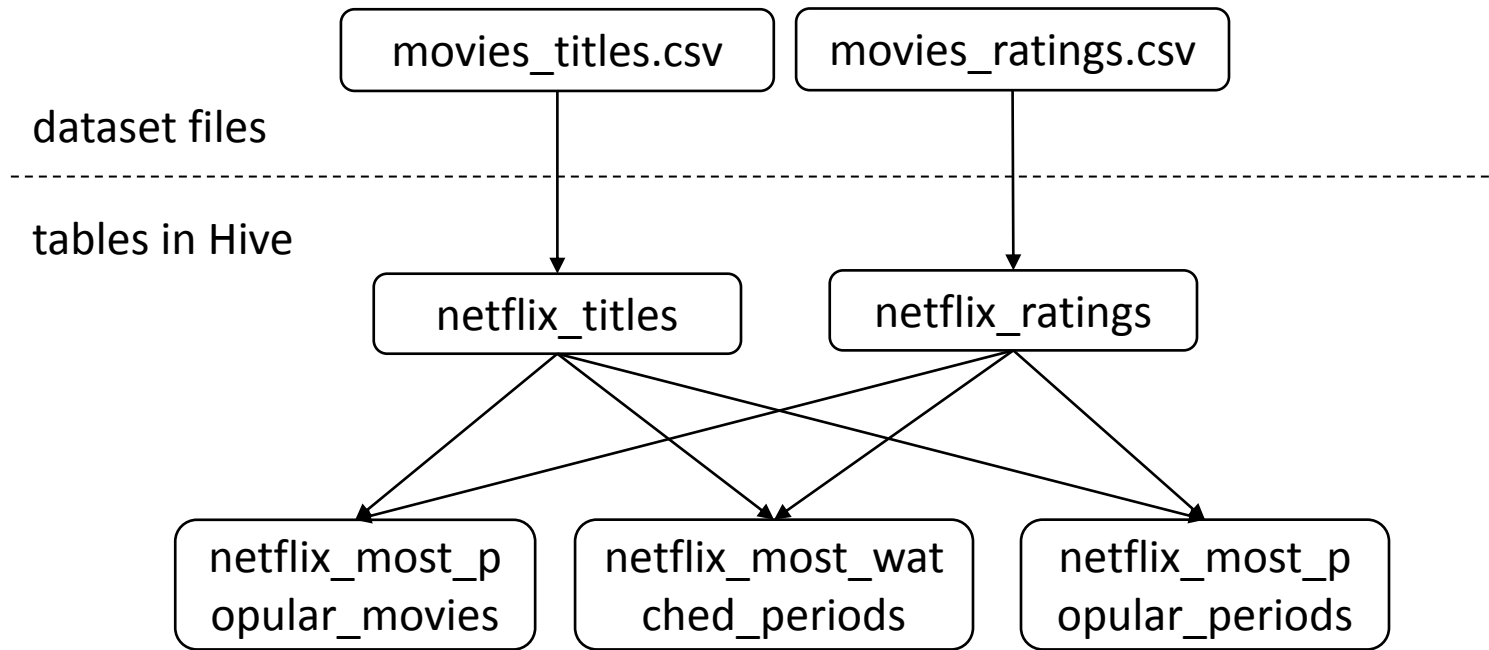


# Enron results (10)

Top 30 most frequently used words from the content of Steve's emails

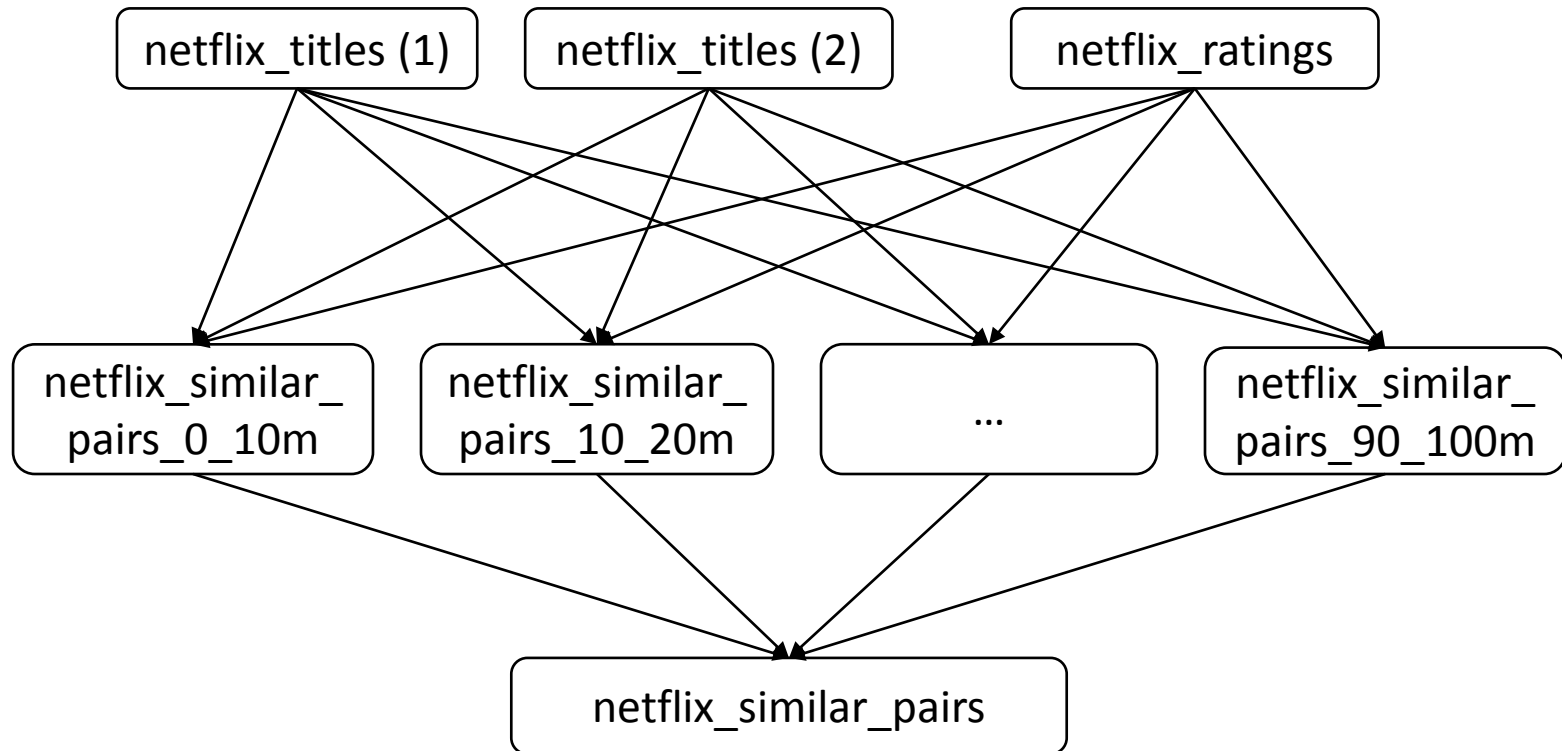


# Netflix data processing pipeline (1)



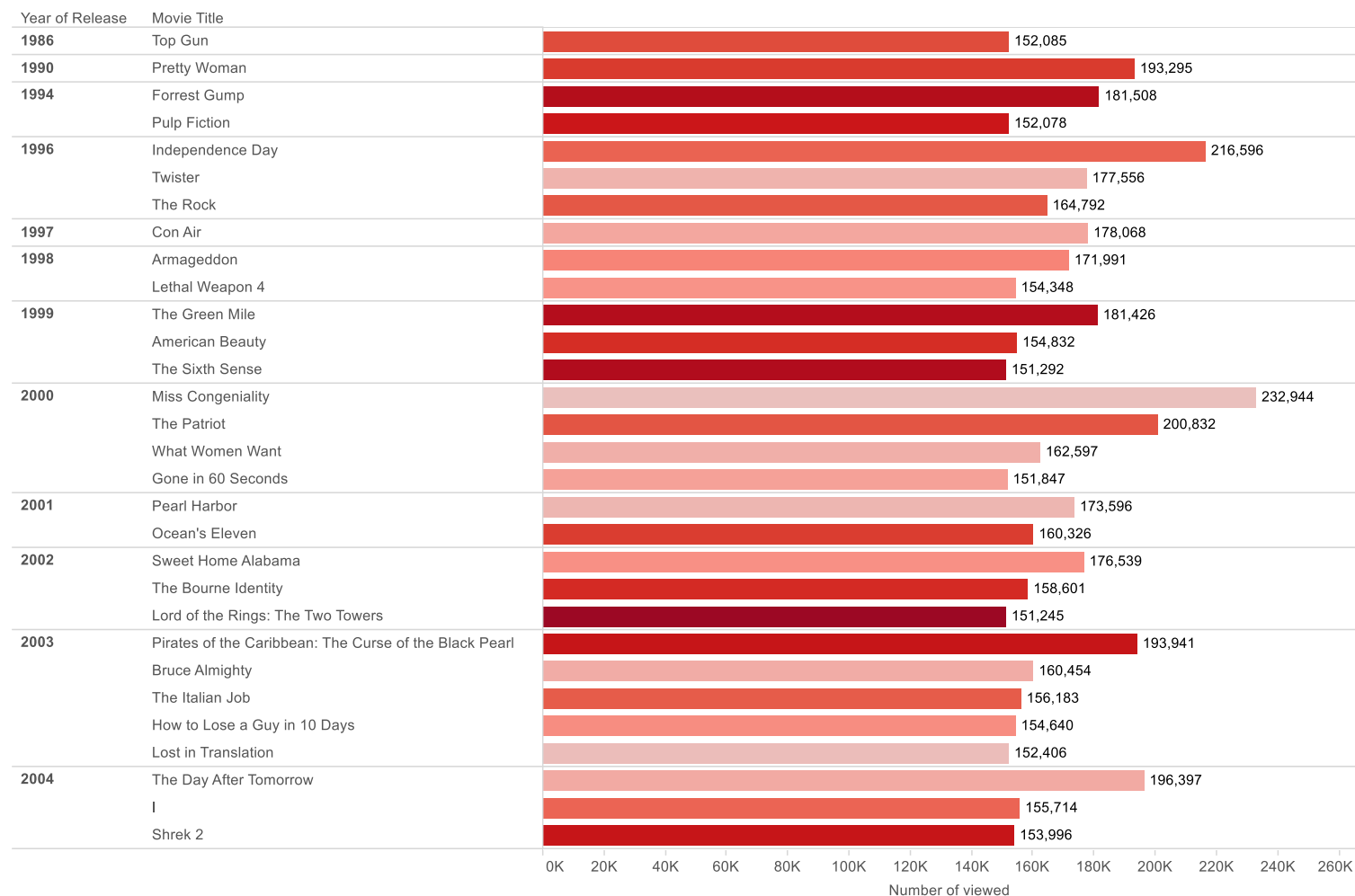
# Netflix data processing pipeline (2)

tables in Hive for most similar movie pairs



# Netflix results (1)

## Most popular movies



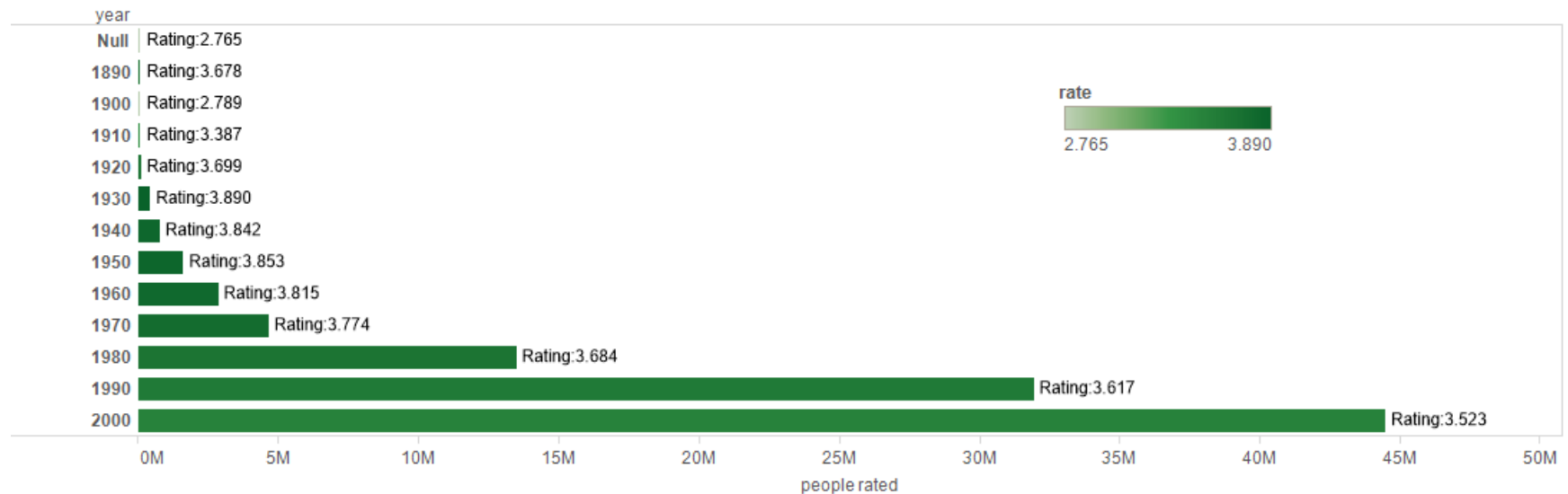
Rating





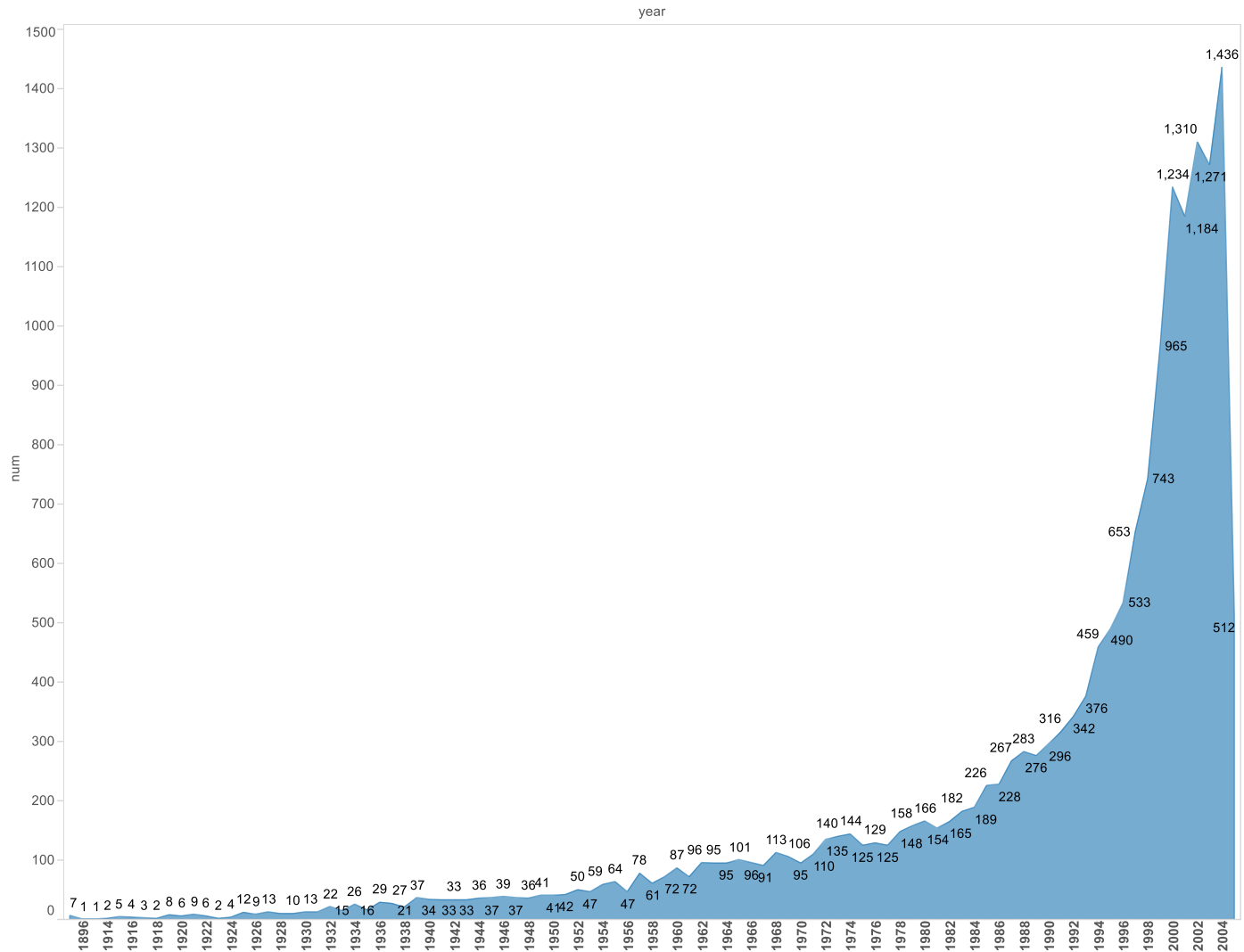
# Netflix results (2)

most\_watched\_periods



# Netflix results (3)

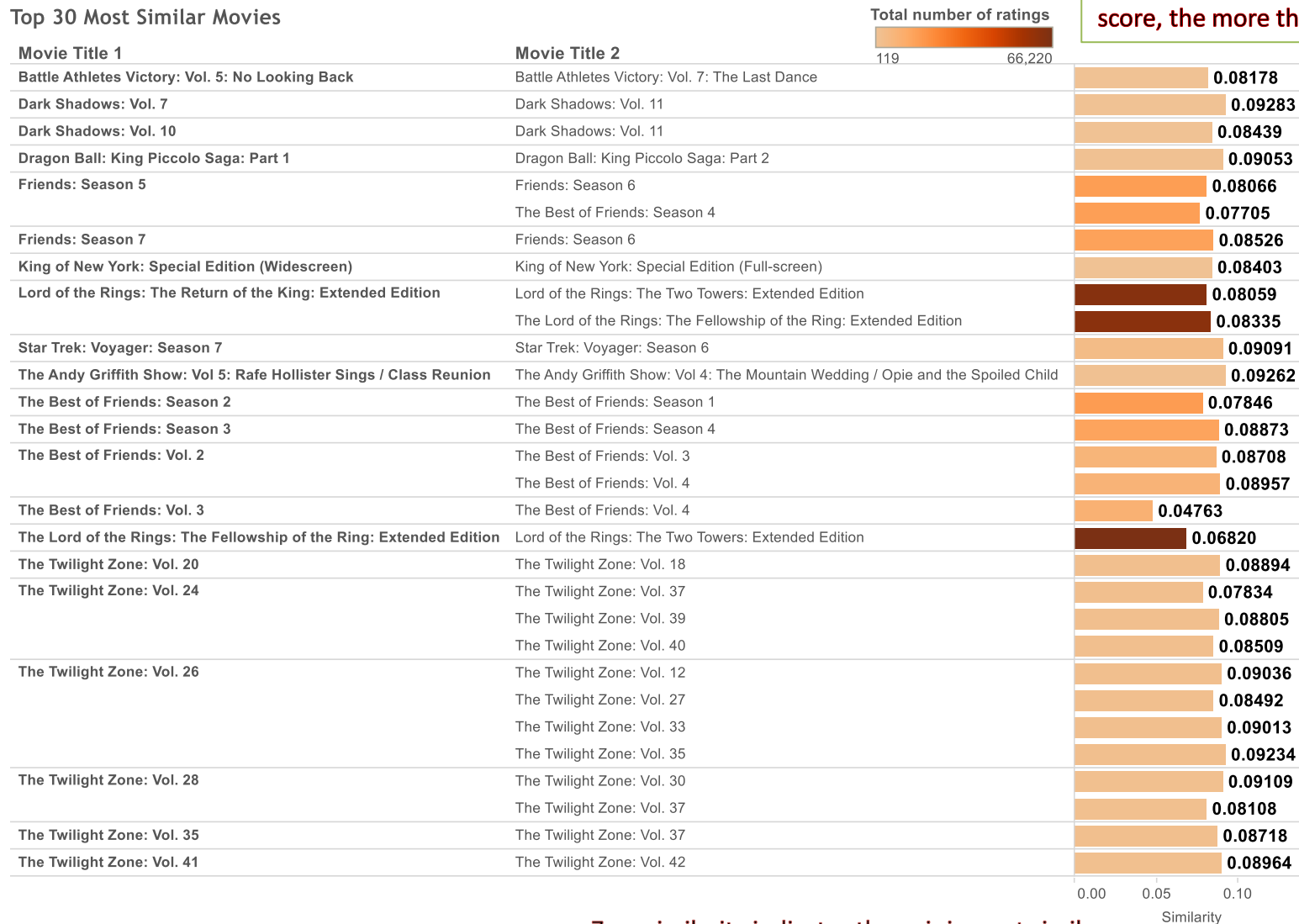
movies per year



# Netflix results (4)

The smaller the similarity score, the more the similar

## Top 30 Most Similar Movies



Zero similarity indicates the pair is most similar.