Data Cleaning

Princeton AI4ALL

Review/Quiz

Find the TF-IDF values of the following. (Hint: this one's easier than you think.)

Keywords: "Prince Jonathan"

Documents:

"The prince's voice was of course familiar to us! Jonathan's speeches had been so memorable."

"The princely figure was none other than our beloved Jonnie."

"Our nation has had many princes, but among the recent ones, Jon has been the most well-known."

Do your answers make sense though?

Today’s Challenge

- Our methods stink on real sentences!
  - Why?
- How do we fix this?

Ideas?

Discuss with a partner:

- Find 2 or more problems with TF-IDF on real sentences
- Propose a few ways to improve TF-IDF for real sentences
  - Remember: the goal of TF-IDF is to find sentences related to the keywords
  - Hint: you may find it easier to modify the sentences themselves, rather than change the TF-IDF algorithm!

What were your ideas?

Example Problems/Solutions

- Tokenize
  - Break into separate words
- Cases (uppercase vs. lowercase)
  - Make everything lowercase!
- Conjunctions
  - Break conjunctions up!
- Multiple word forms
  - Turn everything into the base form!
  - This is hard
- Words that don’t matter (“the”, “and”, etc.)
  - Remove these words!