TF-IDF

Princeton AI4ALL

Today's Challenge

• How to find the **relevance** of a word in a document?

Relevance: Problem Setting

- Ex: You have 10,000 newspaper articles. Find the articles most relevant to some phrase (e.g. "the brown cow").
- How could you approach this problem?
 - Discuss ideas with a partner!

Relevance: Ideas

- Return any article which contains all the keywords
- Return any article containing one keyword or more
- Return the article with the most mentions of the keywords
- Other ideas?

Relevance: Challenges

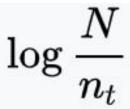
- Filter out common words like "the"
 - O How to do this?
- Don't favor longer articles
 - Output How to do this?

Relevance: Term Frequency (TF)

- The proportion of words which equal the keyword
- Calculate the TF of "cow" for these examples:
 - "The brown cow mooed"
 - o "Cow is a cow"
 - "Cow cow cow cow"
- Higher is better

Relevance: Inverse Document Frequency (IDF)

- What proportion of documents actually contain this word?
 - o "the" gives a high value
 - o "cow" gives a low value
- Take the reciprocal of this, and take the log base 2
 - "the" gives a low value, as desired
 - o "cow" gives a high value



Relevance: TF-IDF

- Combine the two scores by multiplying TF with IDF
- Then, add the score for each word in the query
- Example 1:
 - Query: "cow"
 - Documents: "the cow moos", "the dog barks", "dog dog"
 - o TF for "cow": 1/3, 0, 0
 - O IDF for "cow": log(3/1)≈1.58
 - o TF-IDF scores: 0.52, 0, 0
- Example 2:
 - Query: "the cow"
 - Documents: "the cow moos", "the dog barks", "dog dog"
 - o TF for "the": $\frac{1}{3}$, $\frac{1}{3}$, 0
 - o IDF for "the": log(3/2)≈0.58
 - \circ Total TF-IDF scores: 0.19 + 0.52 = 0.71, 0 + 0.19 = 0.19, 0 + 0 = 0

Practice Time!

Practice Problems:

Calculate the TF-IDF scores for each of the following:

Phrase: "panther" Documents:

- 1) "I like that panther."
- 2) "Panther panther panther."
- 3) "TF-IDF is awesome."
- 4) "Nothing here."
- 5) "Hello."

Phrase: "the computer broke" Documents:

- 1) "We fix computers, we're the best!"
- 2) "If you're broke, you can't buy that."
- 3) "My hamster broke the computer screen."
- 4) "The cat is cute."
- 5) "The book is funny."

Practice Coding

• Time to code TF-IDF in Python!