Discussion 8

Quiz 4 Review

Welcome! Information Retrieval Discussion

November 17, 2021

Today's Itinerary

- Quiz 4 Review

Upcoming Deadlines

- **Wednesday 11:59pm**: Assignment 2 Late deadline
- **Friday 11:59pm**: Assignment 3, Milestone 2
- **Monday in class**: Quiz 4 (due to Thanksgiving holiday)

Quiz 4 Review

Team Trivia

Icebreaker

Instructions

Introduce yourself (or say hello) to your group members!

Going to work in groups for solving the practice quiz problems.

⚠ warning ⚠

FYI I did my best on these problems, however it is possible that I may have made calculation errors, etc. So please double check the work, and make sure to let me know if there are any errors so I can fix them!

Quiz 4 Content

18% Review

- 2 questions from each past quiz material, randomly selected
- 3 points per question

100% turkey puns 🦃

82% New Material

- Jaccard coefficient
- Document frequency
- Inverse document frequency
- tf-idf
- Vector space model
- Cosine similarity
- Calculate window
- Boolean Querying

Definitions & Concepts to Know

- Vector space model
- Cosine similarity
- tf-idf
- Tiered indexes

Jaccard Coefficient

Calculate the Jaccard Coefficient between…

1. Q and S1

2. Q and S2

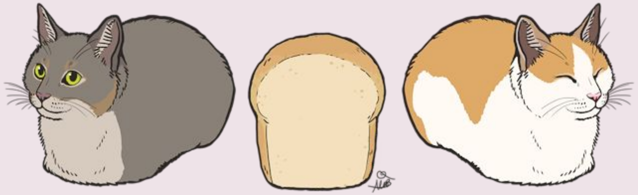
3. Q and S3

S1: "fat cat loaf is the best loaf"

S2: "I like chonk chonk cute cat"

S3: "cat that is black is a burnt loaf"

Q: **kitty loaf**



Jaccard Coefficient

1. **Q and S1**

A = {kitty, loaf}

B = {fat, cat, loaf, is, the, best}

A ∩ B = loaf

|A| = 2

|B| = 6

|A ∩ B| = 1

→ 1 / (2 + 6 - 1) = **1/7**

```
S1: "fat cat loaf is the best loaf"

S2: "I like chonk chonk cute cat"

S3: "cat that is black is a burnt
loaf"

Q: kitty loaf
```

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Document Frequency

What is the document frequency of...

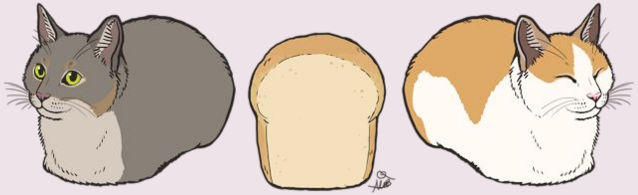
1. cat

2. loaf

S1: "fat cat loaf is the best loaf"

S2: "I like chonk chonk cute cat"

S3: "cat that is black is a burnt loaf"



Document Frequency

1. **cat**

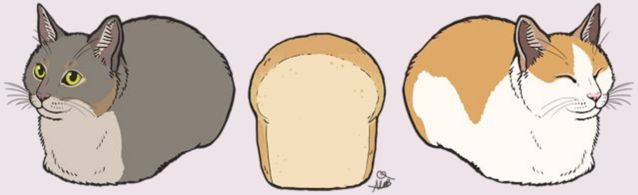
cat appears in all three documents

= 3

S1: "fat **cat** loaf is the best loaf"

S2: "I like chonk chonk cute **cat**"

S3: "**cat** that is black is a burnt loaf"



Inverse Document Frequency

What is the inverse document frequency of...

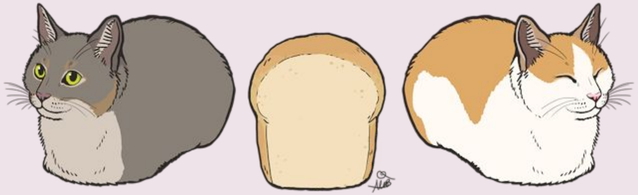
1. cat

2. loaf

S1: "fat cat loaf is the best loaf"

S2: "I like chonk chonk cute cat"

S3: "cat that is black is a burnt loaf"



Inverse Document Frequency

1. cat

$$\mathrm{idf}_t = \log_{10}(N/\mathrm{df}_t)$$

$$df_t = 3$$

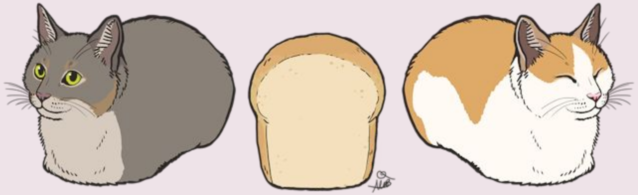
$$N = 3$$

$$log_{10}(3/3) = 0$$

S1: "fat cat loaf is the best loaf"

S2: "I like chonk chonk cute cat"

S3: "cat that is black is a burnt loaf"



tf-idf

What is the tf-idf of...

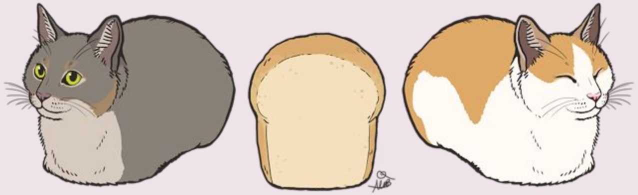
1. The word **loaf** in S1?

2. The word **that** in S3?

S1: "fat cat loaf is the best loaf"

S2: "I like chonk chonk cute cat"

S3: "cat that is black is a burnt loaf"



tf-idf

The word **loaf** in S1?

1. Calculate tf in S1: **2**

2. Calculate idf $log(N/df_t) = idf_t$

$$log(3/2) = idf_t$$

3. Plug n chug

$$w_{t,d} = (1 + log(tf_{t,d})) \times idf_t$$

$$w_{t,d} = (1 + log(2)) \times log(3/2)$$

$$w_{t,d} = (1+\log(tf_{t,d})) \text{ x } \log(N/df_t)$$

S1: "fat cat **loaf** is the best **loaf**"

S2: "I like chonk chonk cute cat"

S3: "cat that is black is a burnt loaf"

Questions?

Unmute yourself or type in the chat.

Otherwise, give a 👍 Reaction if you understand!

Cosine Similarity

Using **cosine similarity** as the ranking formula, what is the relative ranking of these documents for a query with coordinates [1, 1, 1, 1]?

Consider a vocabulary of 4 words.

Two of the documents have coordinates in that space:

D1: [0, 2, 1, 0]

D2: [1, 0, 1, 1]

Cosine Similarity

Determine the similarity of the documents to the query:

	D1	D2
q	0.671	0.866

However for time purposes in the quiz, you can kind of just look at the two documents and see that D2 is more similar to Q (only 2nd element differs).

Answer: D2, D1

Consider a vocabulary of 4 words.

Two of the documents have coordinates in that space:

D1: [0, 2, 1, 0]

D2: [1, 0, 1, 1]

Using cosine similarity as the ranking formula, what is the relative ranking of these documents for a query with coordinates [1, 1, 1, 1]?

$$\cos(\vec{q},\vec{d}) = \frac{\overbrace{\vec{q}\bullet\vec{d}}^{\text{Dot product}}}{|\vec{q}||\vec{d}|} = \overbrace{\frac{\vec{q}}{|\vec{q}|}\bullet\frac{\vec{d}}{|\vec{d}|}}^{\text{Unit vectors}} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

Boolean Query

Determine the most efficient processing order, if any, for the Boolean query **Q** considering the document frequency information from the table →

Q: T1 AND T2 AND T3

Term	Document Frequency
T1	154,383
T2	623,146
T3	483,259

Boolean Query

Answer:

(T1 AND T3) first, then merge with T2

Q: T1 AND T2 AND T3

Term	Document Frequency
T1	154383
T2	623146
T3	483259

Determine the most efficient processing order, if any, for the Boolean query Q considering the document frequency information from the table.

Temperature Check

Give an 👏👍❤️😀😂😮👎😰😭 Emoji Reaction that shows how comfortable you are with your understanding.

Next Week's Discussion

Tentative plan for next week's discussion based on upcoming course deadlines.

- 📢 No discussion due to Thanksgiving holiday
  - 📢 Go home and bake some brownies!

Recommended Homework

To best prepare for next week's session, I recommend you do the following.

- 📓 Finish Assignment 3 Milestone 2
- 📔 Study for the quiz!
  - 📢 **Pro tip**: Combine lecture slides into 1 PDF for easy searching during quiz
  - 📢 Students get [Adobe Acrobat for free]
  - 📢 Use the "Combine PDF" tool
- 📓 Get started on Milestone 3