Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Annotation and Feature Engineering

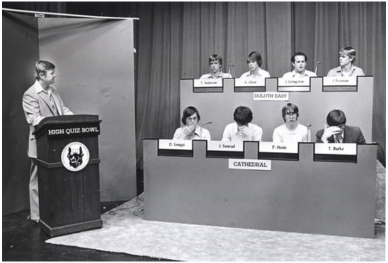
Introduction to Data Science Algorithms

Jordan Boyd-Graber and Michael Paul

HOUSES, SPOILERS, AND TRIVIA

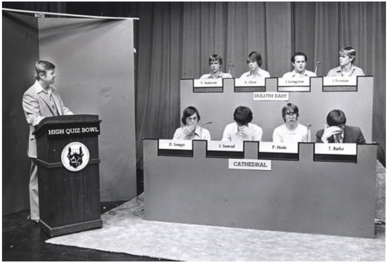
Humans doing Incremental Classification

- Game called “quiz bowl”
- Two teams play each other
  - Moderator reads a question
  - When a team knows the answer, they signal (“buzz” in)
  - If right, they get points; otherwise, rest of the question is read to the other team
- Hundreds of teams in the US alone



Humans doing Incremental Classification

- Game called “quiz bowl”
- Two teams play each other
  - Moderator reads a question
  - When a team knows the answer, they signal (“buzz” in)
  - If right, they get points; otherwise, rest of the question is read to the other team
- Hundreds of teams in the US alone
- Example . . .



Sample Question 1

With Leo Szilard, he invented a doubly-eponymous

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the special and general theories of relativity?

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the special and general theories of relativity?

Albert Einstein

Humans doing Incremental Classification

- This is **not** Jeopardy
- There are buzzers, but players can only buzz at the end of a question
- Doesn't discriminate knowledge
- Quiz bowl questions are pyramidal

Research Question: How do we know if a guess is correct?

- Turn (question, guess) into features
- Treat it as a binary classification problem
- What features help us do this well?

Research Question: How do we know if a guess is correct?

- Turn (question, guess) into features
- Treat it as a binary classification problem
- What features help us do this well?
- Subject of HW3

Provided Dataset

- **text**: the clues revealed so far
- **page**: a guess at the answer
- **answer**: the actual answer (closest Wikipedia page)
- **body_score**: IR measure of how good a match the text is

Baseline

- What if we always say that the answer is wrong?
- Performance: 0.54
- Every feature should do better than this (otherwise, it’s useless)

Page Name

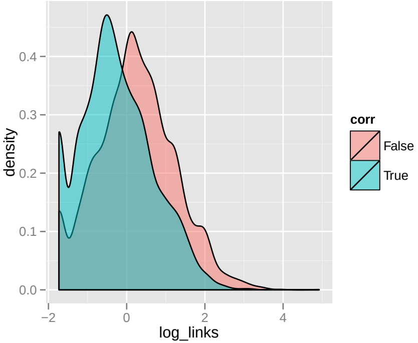
- The title of wikipedia pages often have disambiguation in parentheses
  1. Paris (mythology)
  2. Paris (song)
  3. Paris (genus)
  4. Paris (band)

Page Name

- The title of wikipedia pages often have disambiguation in parentheses
  1. Paris (mythology)
  2. Paris (song)
  3. Paris (genus)
  4. Paris (band)
- Feature is 1 if the page has disambiguator in the text
  - "This band performed . . . ", Paris (band) → `True`
  - "This band performed . . . ", Paris (mythology) → `False`
- Slight improvement: 0.58

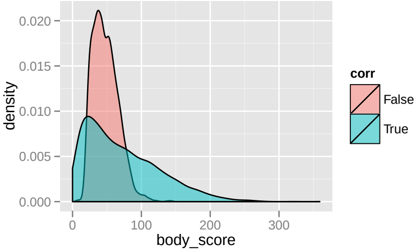
Links

- The more more links a Wikipedia page has, the more popular it is
- Popularity is often a sign of a **wrong answer**
- By itself, doesn’t do so well: 0.56
- But improves if we take the log of the value: 0.61



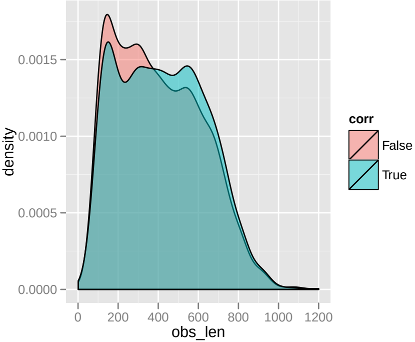
Score

- We can see how similar the text of a Wikipedia page is
- Higher, the better
- This feature alone gives accuracy of 0.75



Length

- The more text we see, the more confident we should be
- By itself, doesn't do so well: 0.56
- But when combined with the IR score, does great: 0.82 (best so far)



Others . . .

- Tournament the question was used in
- The type of thing the answer is
- Try your own, be creative!
- Last year’s feature engineering assignment