WIKIPEDIA PAGES CLASSIFICATION FOR ANSWERING JEOPARDY-LIKE QUESTIONS

Tudor Lechințan, Mădălin Raiu, Andrei Lupșa

Code Description

Uses Java 17 and Apache Lucene

Document Parser

- Splits files into documents.

- Each document has a title, content, chapters and sub-chapters.

- Only one document is loaded in memory at any given time.

Index Builder & Searcher

- Lucene's EnglishAnalyzer was used

- Queries are made up of question text and category.

- Only the document content is used in searching.

Command Line Interface

- Offers options to create and delete the document index.

- Allows users to run an arbitrary Lucene query.

- Has an option to answer questions from a questions file.

Performance Measurement

1. Precision at 1 (P@1)

- Measures the relevance of the top result

- Useful for tasks with an all-or-nothing requirement

2. Correctly answered questions

- Our program yielded 20 correct answers out of 100 questions

- This indicates that when the clues are straightforward the system can perform reliably

Error Analysis

1. Misinterpretation of Clue – answer is related but not correct

Clue: The Naples Museum of Art

Answer : Michael Bierut

Correct: Florida

2. Overgeneralization – answer is related but too general

Clue: 1983: "Beat It"

Answer : Fontana Records

Correct: Michael Jackson

Error Analysis

3. Misinterpretation of Context – answer relates to wrong parts of the clue

Clue: The dominant paper in our nation's capital, it's among the top 10 U.S. papers in circulation

Answer : Media of the United States

Correct: The Washington Post

4. Lack of sufficient information - answer is incorrect

Clue: Song that says, "you make me smile with my heart; your looks are laughable, unphotographable"

Answer: Bess, You Is My Woman Now

Correct: My Funny Valentine