2015

Text Analytic System: Document Similarity

Ryan Murphy
*Virginia Commonwealth University*

James Cecil
*Virginia Commonwealth University*

Joseph Contarino
*Virginia Commonwealth University*

Follow this and additional works at: [https://scholarscompass.vcu.edu/capstone](https://scholarscompass.vcu.edu/capstone)

Part of the [Computer Engineering Commons](https://scholarscompass.vcu.edu/capstone)

© The Author(s)

Downloaded from [https://scholarscompass.vcu.edu/capstone/10](https://scholarscompass.vcu.edu/capstone/10)
Abstract

- Text analytics is a critical function to knowledge discovery.
- Our algorithm processes web-based texts embedded in HTML pages and analyzes them to determine similarity.
- By analyzing the similarity of these HTML documents, we are helping the Idaho National Laboratory to keep redundant data out of the database. Without proper parsing of similar data, repetitive entries may clog the system with unneeded information.

Term-Document Matrix: A method of analyzing the frequency of terms among documents. Rows correspond to documents, and columns correspond to terms.

Cosine Similarity: One way of describing how similar two documents are is to treat their rows in the term-document matrix as vectors, and then calculate the angle between them.

\[ \cos(\theta) = \frac{\text{Vector}_1 \cdot \text{Vector}_2}{||\text{Vector}_1|| ||\text{Vector}_2||} \]

The result of the above matrix is:

\[ \cos(\theta) = \frac{1 \cdot 1 + 1 \cdot 1 + 0 \cdot 0 + 1 \cdot 0}{\sqrt{1^2 + 1^2 + 0^2 + 0^2} \cdot \sqrt{1^2 + 1^2 + 0^2 + 1^2}} \]

Our program takes in two URLs as input parameters. We remove the stop words from the text. Next it parses out the article text from html. A term document is created based on the frequency of the words. We insert the unique words into a wordmap and count the frequencies. A similarity probability is generated for the three text features. The outputs are then fed into a neuron for categorization. A final overall similarity probability for the articles is created as a decimal.