# Text Analytic System: Document Similarity 

Ryan Murphy<br>Virginia Commonwealth University<br>James Cecil<br>Virginia Commonwealth University<br>Joseph Contarino<br>Virginia Commonwealth University

Follow this and additional works at: http:// scholarscompass.vcu.edu/ capstone
Part of the Computer Engineering Commons
© The Author(s)

## Downloaded from

http://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.


## Overview

| Our program takes in two URLs as input parameters | rticle text from html | We remove the stop words from the text |
| :---: | :---: | :---: |
|  |  |  |
| The cosine similarity is calculated based on term document matrix $\square$ | 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 $\qquad$ The outputs are then is generated for the fed into a neuron for three text features categorization |  | A final overall similarity obability for the articles is created as a decimal |

Natural Language Processing

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

Consider the sentences "I like apples" and "I like oranges".

|  | I | like | apples | oranges |
| :--- | :---: | :---: | :---: | :---: |
| Sentence 1 | 1 | 1 | 1 | 0 |
| Sentence 2 | 1 | 1 | 0 | 1 |

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

$$
\operatorname{Cos}()=\frac{\text { Vector } 1 \cdot \text { Vector } 2}{\mid \text { Vector } 1| | \text { Vector } 2 \mid}
$$

The result of the above matrix is

$$
\operatorname{Cos}()=\frac{1 * 1+1 * 1+1 * 0+0 * 1}{\sqrt{1^{2}+1^{2}+1^{2}+0^{2}} * \sqrt{1^{2}+1^{2}+0^{2}+1^{2}}}
$$

Idaho National Laboratory

- Results from cosine similarity analysis fed into an artificial neuron, which separates data into two fields: similar and dissimilar

- The neuron uses a soft activation function, which biases the output naturally towards either 0.0 or 1.0, representing dissimilar and similar comparisons, respectively

