Recitation 6

CS435: Introduction to Big Data

GTA: Bibek R. Shrestha
Email: cs435@cs.colostate.edu

Feb 23, 2018
Today...

- Introduction to Programming Assignment 2
Term Frequency (TF) Calculation

- Term Frequency (TF)

\[ TF_{ij} = 0.5 + 0.5 \left( \frac{f_{ij}}{\max_k f_{kj}} \right) \]

- Suppose we have a collection of documents written by M authors. This collection of documents may contain multiple books written by an author. Let's define a set of documents written by same author as a sub-collection \( j \).

- We define \( f_{ij} \) to be the frequency (Number of occurrences) of term (word) \( i \) in sub-collection \( j \).

- And, \( \max_k f_{kj} \) is the maximum raw frequency of any term \( k \) in the sub-collection \( j \).

- The most frequent term in the sub-collection will have a augmented TF value of 1.
Inverted Document Frequency (IDF) Calculation

- Suppose that term $i$ appears in $n_{ij}$ sub-collections within the corpus. Then,

$$IDF_i = \log_{10}(N/n_i)$$

- where, $N$ is the total number of sub-collections (number of authors).
TF-IDF Calculation

- The TF-IDF score is defined as:

\[ TF_{ij} \times IDF_i \]

- The terms with the highest **TF-IDF** score are considered the best words that characterize the document.
The result of performing calculations above is that each set of books written by the same author in your corpus will have an Author Attribute Vector (AAV),

$$AAV_m = (TF.IDF_{word1}, TF.IDF_{word2}, TF.IDF_{word3}, \ldots, TF.IDF_m)$$
The AAVs will be used to calculate the Cosine Distance to measure the similarity between the authors’ writing styles.

Suppose that we have two authors with vectors,

\[ AAV_1 = [x_1, x_2, \ldots, x_m] \] and \[ AAV_2 = [y_1, y_2, \ldots, y_m]. \]

The Cosine Similarity between them is defined as,

\[
\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}
\]

where, \( A_i \) and \( B_i \) are components of \( AAV_1 \) and \( AAV_2 \) respectively.

Finally, this approach will be able to provide a ranked list of authors (top 10 most similar authors) for a document with unknown authorship.
Demo

- Demo on using Hadoop & Spark maven project
Reading

- http://www.tfidf.com/
In the next recitation...

- Continue Programming Assignment 2.
- Discuss common issues that might occur during MapReduce implementation.