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 Wikipedia articles.
- For a Wikipedia article \( j \), we define \( f_{ij} \) to be the frequency (Number of occurrences) of term/word \( i \) in the article \( j \).
- And, \( \max_k f_{kj} \) is the maximum raw frequency of any term \( k \) in the article \( j \).
- The most frequent term in the sub-collection will have a augmented TF value of 1.
Suppose that term $i$ appears in $n_i$ articles within the corpus. Then,

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

where, $N$ is the total number of articles.
TF.IDF Calculation

- For term \( i \) in articles \( j \), TF.IDF value is:
  \[
  \text{TF.IDF} = \text{TF}_{ij} \times \text{IDF}_i
  \]
Scoring Each Sentence

- We score each sentence in the article.
- For each sentence $S_k$, calculate $\text{Sentence.TF.IDF}$ as:

$$\text{Sentence.TF.IDF}(S_k) = \sum \text{top n TF.IDF (word)}$$
Reading

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

- Continue Programming Assignment 2.
- Discuss common issues that might occur while programming.