Text Summarization for Compressed Inverted Indexes and Snippets

Mangesh Dahale
Text Summarization using Intersection Function

- Intersection Function:
  
  \[ f(s_1, s_2) = \left| \{ w \mid w \text{ in } s_1 \text{ and } w \text{ in } s_2 \} \right| / ((|s_1| + |s_2|) / 2) \]

- If two sentences have a good intersection, they probably hold the same information.
- If one sentence has a good intersection with many other sentences, it probably holds some information from each one of them.
What is a centroid?

- A centroid is a set of words that are statistically important to a cluster of documents.

Each document is represented as a weighted vector of TF-IDF.
Centroid Algorithm

- It first generates a centroid by using only the first document in the cluster. As new documents are processed, their TF–IDF values are compared with the centroid using the formula

$$sim(D, C) = \frac{\sum_k (d_k * c_k * idf(k))}{\sqrt{\sum_k (d_k)^2} \sqrt{\sum_k (c_k)^2}}$$
Three features to compute the quality of a sentence

- Centroid value
  \[ C_i = \sum_w C_{w,i} \]

- Positional value
  \[ P_i = \frac{(n-i+1)}{n} * C_{\text{max}} \]

- First-sentence overlap
  \[ F_i = \bar{S}_1 \bar{S}_i \]
Combining three parameters

\[ \text{SCORE}(s_i) = w_c C_i + w_p P_i + w_f F_i \]

- **INPUT**: Cluster of \( d \) documents with \( n \) sentences (compression rate = \( r \))
- **OUTPUT**: \((n \times r)\) sentences from the cluster with the highest values of \( \text{SCORE} \).
Text Summarization using TF-IDF

- Represent the document collection as the set of sentences from all the documents
- \( S = \{s_1, s_2, \ldots, s_n\} \)
- \( T = \{t_1, t_2, \ldots, t_m\} \) represents all the terms in \( S \)
- \( w_{ij} \) associated with term \( t_j \) in sentence \( s_i \) is calculated by the scheme \( \text{tf–isf} \).

\[
w_{ik} = tf_{ik} \times \log\left( \frac{n}{n_k} \right)
\]
The Cosine Similarity

\[ sim(s_i, s_j) = \frac{\sum_{k=1}^{m} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{m} w_{ik}^2 \cdot \sum_{k=1}^{m} w_{jk}^2}}, \quad i, j = 1, \ldots, n \]

- **Coverage:**
  Coverage means that the generated summary should cover all subtopics as much as possible

- **Diversity**
  Sentences in a summary should have little overlap with one another in order to increase diversity


Rasim M. Alguliev, Ramiz M. Aliguliyev, Nijat R. Isazade, Formulation of document summarization as a 0–1 nonlinear programming problem, *Computers & Industrial Engineering, Volume 64, Issue 1, January 2013, Pages 94–102, ISSN 0360–8352,*