Vector Space Scoring

Introduction to Information Retrieval
INF 141
Donald J. Patterson

Content adapted from Hinrich Schütze
http://www.informationretrieval.org
Alternatives to tf-idf

- Sublinear tf scaling
- 20 occurrences of "mole" does not indicate 20 times the relevance
- This motivated the WTF score.

\[
\text{WTF}(t, d) =
\begin{align*}
1 & \text{ if } t\!f_{t,d} = 0 \\
2 & \text{ then return}(0) \\
3 & \text{ else return}(1 + \log(t\!f_{t,d}))
\end{align*}
\]

- There are other variants for reducing the impact of repeated terms
TF Normalization

- Normalize tf weights by maximum tf in that document

\[ ntf_{t,d} = \alpha + (1 - \alpha) \frac{tf_{t,d}}{tf_{max}(d)} \]

- \( \alpha \) is a smoothing term from (0 - 1.0) \~0.4 in practice
- This addresses a length bias.
- Take one document, repeat it, WTF goes up
TF Normalization

- Normalize tf weights by maximum tf in that document

\[ n_{tf_{t,d}} = \alpha + (1 - \alpha) \frac{tf_{t,d}}{tf_{\text{max}}(d)} \]

- a change in the stop word list can change weights drastically - hard to tune
- still based on bag of words model
- one outlier word, repeated many times might throw off the algorithmic understanding of the content
## Laundry List

<table>
<thead>
<tr>
<th>Term Frequency</th>
<th>Document Frequency</th>
<th>Normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n)atural</td>
<td>1 + log(tf_{t,d})</td>
<td>1/(1 + log(tf_{t,d}))</td>
</tr>
<tr>
<td>(l)ogarithm</td>
<td>\alpha + (1 - \alpha) \frac{tf_{t,d}}{df_{max}(d)}</td>
<td>\log \left( \frac{</td>
</tr>
<tr>
<td>(a)ugmented</td>
<td>tf_{t,d} &gt; 0 ? 1 : 0</td>
<td>\max { 0, \log \left( \frac{</td>
</tr>
<tr>
<td>(b)oolean</td>
<td>\frac{1 + \log(tf_{t,d})}{1 + \log(\text{average}(tf_{t,d}))}</td>
<td>\frac{1}{\sqrt{w_1^2 + w_2^2 + ... + w_m^2}}</td>
</tr>
<tr>
<td>(L)ogaverage</td>
<td>\frac{1}{1 + \log(tf_{t,d})}</td>
<td>\frac{1}{u}</td>
</tr>
</tbody>
</table>

• SMART system of describing your IR vector algorithm

• ddd.qqq (ddd = document weighting) (qqq = query weighting)

• first is term weighting, second is document, then normalization

• lnc.ltc is what?
Efficient Cosine Ranking

- Find the k docs in the corpus “nearest” to the query
- the k largest query-doc cosines
- Efficient ranking means:
  - Computing a single cosine efficiently
  - Computing the k largest cosine values efficiently
- Can we do this without computing all n cosines?
  - n = number of documents in corpus
Efficient Cosine Ranking

- Computing a single cosine
- Use inverted index
- At query time use an array of accumulators \( A_j \) to accumulate component-wise sum
- Accumulate scores as postings lists are being processed (numerator of similarity score)

\[
A_j = \sum_t (w_{q,t} w_{d,t})
\]
Efficient Cosine Ranking

• For the web
  • an array of accumulators in memory is infeasible
  • so only create accumulators for docs that occur in postings list
    • dynamically create accumulators
  • put the tf_d scores in the postings lists themselves
  • limit docs to non-zero cosines on rare words
    • or non-zero cosines on all words
      • reduces number of accumulators
**Efficient Cosine Ranking**

\[ \text{COSINE\text{SCORE}}(q) \]

1. \text{INITIALIZE}(\text{Scores}[d \in D])
2. \text{INITIALIZE}(\text{Magnitude}[d \in D])
3. \text{for each term}(t \in q)
   4. \text{do } p \leftarrow \text{FETCH\text{POSTINGS\text{LIST}}}(t)
   5. \quad df_t \leftarrow \text{GET\text{CORPUS\text{WIDE\text{STATS}}}(p)}
   6. \quad \alpha_{t,q} \leftarrow \text{WEIGHT\text{IN\text{QUERY}}}(t, q, df_t)
   7. \quad \text{for each } \{d, tf_{t,d}\} \in p
   8. \text{do } \text{Scores}[d] + = \alpha_{t,q} \cdot \text{WEIGHT\text{IN\text{DOCUMENT}}}(t, q, df_t)
9. \text{for } d \in \text{Scores}
10. \text{do } \text{NORMALIZE}(\text{Scores}[d], \text{Magnitude}[d])
11. \text{return top } K \in \text{Scores}
Use heap for selecting the top K Scores

- Binary tree in which each node’s value > the values of children
- Takes 2N operations to construct
- Then each of k “winners” read off in 2logn steps
- For n =1M, k=100 this is about 10% of the cost of sorting