605.744: Information Retrieval
Final Exam

Sabbir Ahmed

December 4, 2022

1. (a) Suppose the document collection is very large, and that the index will not fit in the available RAM. Describe an indexing algorithm that works when memory is small compared to the size of the index.

   (b) Once an inverted file has been created, it is possible to calculate document vector lengths for a TF/IDF cosine model. This pre-calculation makes query-time performance much more efficient. Explain how right after creating the inverted file document vector lengths can be efficiency computed for all docids in parallel using one traversal (i.e., one single pass) over the inverted file.

Table 1: Cardinalities of set-difference sets with various n-grams and normalization parameters

N-Gram	**Normalized**	$\|G - O\|$	$\|O - G\|$
6	True	10	12
6	False	10	12
5	True	3	6
5	False	4	9
4	True	3	6
4	False	3	6
3	True	2	4
3	False	2	4
2	True	1	2
2	False	3	2
1	True	2	1
1	False	2	2

2. The three major problems in text retrieval are: (a) polysemy; (b) synonymy; and, (c) morphology. Briefly explain each issue and how it can lower performance. Give an example of each phenomena.

   Answer:

   (a) Polysemy refers to words that can have multiple meanings depending on the context. For example, *space* can refer to its noun version of unoccupied area. The unoccupied area can be physical or abstract, i.e. “the space between planets” or “a teenager needing their own personal space”. The word can also be used as a verb to refer to a person physically or emotionally distancing themselves from a situation, i.e. “spacing out during lectures”.

Polysemy introduces ambiguity to a retrieval if a query is not given enough context and the system retrieves the undesired version of the word.

(b) Synonymy refers to different words addressing the same meaning. For example, *colossal*, *giant* and *huge* all describe the size of an object to be very big. Numerous other words also act as synonyms for *big*. Synonymy can lower performance of a retrieval system if it is not aware of the numerous synonyms a query word may have. If a user queries for "big company" but the system only contains documents with the numerous synonyms of *big*, the ranked documents may not be what the user implied.

(c) Morphology refers to the various conjugations of a word. In English, adding suffixes such as "s" or "es" transforms a noun into its plural form. Adding suffixes such as "d", "ed", and "ing" transforms a present tense verb to a different tense. Morphology is not only limited to suffixes or prefixes, since there are special cases of words needing a replacement in a character, i.e. *sang* is the past tense form of *swim*, while *sung* is its past participle form. Morphology can introduce issues in a retrieval system if the different variations of the words in a query are not accounted for. These systems often employ some levels of stemming in their dictionary and the query processing to normalize the words to their base forms. However, stemming can introduce additional ambiguity when different words get stemmed to a common word. i.e. "transparent" and "transparency" get stemmed to "transpar" using a Porter stemmer.

3. Short answers about text classification.

(a) What is negative evidence in Binomial (also called Bernoulli) Naïve Bayes text classification?

Answer:

(b) For a class that attains precision of 0.5 and recall of 0.6, what is the corresponding F1 score?

Answer: The F1 score is computed as $2 \times \frac{P \times R}{P+R}$. Therefore,

$$\begin{aligned} F1 &= 2 \times \frac{P \times R}{P + R} \\ &= 2 \times \frac{0.5 \times 0.6}{0.5 + 0.6} \\ &= 0.54 \end{aligned}$$

(c) For the three classes below (business, local, and sports) with the indicated system predictions, calculate precision for the 12 news articles in two ways: using micro averaging and macro averaging.

Answer: Precision is the ratio of the true positives over the total number of classes labeled positive (both true and false positive classes). In the table, there are a total of 8 true positives, with 2, 3, and 3 true positives and 2, 0, and 2 false positives in the 3 respective classes.

The micro average can be computed as:

$$\begin{aligned}
P_\mu &= \frac{\sum_{i=0}^{n}(TP_i)}{\sum_{i=0}^{n}(TP_i + FP_i)} \\
&= \frac{2+3+3}{(2+2)+(3+0)+(3+2)} \\
&= \frac{8}{12} \\
&= 0.67
\end{aligned}$$

The macro average is computed by taking the expected value of the individual precision scores of the 3 classes.

$$\begin{aligned}
P_b &= \frac{TP_i}{TP_i + FP_i} \\
&= \frac{2}{4} \\
&= 0.5
\end{aligned}$$

$$\begin{aligned}
P_l &= \frac{TP_i}{TP_i + FP_i} \\
&= \frac{3}{3} \\
&= 1
\end{aligned}$$

$$\begin{aligned}
P_s &= \frac{TP_i}{TP_i + FP_i} \\
&= \frac{3}{5} \\
&= 0.6
\end{aligned}$$

$$\begin{aligned}
P_M &= \frac{P_b + P_l + P_s}{3} \\
&= \frac{0.5+1+0.6}{3} \\
&= 0.70
\end{aligned}$$