605.744: Information Retrieval
Problem Set (Module 8)

Sabbir Ahmed

October 19, 2022

1. (20%) Name three significant issues that arise when using dictionaries to translate queries in cross-language information retrieval and briefly explain why they create a problem.

   Answer:

2. (20%) Give two advantages and one disadvantage of using character n-gram tokenization for multilingual text retrieval.

   Answer:

3. (20%) For this question consider an English alphabet to consist of just 26 (lower-cased) letters, 10 digits, and a space character. And consider there to be exactly 10,000 characters in Chinese. Note, spaces are not used in written Chinese.

   (a) How many possible character 4-grams are there in English? Using Table 5.1 (IIR) how does this number compare to a typical vocabulary size when words are used?

   Answer:

   (b) How many possible indexing terms will there be if 2-gram indexing is used for Chinese? What if 3-grams are used?

   Answer:

   (c) What difficulties might occur when indexing a document collection if the vocabulary size (i.e., number of indexing terms) is extremely large?

   Answer:

4. (20%) What advantages does query translation have over document translation in cross-language information retrieval (CLIR)?

   Answer:

5. (20%) Briefly describe what pre-translation query expansion (sometimes called pre-translation feedback) is and then explain why it is helpful in dictionary-based cross-language information retrieval.

   Answer: