Module 1 Programing Assignment

Max Robinson

1 Problem 1

Question: Approximately how many distinct words are there in the English language? Briefly justify your response.

There are approximately 470,000 distinct words in the English Language. This is according to the "help" site for the Merriam-Webster online dictionary (https://www.merriam-webster.com/help/faq-how-many-english-words). They state that there are approximately 470,000 entries in the *Webster's Third New International Dictionary*, Unabridged, along with a 1993 Addenda Section. These entries try and capture words in their original form, disregarding how they might be used for multiple different meanings, or have additional prefixes or suffixes.

2 Problem 2

Question: Give a specific example of a word where case-folding (i.e, lower-casing text) can cause an IR system to make an error that would not normally occur if case distinctions were retained. Briefly explain the error.

A specific example of a word where case-folding can cause an IR system to make an error that would not normally occur if case distinctions were retained is 'Bob' vs. 'bob'. 'Bob' with a capital 'B' usually refers to a name of a person. 'bob' with a lower case 'b' is a verb and can be interpreted as an action rather than a person.

Another example is, 'US' vs 'us' where 'US' is an abbreviation for 'United States' and 'us' is a word referring to a collection of people from an inclusive standpoint.

3 Problem 3

Question: For a conjunctive query, is processing postings lists in order of size guaranteed to be optimal? Explain why it is, or give an example where it isnt.

Processing postings lists in order of size is not guaranteed to be optimal. Consider the following lists:

```
word1: 10, 13, 15
word2: 10, 13, 45, 39
word3: 0, 17, 28, 99, 145
```

Looking for intersections in the posting size order would mean comparing word1, word2, then word3. The number of steps required is 3 + 4 + 2 + 5 = 14 since the intersection of word1 and word2 is 2.

However, if the order checked was word1, word3, word2 it would have been seen that there is no intersection between word1 and word3 in 8 steps (3 + 5) and then the query is then complete. Thus, this order of comparison would be more efficient even though the postings are not in order from smallest to largest.

4 Problem 4

Question: Assume a biword index. Give an example of a document which will be returned for a query of *New York University* but is actually a false positive which should not be returned.

The query assuming a biword index for *New York University* would be: **"new york" AND "york university"**.

From this an example of a document that could be returned as a false positive to this query could be a document such as the following; *It is slightly amusing that Matilda lives in* ***New York*** *now but studied at* ***York University***.