4. Assignment

Assignment 4.1

Consider the following text where each sentence should be seen as a document for simplicity:

> Today is sunny. She is a sunny girl. To be or not to be. She is in Berlin today.
> Sunny Berlin! Berlin is always exciting!

a) Create an inverted index for this document collection with dictionary and postings.

b) For pre-processing only spilt the text in a document based on non-letter characters and then lowercase.

c) Use this index and the Intersection algorithm to find all the documents that contain the words "sunny" and "exciting".

Assignment 4.2

A query "$Term_1$ AND $Term_2$" and postings lists for each term are given:

> $Term_1$: [4, 6, 10, 12, 14, 16, 18, 20, 22, 32, 47, 81, 120, 122, 157, 180]
> $Term_2$: [47]

Work out how many comparisons would be done to *intersect* the two postings lists with the following two strategies. Briefly justify your answers:

a) Using standard postings lists.

b) Using postings lists stored with skip pointers, with a skip length of $\sqrt{L}$[1] where L is the length of a posting list.

Why are skip pointers not useful for queries of the form "$Term_1$ OR $Term_2$"?

Assignment 4.3

Learn about different Index variations:

a) 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.

b) Consider the following fragment of a positional index with the format:

```
word: document: [position, position, ...]; document: [position, ...] ...

Gates: 1: [3]; 2: [6]; 3: [2,17]; 4: [1];
IBM: 4: [3]; 7: [14];
Microsoft: 1: [1]; 2: [1,21]; 3: [3]; 5: [16,22,51];
```

[1] $\sqrt{L}$ elements are skipped in the list.

The /k operator, $word_1$ /k $word_2$ finds occurrences of $word_1$ within k words of $word_2$ (on either side), where k is a positive integer argument. Thus $k = 1$ demands that $word_1$ be adjacent to $word_2$.
Find the set of documents that satisfy the query Gates /2 Microsoft! Explain the algorithm.