Due: 2019/3/8

Homework 1

Student Number:
Name:

Problem 1. (20 points) The following pairs of words are stemmed to the same form by the Porter stemmer. Which pairs would you argue shouldnt be conflated. Give your reasoning.

a. abandon/abandonment

b. absorbency/absorbent

c. marketing/markets

d. university/universe

e. volume/volumes

Problem 2. (30 points)

Doc 1: new home sales top forecasts

Doc 2: home sales rise in july

Doc 3: increase in home sales in july

Doc 4: july new home sales rise

Consider the documents above,

a. Draw the term-document incidence matrix for this document collection.

b. Draw the inverted index representation for this collection.

c. For the document collection, what are the returned results for these queries:

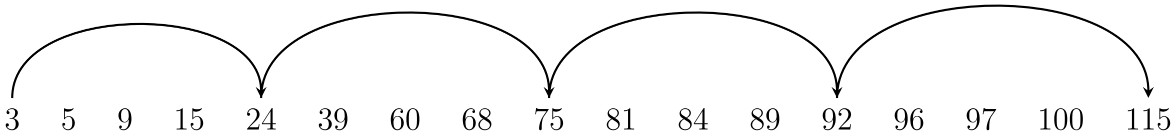
  i july AND rise
  ii (NOT increase) AND (home OR sale)

Problem 3. (30 points) Write out a postings merge algorithm, in the style of Algorithm 1, for an x OR y query.

Algorithm 1: INTERSECT(p_1, p_2)

```
$answer \leftarrow ()$
while $p_1 \neq NIL$ and $p_2 \neq NIL$ do
    if $docID(p_1) = docID(p_2)$ then
        ADD($answer$, $docID(p_1)$)
        $p_1 \leftarrow next(p_1)$ $p_2 \leftarrow next(p_2)$
    else
        if $docID(p_1) < docID(p_2)$ then
            $p_1 \leftarrow next(p_1)$
        else
            $p_2 \leftarrow next(p_2)$
return $answer$
```

Problem 4. (30 points) Consider a postings intersection between this postings list, with skip pointers:



3 5 9 15 24 39 60 68 75 81 84 89 92 96 97 100 115

and the following intermediate result postings list (which hence has no skip pointers):
3 5 89 95 97 99 100 101
Trace through the postings intersection algorithm(pdf of lecture 1, page 39)

a. How often is a skip pointer followed?

b. How many postings comparisons will be made by this algorithm while intersecting the two lists?

c. How many postings comparisons would be made if the postings lists are intersected without the use of skip pointers?