Due: 2018/3/9

Homework 1

Problem 1. (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)

Solution:

(*a*) Incidence Matrix

	Doc1	Doc 2	Doc3	Doc4
increase	0	0	1	0
forecast	1	0	0	0
top	1	0	0	0
in	0	1	1	0
rise	0	1	0	1
new	1	0	0	1
july	0	1	1	1
sales	1	1	1	1
home	1	1	1	1

(*b*) Inverted Index Representation

Terms	Posting List
forecasts	1
home	$1 \leq 2 \leq 3 \leq 4$
in	$2 \leq 8$
increase	3
july	$2 \leq 3 \leq 4$
new	$1 \leq 4$
rise	$2 \leq 4$
sales	$1 \leq 2 \leq 3 \leq 4$
top	1

(c)
(i) July AND Rise
$\boxed{2}$ $\boxed{4}$

DOC1: 0 DOC2: 1 DOC3: 0 DOC4:1
(ii) (NOT increase) AND (home OR sale)

$\boxed{1}$ $\boxed{2}$ $\boxed{4}$
DOC1: 1 DOC:1 DOC3:0 DOC4: 1

Problem 2. (20 points) The following pairs of words are stemmed to the same form by the Porter stemmer. Which pairs would you argue shouldn't be conflated. Give your reasoning.

a. abandon/abandonment

b. absorbency/absorbent

c. marketing/markets

d. university/universe

e. volume/volumes

Problem 3. ***<u>Solution:</u>***

a. abandon/abandonment
Conflate : Because of similar semantics

b. absorbency/absorbent
Conflate: Because of similar semantics

c. marketing/markets
Don't Conflate: Marketing is a subject (taught in business schools)
Market is different from marketing as it can be use for farmers market , stock market e.t.c

d. *university/universe*
Don't Conflate: University is a higher learning place for bachelors, masters and doctrate and post-doctrate and research learning place.
However Universe is meant as for the world and the galaxy and the planets in it in general.

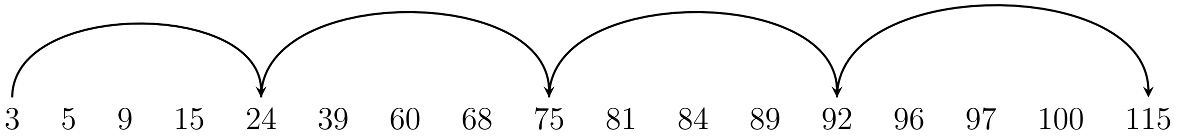
e. *volume/volumes*
Conflate: As they have same semantics in most contexts however in some cases it's different like degree of loudness

Problem 4. (20 points) Write a query using Westlaw syntax which would find any of the words professor, teacher, or lecturer in the same sentence as a form of the verb explain.

Solution:

$\boxed{QUERY : professor\ teacher\ lecturer\ /s\ explain!}$

Problem 5. (30 points) Consider a postings intersection between this postings list, with skip pointers:



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?

Solution:
(a) The skip pointer is followed once. from (24 to 75)
(b) 19 comparisons are made.
Let (x,y) denote a posting comparison.
The comparisons are : (3,3),(5,5),(9,89),(15,89),(24,89),
(75,89),(75,89), (92,89),(81,89),(84,89),(89,89),(92,95),
(115,95),(96,95),(96,97),(97,9),(100,99), (100,100),(115,101)
(c) 19