Homework
Information Retrieval

TA. Wang Haiwen, Qi Jiexing

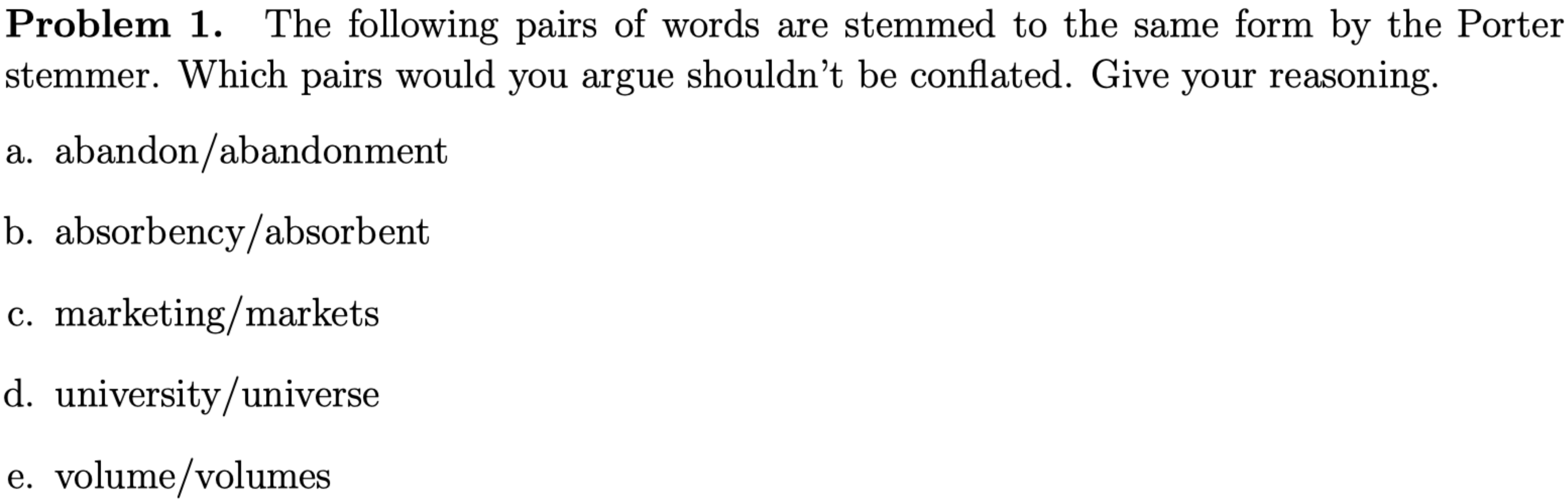
Dept. Computer Science & Engineering,

Shanghai Jiaotong University

Wang Haiwen, wanghaiwencn@foxmail.com

Qi Jiexing, qi_jiexing@qq.com

Part 1



Problem 1. 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

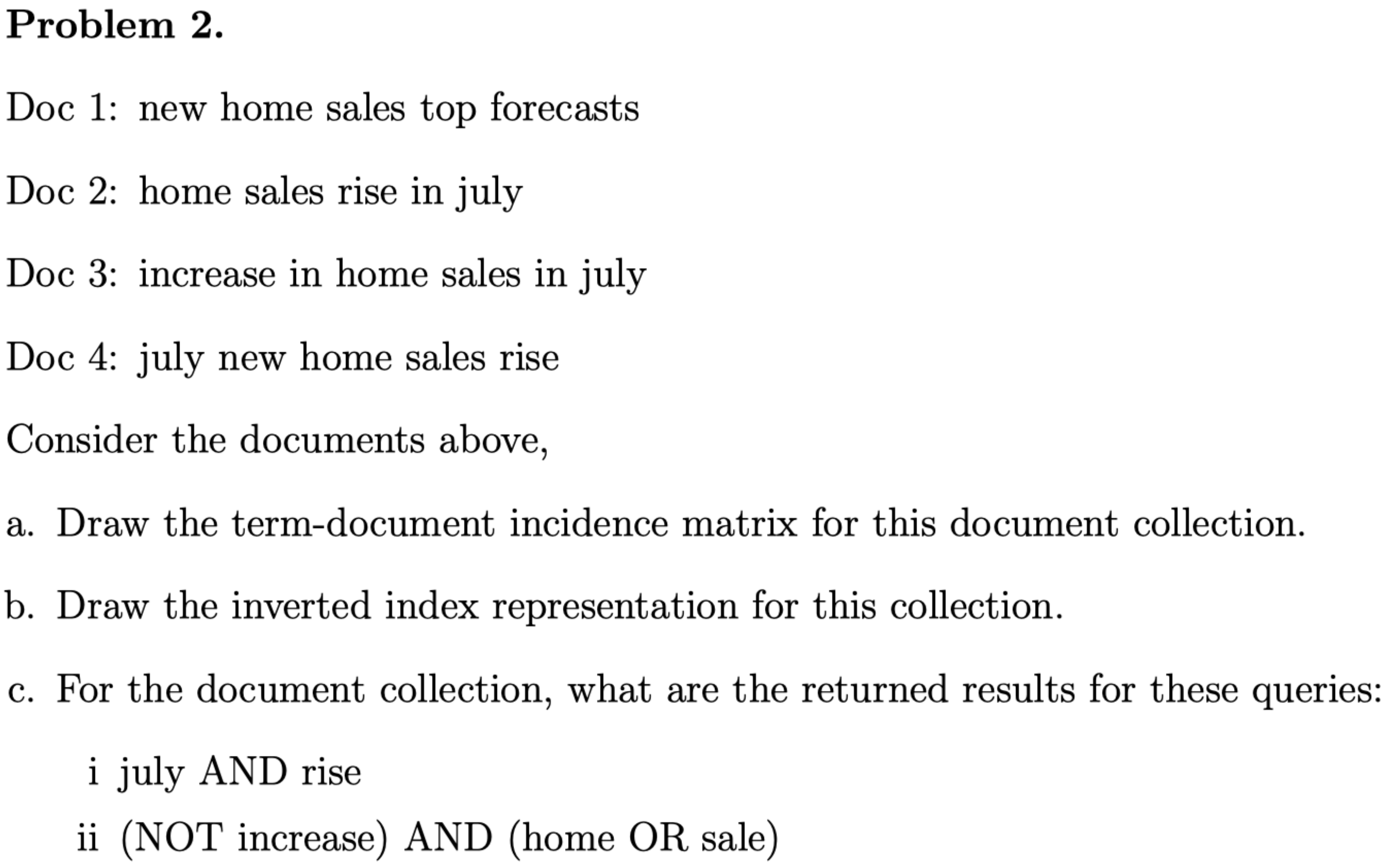
b. absorbency/absorbent

c. marketing/markets

d. university/universe

e. volume/volumes

Part 1



Problem 2.

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.

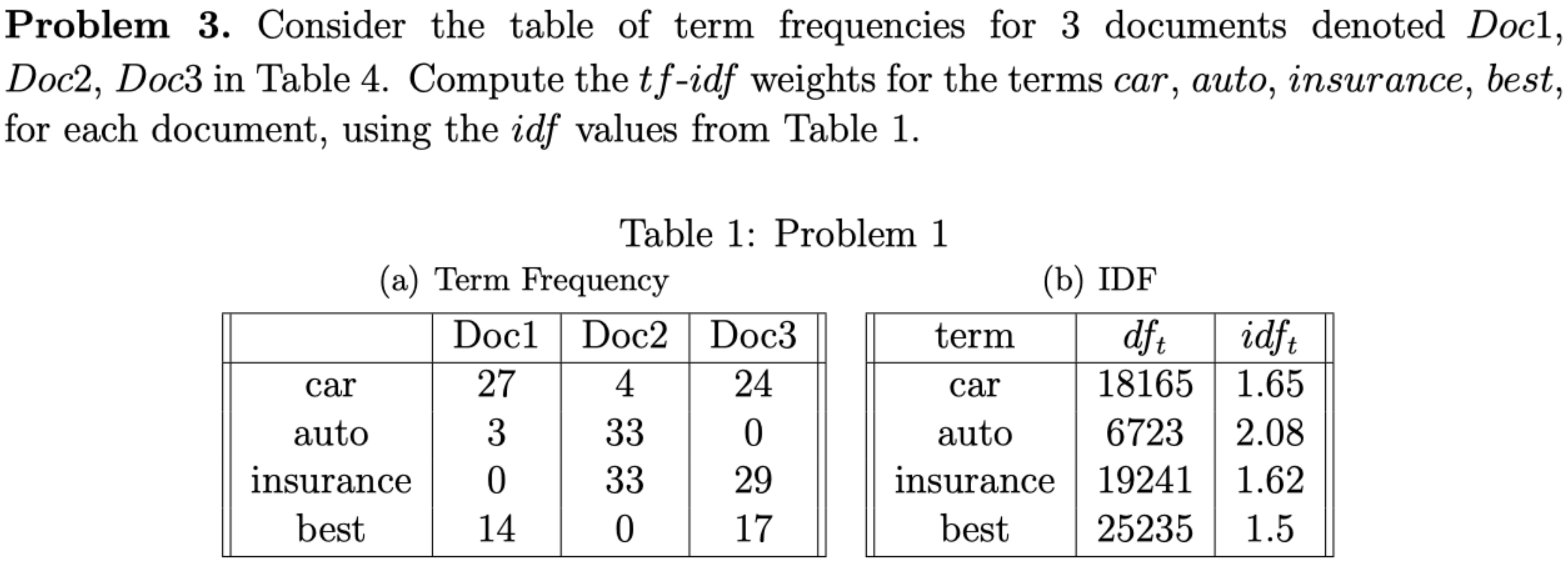
b. Draw the inverted index representation for this collection.

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

  i july AND rise

  ii (NOT increase) AND (home OR sale)

Part 1



Problem 3. Consider the table of term frequencies for 3 documents denoted *Doc*1, *Doc*2, *Doc*3 in Table 4. Compute the *tf*-*idf* weights for the terms *car*, *auto*, *insurance*, *best*, for each document, using the *idf* values from Table 1.

Table 1: Problem 1

(a) Term Frequency

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

(b) IDF

term	df_t	idf_t
car	18165	1.65
auto	6723	2.08
insurance	19241	1.62
best	25235	1.5

Part 2

给定数据：

- MySQL数据库连接方式：
- Host：server.acemap.cn
- Port：13306
- Username：ee101_web
- Password：ee101_web
- Charset：utf-8
- 其中包含1万篇论文的信息，论文id从0-9999

Introduction

- *Introduction* ：
  - 安装部署es或者solr（二选一）。
  - 分别将数据中需要建立索引的不同字段添加到不同的core中，建立索引（Solr）。
  - 配置合理的mapping，构建index，实现基本的文本检索功能。（ES）
  - 参考课程ppt。

Lab

- ***lab*：**
  - 面向Paper数据创建index。
    - 需要在report中描述具体多少条数据。
  - 配置mapping（包含Author, Keywords, Venue *etc.*）。
    - 需要设置合理的分词器配置、类型设置等
    - 需要记录在report中。
  - 进行数据检索实验
    - 需要在report中写返回结果。
    - 需要对返回结果评分与排序进行评价，是否实现了预期的功能

Task

- ***实现：***
  - 进行数据检索
  - 给定**论文名称**关键词，返回搜索结果
  - 给定**作者名称**关键词，返回搜索结果
  - 给定**会议**关键词，返回搜索结果
  - 给定**keywords**关键词，返回搜索结果

作业提交

- 提交方式:上传ftp
- 地址:ftp://public.sjtu.edu.cn
- 用户名:hnxxjyt 密码:public
- 打开方式：文件资源管理器/filezilla等软件
- 报告上传：/upload/文本信息检索
- 报告命名格式：HW2_组号_version
- 如： HW2_ 1_v1.pdf
- **截止日期：4.20 23:59！**