This docx will illustrate what implementation I have done to the project. They are: 1.probabilistic retrieval model, 2. Stemmed words reconvert

For group member draw evaluation figures:

All query result file and evaluation result file are under trec\_eval/test

Please find them and do comparison according to below illustration in I.4.result and II.3.result

Please add below to the correct section in the final report:

I.probabilistic retrieval model

**1. Intuition**

The intuition of the thought of probabilistic retrieval model comes from bayesian network. Both bayesian network and information retrieval are attempting to solve the “uncertainty” problem.

Information retrieval problem can be regard as a probabilistic problem. The relevance between document and query will lead to a certain probability that the document is retrieved by that query. So equivalent to cosine similarity in VSM, the probabilitic retrieval model aims to calculate :

the probability that certain document D is retrieved by a certain query Q,

which is also equivalent to:

the probability of (decision = retrieve it!) given the document D and query Q, that is:

P(R|(D,Q))

We can further expand it using bayesian conditional formula

Also for ~R:

Evaluation (ranking) formula (After we eliminate P(D|Q) because it hard to calculate)

For given query, is constant

Assume all words appear independently in a document (This assumption is from Internet)

So that the probability that document d exists is: all words in D appear and all words not in D disappear: (x represent words or stems)

After some simplification and prove:

where:

**2. Problems:**

<1>. It assume every words appear independently in the document, which is actually not the case

<2>. It need relevance feedback to adjust itself to a convergence. This lead to problems in RF and time complexity for convergence.

**3. A better model – BM25 (from internet)**

Best advantage: only need to change the similarity function in the VSM

=>

where:

k1: tunable parameter, usually 1.2

b: tunable parameter, usually 0.75

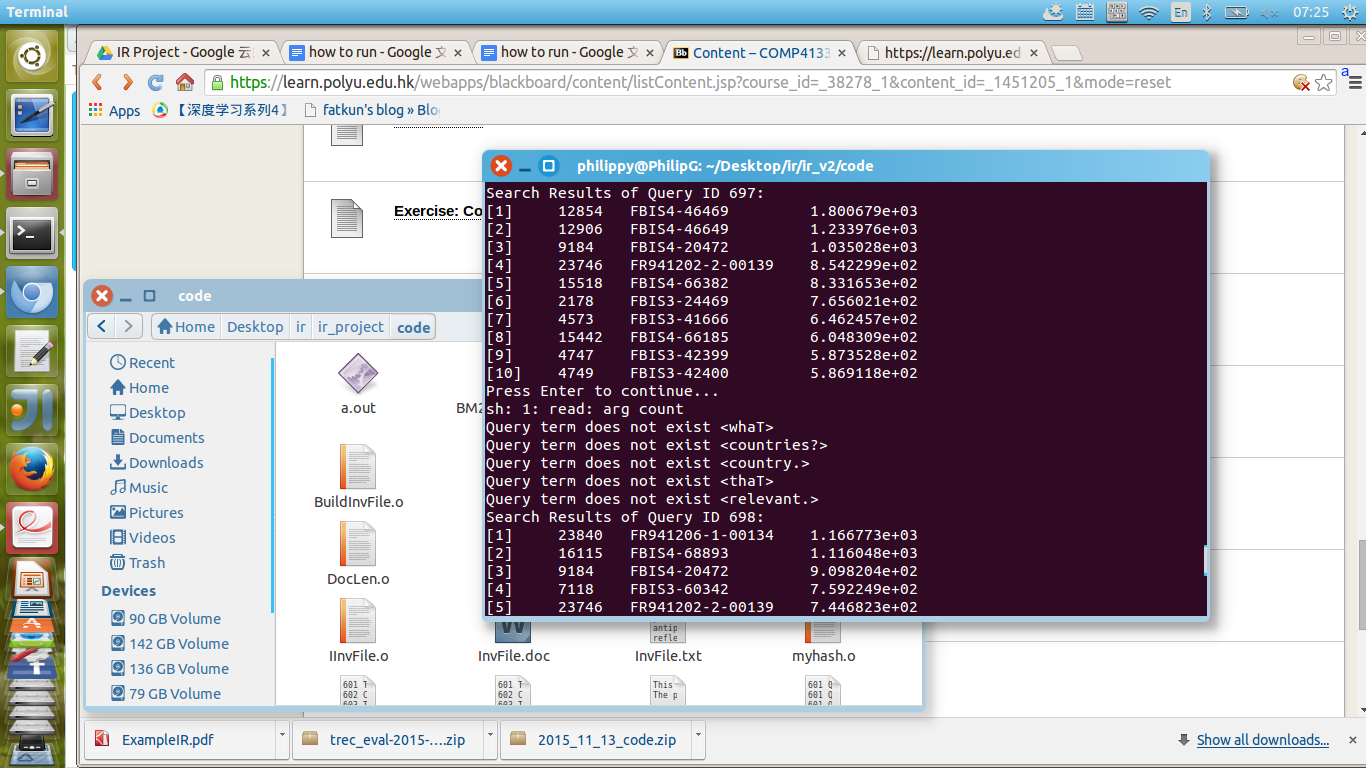
|D|: document length

Avgdl: average document length

II. stemmed word reconvert

**1. Problem**

After stemming the word: some problem still exists, most of which are related to punctuations:



As you can see, some words including punctuations like ? or . seriously influence the retrieval result.

**2. Solution**

If the stem can not be found in the document, remove the punctuation and find it again.

**III. Evaluation:**

**1. Interpolated precision-recall curve**

Analysis:

Overall speaking, BM25 model perform better than VSM model.

Reconvert of stemmed words will help the performance

**2. Precision on top N retrieved results**

Analysis:

For this analysis, we should focus on the top records (for example, top 5-top20) being retrieved, because irrelevant documents will be inevitably being retrieved at bottom.

This figure reflects whether user will easily get what they want at the top N records of retrieved results.

Result:

BM25 model perform better than VSM model.

Reconvert of stemmed words will help the performance

**3. Recall on top N retrieved results**

Analysis

For this analysis, we should focus on the figure on all records because it reflects whether all relevant records are exhausively being presented.

Result:

BM25 model perform better than VSM model.

Reconvert of stemmed words slightly help the performance