Task 1

1. Explain the difference between automatic local analysis and automatic global analysis.

To say that the local or global analysis is automatic means in this context that it's part of an implicit relevance feedback cycle. The feedback is derived from the system with no user participation.

Automatic local analysis means that we derive the feedback information from the top ranked documents in the result set.

If we use automatic global analysis we derive the feedback information from external sources such as a thesaurus or from term relations extracted from the document collection.

2. What is the purpose of relevance feedback? Explain the terms Query Expansion and Term Re-weighting. What separates the two?

The purpose of relevance feedback is to improve the initial query formulation and to help us retrieve documents that are likely to be considered relevant to the initial query.

Query Expansion:

Is to try to change (expand) the initial query to another query that will give better results.

Term re-weighting:

Is to analyze our global set of documents and try to find correlations with terms in our initial query that will improve our results. We generalize quite much as we don't look at document term weights or query term weights.

Task 2

1. Explain the language model, what are the weaknesses and strengths of this model?

The idea of the a language model is to use the text in the document to predict the probability of observing a given query.

Strengths with using Language models is that its conceptually simple and explanatory and that is has a natural use of collection statistics.

Weaknesses can be that the language models aren't accurate representations of the data or that the term distribution doesn't match the users idea of the term distribution.

2. Given the following documents and queries, build the language model according to the document collection.

For each query, rank the documents using the generated scores.

Assumption: Document has been preprocessed, which includes stop words being eliminated and terms being categorized

$$\hat{P}(t|M_d) = (1-\lambda)\hat{p}_{mle}(t|M_d) + \lambda\hat{p}_{mle}(t|_C), \lambda = 0.5. \qquad (1)$$

$t|M_d$ er hvor mange ganger termen i queryen dukker opp i dokumentet delt på hvor mange termer det er i dokumentet.

$t|_C$ er hvor mange ganger termen dukker opp i alle dokumentene i hele kolleksjonen delt på alle termer i alle dokumentene i kolleksjonen.

Original:

d1 = failure is the opportunity to begin again more intelligently.

d2 = intelligence is the ability to adapt to change.

d3 = lack of will power leads to more failure than lack of intelligence or ability

After preprocessing:

d1 = failure opportunity begin intelligently.

d2 = intelligence ability adapt change.

d3 = lack will power leads failure intelligence ability

term collection: {ability, adapt, change, failure, intelligence, lack, opportunity, power, will}

d1 word count: 4, d2 word count: 4, d3 word count: 7, total: 15, tc size: 9

q1 = failure

$P\hat{}(q1|d1) = 0.5 * (1/4) + 0.5 * (2/15) = 0.192$

$P\hat{}(q1|d2) = 0.5 * (0/31) + 0.5 * (2/15) = 0.067$

$P\hat{}(q1|d3) = 0.5 * (1/7) + 0.5 * (2/15) = 0.138$

Ranking:

1. d1
2. d3
3. d2

term collection: {ability, adapt, change, failure, intelligence, lack, opportunity, power, will}

d1 word count: 4, d2 word count: 4, d3 word count: 7, total: 15, tc size: 9

q2 = intelligence opportunity

$P\hat{}(q2|d1) = (0.5 * ((1/4) + (3/15))) * (0.5 * ((1/4) + (1/15))) = 0.0356$

$P\hat{}(q2|d2) = (0.5 * ((1/4) + (3/15))) * (0.5 * ((0/4) + (1/15))) = 0.0075$

$P\hat{}(q2|d3) = (0.5 * ((1/7) + (3/15))) * (0.5 * ((0/7) + (1/15))) = 0.0057$

Ranking:

1. d1
2. d2
3. d3

term collection: {ability, adapt, change, failure, intelligence, lack, opportunity, power, will}

d1 word count: 4, d2 word count: 4, d3 word count: 7, total: 15, tc size: 9

q3 = intelligence failure

$P\hat{}(q3|d1) = (0.5 * ((1/4) + (3/15))) * (0.5 * ((1/4) + (2/15))) = 0.0431$

$P\hat{}(q3|d2) = (0.5 * ((1/4) + (3/15))) * (0.5 * ((0/4) + (2/15))) = 0.0150$

$P\hat{}(q3|d3) = (0.5 * ((1/7) + (3/15))) * (0.5 * ((1/7) + (2/15))) = 0.0237$

Ranking:

1. d1
2. d3
3. d2

3. Explain what smoothing means and how it affects retrieval scores. Describe your answer using a query from the previous subtask.

Smoothing is a method to give some probability mass to zero probabilities in documents lacking terms in our queries. It affects retrieval score by increasing probability ranking of documents lacking some terms in the queries, but only by a small amount (comparatively to all terms in all documents). Its also going to increase probability ranking of documents containing any other number of terms by the same principle, smoothing out the extra probabilities over all documents and queries.

Task 3:

1. Explain the terms Precision and Recall, including their formulas. Describe how differently these metrics can evaluate the retrieval quality of an IR system.

Formula precision: $Precision = \frac{Relevant\ and\ Retrieved}{Retrieved}$

Precision is the ability to retrieve top-ranked documents that are mostly relevant.

It includes the percentage of documents that are relevant, probability that a selected document is relevant and how well the system is at filtering out non-relevant stuff

Formula recall: $Recall = \frac{Relevant\ and\ Retrieved}{Total\ Relevant}$

Recall is the ability of the search to find all of the relevant items in the entire document collection

It includes the percentage of all the relevant documents selected, probability that a given relevant document will be retrieved and how complete the results are.

2. Given the following set of relevant documents rel = {82, 21, 45, 271, 72, 300, 94, 56, 88, 150}, and the set of retrieved documents ret = {91, 21, 45, 56, 82, 221, 72, 215}, provide a table with the calculated precision and recall at each level.

d	Relevant	Precision	Recall
91			
21	REL	0.50	0.125
45	REL	0.67	0.250
56	REL	0.75	0.375
82	REL	0.80	0.500
221			
72	REL	0.71	0.625
215			

Task 4:

1. What is interpolated precision?

Interpolated precision at the j-th recall level is the maximum know precision among all recall levels above r_j.

2. Given the example in Task 3.2, find the interpolated precision and make a graph.

Here is the graph: