Question 1

(a) In a general IR system:

(i) Draw a basic structure with all fundamental components, including those inside of the IR black box.

(ii) Suppose we are using a vector space model, label in your structure diagram what each block is, as a function or data.

[10 marks]

(b) Assume a document collection with the following three documents:
D1: Several technologies are important for realizing the semantic web. (Length = 0.619)
D2: Technologies that are important for semantic web are those including RDF and OWL. (Length = 0.781)
D3: Owl is hatching eggs in its nestle. Owl has three pretty eggs. (Length = 0.661)
Assume a Boolean retrieval model and the query "**(semantic AND owl) OR (rdf AND NOT(owl))**". Derive the disjunctive normal form (DNF) $\vec{q}_{dnf}$ for the query. Which are the relevant documents for the query? Explain why.

[5 marks]

(c) Based on the same document collection in (b), assume a vector space model and rank the documents according to the query "**semantic owl**", based on the TF-IDF coefficient:

$$w_{t,q} = (0.5 + 0.5 \times \frac{tf_{t,q}}{max\{tf_{t',q} : t' \in q\}}) \times \log(N/df_t)$$

Use the cosine similarity and show your calculations. $log_{10}(3/1) = 0.477$ and $log_{10}(3/2) = 0.176$.

[10 marks]

Question 2

(a) Illustrate mathematically how the zero-probability problem occurs in the context of unigram Language Model (LM), and how to work around this problem using Jelinek-Mercer smoothing.

[5 marks]

(b) Suppose in a piece of Starwars movie transcript d, bigram probabilities for eight words are listed below. For instance, when the bigram is 'use lightsaber', t_k ='lightsaber', t_{k-1} ='use', then $p(t_k|t_{k-1}) = 0.27$. Within document d, $p_d('jedi') = 0.051$. Consider a query 'Jedi Yoda is Jedi master'.

t_{k-1} \ t_k	jedi	use	lightsaber	Yoda	is	master	force	hope
jedi	0	0.33	0.0011	0.26	0.52	0.12	0.0065	0.0029
use	0	0	0.27	0	0	0	0.17	0
lightsaber	0	0.0022	0	0.00092	0.38	0.00079	0	0.0058
Yoda	0.43	0.29	0.37	0.014	0.58	0.35	0.021	0.065
is	0.67	0	0.15	0.12	0	0.14	0.089	0.36
master	0.22	0.18	0.069	0.66	0.39	0	0.14	0.002
force	0	0.0065	0	0.095	0.44	0.11	0	0.056
hope	0	0	0	0	0.18	0	0	0
Global $P(t_k)$	0.022	0.092	0.055	0.032	0.19	0.0015	0.011	0.0043

(i) What is the probability of this document being relevant to the query in the bigram model without smoothing?

(ii) What is the probability of this document being relevant to the query in the bigram model with Jelinek-Mercer smoothing given $\lambda = 0.6$? No need to calculate the final answer for these questions.

[10 marks]

(c) In a local network of six webpages $\{A, B, C, D, E, F\}$, we have the following Pagerank:
$PR(A) = (1-d) + d(\frac{PR(B)}{N(B)} + \frac{PR(C)}{N(C)})$; $PR(B) = (1-d) + d(\frac{PR(A)}{N(A)} + \frac{PR(D)}{N(D)})$;
$PR(C) = (1-d) + d(\frac{PR(B)}{N(B)})$; $PR(D) = (1-d) + d(\frac{PR(E)}{N(E)})$;
$PR(E) = (1-d) + d(\frac{PR(B)}{N(B)} + \frac{PR(D)}{N(D)})$; $PR(F) = (1-d) + d(\frac{PR(C)}{N(C)})$.

(i) Provide a diagram to illustrate the network structure.

(ii) Set $d = 0.5$. Complete the Pagerank functions with the correct $N(\cdot)$ values.

(iii) Use any approach to calculate the $PR(\cdot)$ values for the six pages. Two decimal places are adequate.

[10 marks]

Question 3

(a) In the following table, if a term appears in a document, it is represented by a 1; otherwise, 0.

Documents	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
D_0	1	0	0	1	1	0	0	0	0	0
D_1	1	1	0	1	0	0	1	1	1	0
D_2	0	0	0	0	1	1	0	1	0	1
D_3	1	0	1	0	1	1	1	0	0	1
D_4	0	1	1	0	0	1	0	0	0	0
D_5	0	1	0	0	0	0	0	1	0	1
D_6	1	1	0	1	0	0	1	1	0	0
D_7	0	1	1	0	1	0	1	0	0	1
D_8	1	0	0	1	1	0	0	1	0	0
D_9	1	0	1	1	0	1	0	0	0	1

A term-to-term occurrence matrix is obtained from the document collection:

$C_{i,j}$	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
t_0	1	0.35	0	0.54	0	0	0.17	0.08	0	0.26
t_1	0.35	1	0.72	0	0	0.06	0	0	0.88	0
t_2	0	0.72	1	0	0.33	0	0	0	0.38	0.18
t_3	0.54	0	0	1	0.21	0.28	0	0	0	0.49
t_4	0	0	0.33	0.21	1	0.81	0	0.56	0	0
t_5	0	0.06	0	0.28	0.81	1	0.48	0	0	0.20
t_6	0.17	0	0	0	0	0.48	1	0.82	0.17	0
t_7	0.08	0	0	0	0.56	0	0.82	1	0	0
t_8	0	0.88	0.38	0	0	0	0.17	0	1	0.16
t_9	0.26	0	0.18	0.49	0	0.20	0	0	0.16	1

The sequence of values $\{i, j, k, l, m, n, p, q, s, t\}$ each corresponds to a digit in your student number, which consists of 10 digits and a slash sign '/'. The '/' itself is ignored in this context. For EXAMPLE, if your student number is 202112345/1, then $i = 2, j = 0, k = 2, l = 1, m = 1, n = 2, p = 3, q = 4, s = 5, t = 1$. If you do not know the value of t, you can use $t = n$.

Based on the given data, calculate the relevance scores based on the fuzzy set model: $\mu_i(D_j)$, $\mu_k(D_l)$, $\mu_m(D_n)$, $\mu_p(D_q)$ and $\mu_s(D_t)$.

[15 marks]

(b) Consider the same Boolean query in Question 1.(b): "**(semantic AND owl) OR (rdf AND NOT(owl))**", suppose w_s, w_o and w_r are the weights for "semantic", "owl" and "rdf", respectively.

(i) Write down the similarity function $R(d, q)$ with 2-norm.

(ii) Calculate the similarity values for cases of ($w_s = 1, w_o = 1, w_r = 0$), ($w_s = 1, w_o = 0, w_r = 1$) and ($w_s = 1, w_o = 0, w_r = 0$)

[10 marks]

Question 4

(a) Given an indexed collection of documents and a query, your system retrieves a set of documents ranked as shown in the table below. For this particular query, the set of relevant documents is $\{$d1, d3, d4, d10$\}$.

Rank	Doc	Relevant/Non-relevant	Precision	Recall
1	d5			
2	d4			
3	d1			
4	d7			
5	d20			
6	d3			
7	d8			
8	d9			
9	d10			
10	d11			

(i) Fill in the table with correct values (You can use fractions or decimals for the values).

(ii) Draw an interpolated precision-recall curve for this result.

[5 marks]

(b) Based on the question above, answer the following:

(i) What is the R-precision? Explain your answer.

(ii) What is the Mean average precision (we only have a single query here). Show your steps.

(iii) What is the E measure of this retrieved list, with $b = 0.3$? Show your steps. Is this E measure in favour of precision or recall?

[10 marks]

(c) Suppose we are conducting relevance feedback with the Rocchio algorithm, with $\alpha = 1.0$, $\beta = 0.5$ and $\gamma = 0.5$. In one feedback iteration, we receive one positive document $\vec{d}_{pos} = \{3, 1, 0, 0, 5, 0, 2, 6, 3, 0\}$ and one negative document $\vec{d}_{neg} = \{0, 2, 4, 3, 0, 0, 4, 6, 0, 3\}$. The resulting new query vector contains 10 elements, each corresponds to a digit in your student number, which consists of 10 digits and a slash sign '$\backslash$'. The '$\backslash$' itself is ignored in this context. For example, if your student number is 202112345/1, the new query vector $\vec{q}_{new} = \{2, 0, 2, 1, 1, 2, 3, 4, 5, 1\}$. Please write down the Rocchio algorithm and deduce the original query vector.

[10 marks]

End of questions