Q1. Document vectorization – M1

A. A user is searching for all documents that contain the term ***petroleum***. Since the terms like ***oil, gas, crude, barrels, gasoline*** etc. are also considered relevant you are asked to add an extra feature to the search engine that can handle the task. **[3+2=5 M]**

i. Briefly explain the two main approaches we discussed in this module to implement this feature.

ii. State the pros and cons of each approach.

B. You are asked to design an IR system that allows wildcard characters in between a word. Name the appropriate index(s) that can handle such a query. **[1 M]**

Q2. Part of speech tagging using Hidden Markov Model (HMM) – M2

The following three sentences (S1, S2 and S3) and their corresponding tag sequences (T1, T2 and T3) are given as training data for implementing HMM. **Answer questions A-D. [2X4=8M]**

S1: John cut the paper .	S2: Mary asked for a hair cut .	S3: Sharon asked for a pay cut .
T1: N V D N STOP	T2: N V I P N N STOP	T3: N V I P N N STOP

A. What will be size of Emission and Tag translation matrices?

Note: Include the start and the stop symbols and assume we are working with a bigram model.

B. What will be the Emission Probability P(asked | V) ?

C. What will be the Tag Translation Probability P(I|V) ?

D. If a brute force approach is employed to find the tags for the test sentence "***I had a deep cut.***", how many possible tag sequences need to be evaluated?

Q3. Topic modelling using Latent Dirichlet Allocation(LDA) – M3 [4+2=6M]

A. Explain the mathematical formulation of LDA and each term and its importance in the overall model building.

B. What is the use of conjugate prior in LDA? Name and explain the conjugate prior used in LDA?

Q4. Sentiment Analysis – M4

You are given a list of positive and negative sentiment words as lexicons and expected to design a sentiment analyzer, explain in steps how would you proceed. **[4 M]**

Q5. Recommender Systems – M5

A. Given **Table-2** consisting of user-items ratings, compute the rating R_{31} using a base line approach. What characteristics of the data does it capture and why is it combined with other approaches? **[3 M]**

Item-id/ Userid	**I_1**	**I_2**	**I_3**	**I_4**	**I_5**	**I_6**
U_1	8	2	5	7	1	8
U_2	1	?	4	1	7	3
U_3	ⓒ	5	7	7	4	8
U_4	2	7	4	3	6	?
U_5	7	5	7	8	5	7

Table-2

B. You are given a utility matrix of size 4000X50000 and when Singular Value Decomposition is performed what are matrices it returns and their sizes? Explain the process you would follow if you are asked to reduce it to 5 genres. **[3 M]**