Original Scores

A mean rank of **5.12** and accuracy of 0.3125

Question 1

For this question, the following processes were implemented to reduce the mean rank:

1. **Removing extra whitespaces:** Extra whitespaces increase the text size and are meaningless. Removed the whitespaces in the text. This did not improve mean rank.
2. **Tokenisation:** Split the text into tokens in to perform NLP operations upon the data that is given.
3. **Clean English:** For better parsing of text, the non-alphabetical characters were removed such as numbers and symbols.
4. **Filter stop words:** Insignificant stop words like 'the', 'an' etc were removed. This is to better perform NLP operations. This greatly improves the mean rank.
5. **Stemming:** This reduces words to their root form. This greatly improved mean rank which is why this was chosen over lemmatisation as it produced slightly better results.

Other processes were implemented but they were removed as they increased mean rank.
They are as follows:

1. **Spell correction:** Tried using spell check to fix the spelling mistakes if they exist in the text. This was done to get better results from the model, but this made the mean rank worse, and it slowed increased pre-processing time a lot.
2. **Lowercasing:** Tried lowercasing the text so that the words are treated the same and this can prevent duplication of words however this made the mean rank worse, so it was abandoned.
3. **Lemmatisation:** This removes affixes from words to convert them to their base form and is good for creating better features. However, either Lemmatisation or Stemming could be chosen so Stemming was chosen as it gave better mean rank.

Mean rank 1.75
12 correct out of 16 / accuracy: 0.75

Question 2

The following feature extraction methods were used to improve mean rank:

1. **POS-Tagging:** Words are grouped into different categories of words. The following tags were used (some tags were removed to improve mean rank):
    a. RB (adverb) -> removed,
    b. NN (Noun, singular),
    c. VB (Verb, base form),
    d. JJ (Adjective),
    e. JJR (Adjective, Comparative) -> removed,
    f. JJS (Adjective, Superlative) -> removed,
    g. CC (Coordinating conjunction),
    h. PRP (Personal Pronoun)

2. **N-grams model:** Used the bigram model to approximate probabilities preceding word. Tried the trigram model too, but the bigram gave improved the mean rank whereas the trigram, 4-gram, 5- gram, 6-gram all made the mean rank worse.

Mean rank 1.625
13 correct out of 16 / accuracy: 0.8125

Question 3

Used Line spoken by characters and scene information to add context dialogue and name was used for tracking. The name was necessary to otherwise it does not track. The mean rank stayed the same and so did the accuracy. There seems to be no change whatsoever. Tried adding Scene and Episode, however that made the mean rank worse, and it was abandoned.

Mean rank 1.625
13 correct out of 16 / accuracy: 0.8125

Question 4

TF-IDF (Term Frequency – Inverse Document Frequency) vectorisation was used as an improved vectorisation method which gives numerical representation of words .

tf (t,d) = |no. of times term t appear in document d|

idf (t, D) = |no. of documents| / |no. of documents that contain term t|

The mean score and accuracy was perfect.

Mean rank 1.0
16 correct out of 16 / accuracy: 1.0

Question 5

The comparison between Dictionary vectorisation and TF-IDF vectorisation is given.

Table 1 - Comparison of Vectorisation Methods

	Dictionary	**TF-IDF**
Mean Rank	5.375	1.25
Accuracy	0.375	0.875

The TF-IDF gives very high accuracy and good mean rank whereas Dictionary vectorisation gives bad mean rank and bad accuracy after applying all the methods and techniques.

Overall

Table 2 - Overview of Metrics

	Mean Rank	**Accuracy**
Original	5.12	0.3125
Question 1	1.75	0.75
Question 2	1.625	0.8125
Question 3	1.625	0.8125
Question 4	1.0	1.0
Question 5	1.25	0.875