System Evaluation

We used three evaluation metrics, viz. Precision@K, Recall@K and Mean Average Precision. We believe that these metrics define the performance of the system quite effectively.

Average Precision
This assesses how much relevant between the user's queries and information and symptoms retrieved by the system in a higher ranking in the ranked list

$$P(rank(j)) = \frac{\#\text{ Relevant Symptoms till rank(j)}}{\#\text{ Symptoms till rank(j)}}$$

$$AP = \frac{\sum_{j=1}^{K} P(rank(j))}{K}$$

A higher AP implies that a greater number of relevant symptoms are being assigned higher ranks in the ranked list.

Precision@K
This defines the precision of the search results at a certain position K in the query results. It is defined as the ratio of the number of retrieved items which are relevant at that position to the total items retrieved at position K.

Recall@K
This is computed as the ratio of the number of relevant symptoms in the top K retrieved list to the total number of all relevant symptoms. It is given as follows:

$$Recall@K = \frac{\#\text{ Relevant items till rank(K)}}{\text{Total Number of Relevant items}}$$

Mean Average Precision (MAP)

Mean Average Precision for a set of queries is the mean of the average precision scores for each query. It is given as follows:

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$$

We used 5 common queries for both symptom search and disease search and calculated all the 3 evaluation metrics. We calculated the precision and recall values for the 8th, 9th and 10th ranked query results as they give a more correct understanding of the performance of the model. Moreover, the average values of these give a better sense of how the model is performing.

Precision@K:

Query	P@8	P@9	P@10
"Asthma"	4/8	5/9	6/10
"Diabetes"	5/8	6/9	7/10
"Flu"	5/8	5/9	6/10
"Heart Disease"	4/8	4/9	5/10
"Cancer"	8/8	9/9	10/10
Average	**0.65**	**0.6444**	**0.68**

Query	P@8	P@9	P@10
"Nausea"	7/8	8/9	9/10
"Seizure"	6/8	6/9	7/10
"Memory loss"	6/8	7/9	7/10
"Cough"	4/8	4/9	5/10
"Abdominal pain"	6/8	6/9	7/10
Average	**0.725**	**0.6888**	**0.7**

Recall@K:

Query	P@7	P@8	P@9
“Asthma”	4/6	4/6	5/6
“Diabetes”	4/6	5/7	6/7
“Flu”	4/6	5/6	5/6
“Heart Disease”	3/6	4/5	4/5
“Cancer”	7/10	8/10	9/10
Average	**0.64**	**0.762**	**0.844**

Query	P@7	P@8	P@9
“Nausea”	6/9	7/9	8/9
“Seizure”	6/7	6/7	6/7
“Memory loss”	5/7	6/7	7/7
“Cough”	4/5	4/5	4/5
“Abdominal pain”	5/7	6/7	6/7
Average	**0.750**	**0.829**	**0.880**

Mean Average Precision:

Query	Calculation	Mean Average Precision
“Asthma”	1/6(1 + 2/3 + 3/4 + 4/5 + 5/9 + 6/10)	0.728
“Diabetes”	1/7(1/2 + 2/3 + 3/5 + 4/6 + 5/8 + 6/9 + 7/10)	0.632
“Flu”	1/6(1/2 + 2/3 + 3/5 + 4/7 + 5/8 + 6/10)	0.593
“Heart Disease”	1/5(1 + 2/4 + 3/6 + 4/8 + 4/9 + 5/10)	0.688
“Cancer”	1	1
Average	**=>**	**0.728**

Query	Calculation	Mean Average Precision
“Nausea”	1/9(1/2 + 2/3 + 3/4 + 4/5 + 5/6 + 6/7 + 7/8 + 8/9 + 9/10)	0.785
“Seizure”	1/7(1 + 2/3 + 3/4 + 4/5 + 5/6 + 6/7 + 7/10)	0.801
“Memory loss”	1/7(1/2 + 2/3 + 3/4 + 4/6 + 5/7 + 6/8 + 7/9)	0.689
“Cough”	1/5(1/2 + 2/5 + 3/6 + 4/7 + 5/10)	0.494
“Abdominal pain”	1/7(1/2 + 2/3 + 3/4 + 4/6 + 5/7 + 6/8 + 7/10)	0.678
Average	**=>**	**0.689**

As the precision and recall provide a more specific score for the search results based on the rank, we used the MAP as the third evaluation metric. The MAP considers precision values for all the positions with relevant results and calculates a mean. This gives a more generic score for the entire set of search results rather than a specific position.