1. Short Answers

    a. Vector space models offer weight to vectors based on term frequency, although statistical language models are able to predict and relate words in a query to documents. For example, in statistical language models, the algorithm can predict how likely a user would use a term since it knows the topics the user is already interested in, based on the statistics of previous queries and document selections.

    b. An information need is what the user is interested in learning from a search, while the query is the exact string that is used when searching.

        Information need: “how do I use the python len() function”
        query: “python len function”

    c. Pseudo relevance is taking the top K ranked documents and using them to redefine the user’s initial query. The IR system assumes that the top ranked documents are relevant, so it uses the terms from the top documents to expand the query, this is called query expansion. The query is then re-weighted, containing the initial and new terms.

    d. A user gives more attention to documents in higher positions, and thus these documents receive more clicks. Position-bias of user clicks is referring to this phenomenon. Even if the documents were ranked in reverse order, with irrelevant documents on top, they would still receive the most clicks because they have the higher position.

2. Problem Solving

    a. Precision = TP / (TP + FP) = 5/10 = .5

    b. Recall = TP / (TP + FN) = 5/(5+25) = .167

    c. F-measure = 2 / (1/P + 1/R) = 2/(1/.5 + 1/.167) = .25037

    d. In this scenario:

        Precision = 5/100 = .05
        Recall = 5/30 = .167

        The precision would be the only metric effected. The recall calculation does not rely on false-positive documents. By returning all 100 documents, the false-positive rating goes up significantly.

3. Problem solving

   a. $IDCG = (2^3 - 1)/(\log_2 2) + (2^3 - 1)/(\log_2 3) + (2^2 - 1)/(\log_2 4) + (2^2 - 1)/(\log_2 5) + (2^1 - 1)/(\log_2 6)$
   $IDCG = 14.595$

   $DCG_1 = (2^3 - 1)/(\log_2 2) + (2^2 - 1)/(\log_2 3) + (2^2 - 1)/(\log_2 4) + (2^1 - 1)/(\log_2 6) + (2^3 - 1)/(\log_2 7)$
   $DCG_1 = 13.273$

   $DCG_2 = (2^3 - 1)/(\log_2 2) + (2^3 - 1)/(\log_2 3) + (2^2 - 1)/(\log_2 4) + (2^2 - 1)/(\log_2 5) + (2^1 - 1)/(\log_2 9)$
   $DCG_2 = 14.524$

   $NDCG_1 = 13.273 / 14.595 = .909$
   $NDCG_2 = 14.524 / 14.595 = .995$

   b.

      i. P@3 System 1 = 3/3
      P@3 System 2 = 3/3

      One system is not better than the other based on precision at K=3.

      ii. P@5 System 1 = 4/5
      P@5 System 2 = 4/5

      Again, one system is not better than the other at K=5.

      iii. MAP System 1 = (1/1 + 2/2 + 3/3 + 4/5 + 5/6) / 5 = .9267
      MAP System 2 = (1/1 + 2/2 + 3/3 + 4/4 + 5/8) / 5 = .925

      System 1 is more precise according to the MAP metric. This is because the relevant documents are ordered higher.

   c. Intuitively, System 1 is better for web search. Thinking about how a user focuses their attention on the documents higher on the page, you want to maximize the relevance in the highest documents on a page. System 1 does this better than system 2, as in system 2 a relevant document would be low on a page.