Short-Answer Questions [25 points]:

1. Investigating one search engine (Google or Bing), and figure out what tricks they are using. (15%)

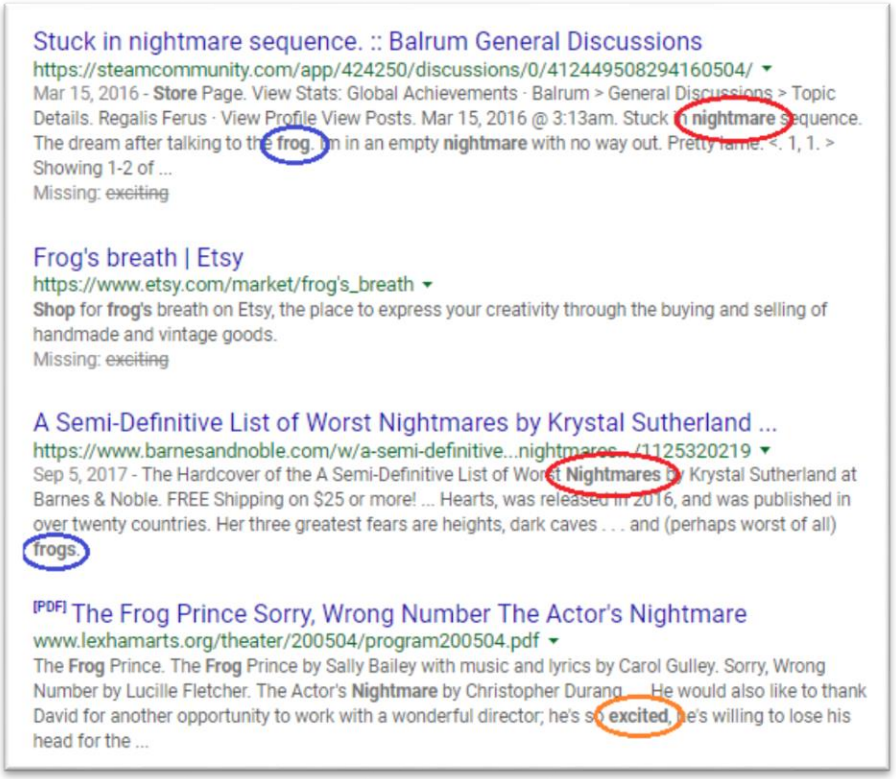
(a) Does it use stemming? How do you know? (Explain your testing method in all the following sub-questions as well)

(b) Does it filter stop words?

(c) Does it use spell correction?

[ANS]

I investigated Google search engine for whether it uses stemming, filters stop words or uses spell correction. The sentence used for examination is "*in a exciting store at Frogs of the nightmares*", a part of the result is as below:

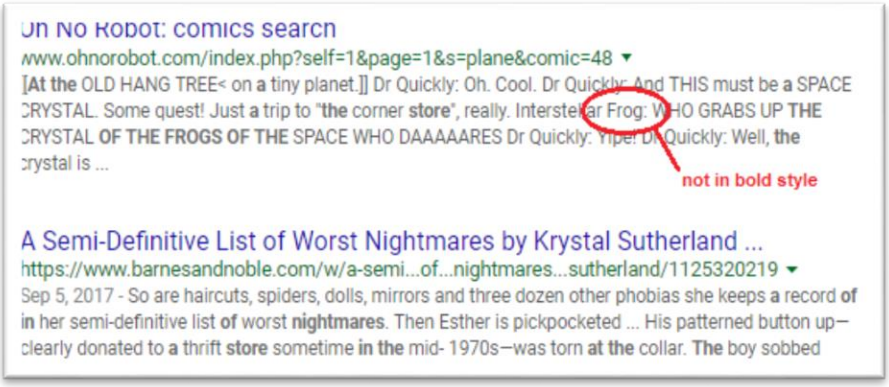


First, Google uses stemming, which is an approximate method for grouping words with a similar basic meaning together. For example, the word set "frog" and "frogs" generally have the same meaning, and so as the word set "nightmare" and "nightmares". These words are all marked in bold style by Google search engine in the result. Obviously, it is an evidence that Google is using stemming method for querying.

Second, Google has filtered stop words. In my query, stop words like “in”, “at”, “of” and “the” is included. While in the returning result, none of these words are marked in bold style even some of them appear in result documents. As a result, we believe that google has filtered these words during querying.

Third, Google is using spell correction for querying, yet it provides option for user to not to use spell correction. For instance, I made a typo “nightmars”. In response, google corrected it for me, returned result documents base on the correct sentence and provided me options if I want to query without auto correction.

On the other hand, Google search engine has an option called “Verbatim”, which removes the functions we discussed above. According to Google Introduces Verbatim Searching written by *Rob D. Young*, while using “Verbatim” option, the search engine will only use the same tense, the same verb form and the same plural/singular form for querying. As a result, only “frogs” will be targeted, only “nightmares” will be targeted, and no defined stop words anymore.



2. For a particular search query, your IR system returns 14 relevant documents and 16 irrelevant documents. There are a total of 80 relevant documents in the collection. (10%)

(a) What is the precision of the system on this search?

[ANS] By definition, precision is the fraction of relevant documents among the result documents. Therefore, the precision is 14 / (14+16) = 46.7%

(b) What is the recall of the system on this search?

[ANS] By definition, recall is the fraction of relevant document that have been retrieved over the total amount of relevant documents. Therefore, the recall is 14 / 80 = 17.5%