Short-Answer Questions

# Question 1

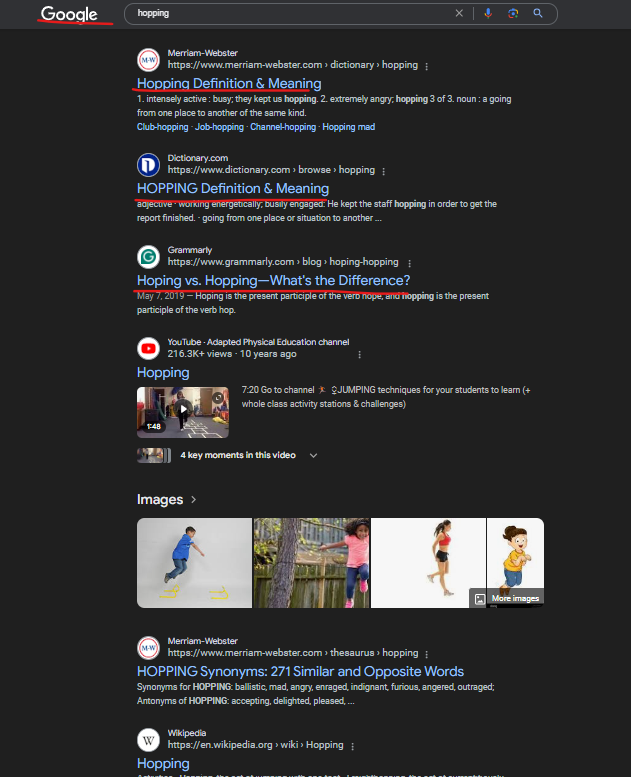
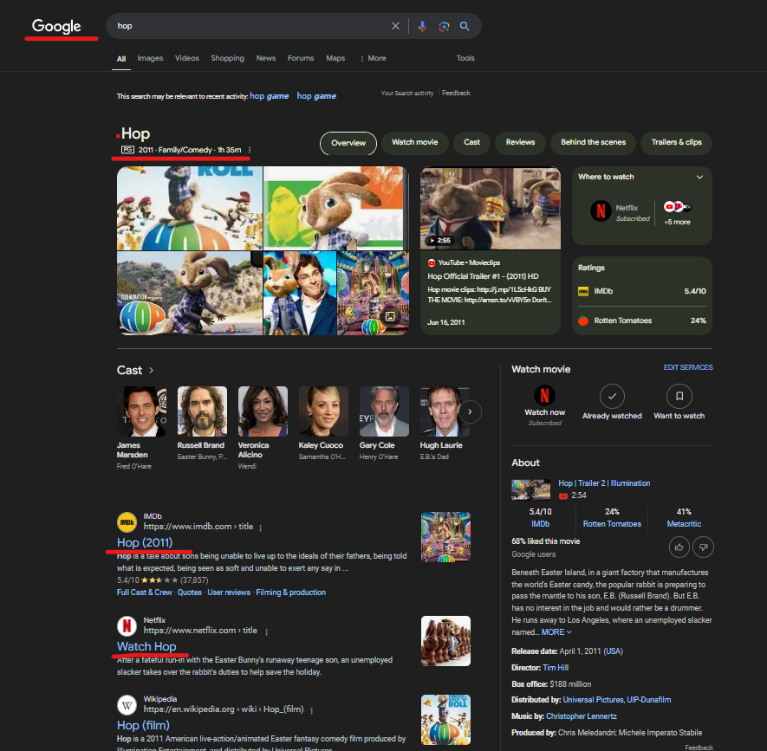
### Investigate two search engines (e.g., Google and Bing), and figure out what tricks they are using by explaining your testing method. (15%)

#### Do both search engines use stemming?

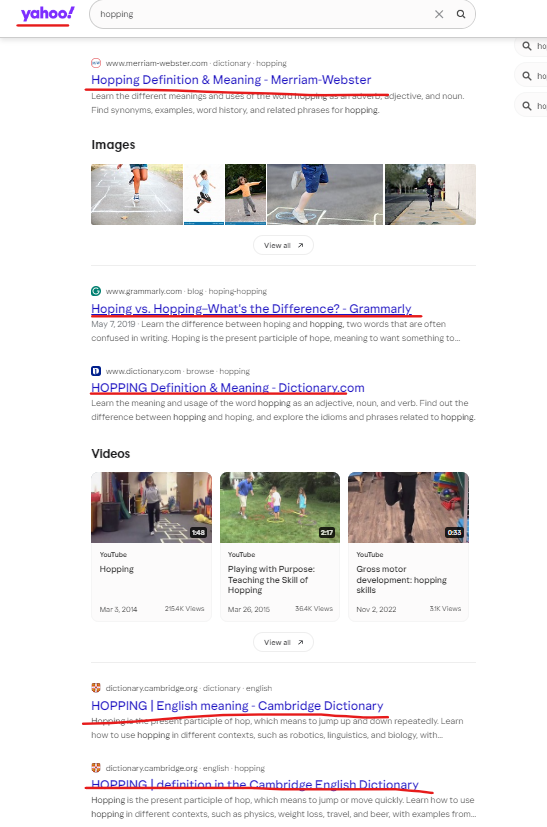
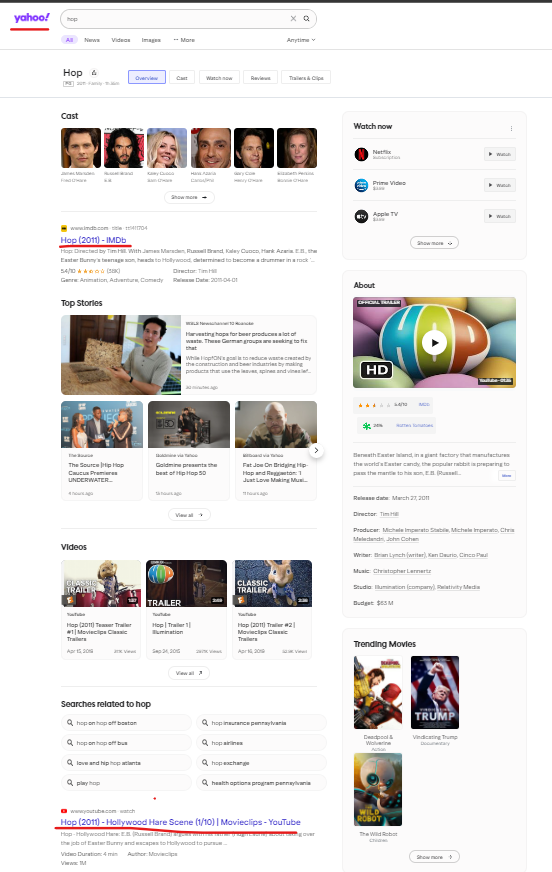
**Stemming** is a process used by search engines to reduce words to their base or "root" form so that variations of the word are treated the same. For instance, words like “jumpping,” “jumps,” and “jumper” can all be reduced to “jump.” This helps the search engine return more comprehensive results by treating similar words as equivalents.

**Testing Method**:

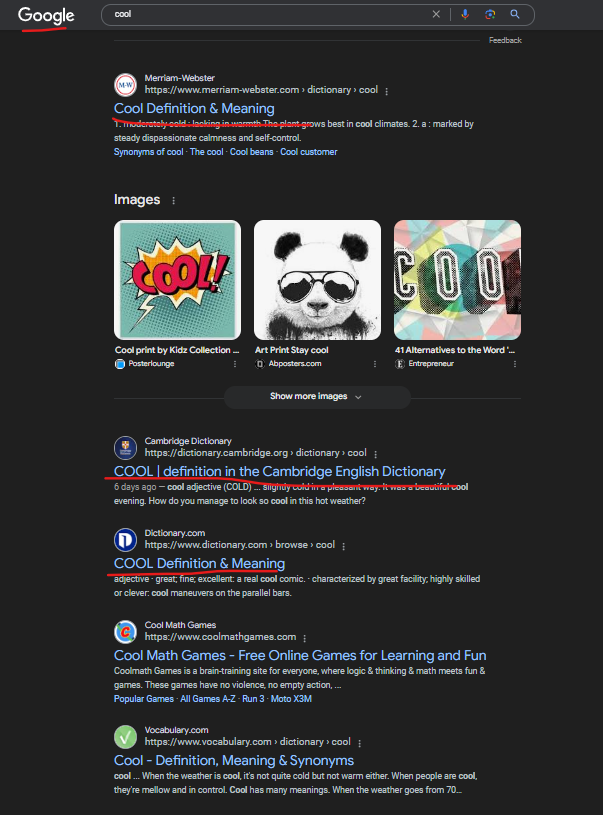
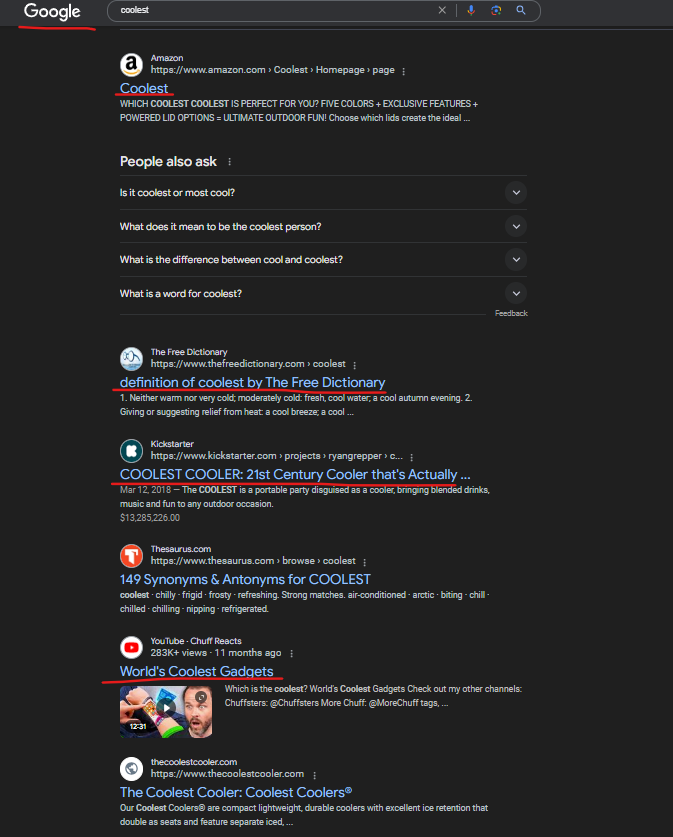
1. By performing searches on both **Google** and **Yahoo** using variations of a word I got below results:
   * 1. Hop, hopping, hopper
        1. Google-



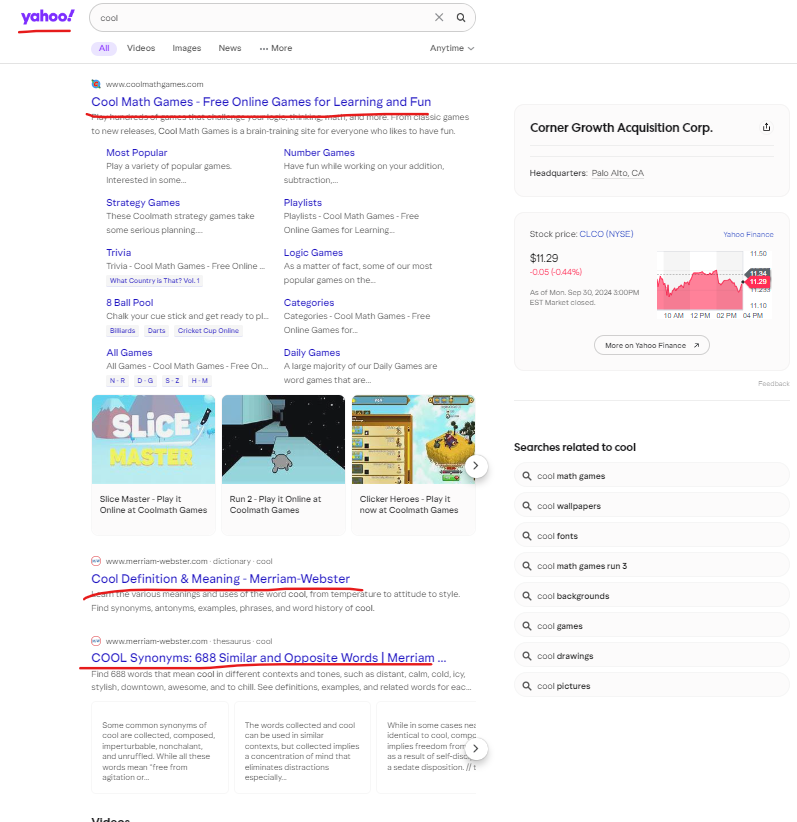
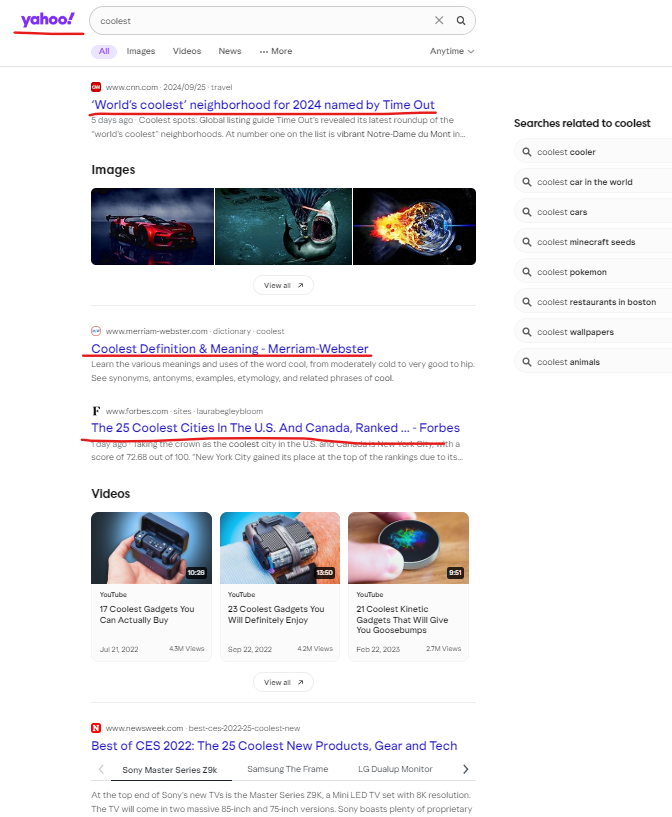
* + 1. Yahoo -



* 1. Cool, cooler, coolest
     1. Google –



* + 1. Yahoo –



1. The goal was to see if both engines returned similar results regardless of the word variation used, which would indicate they were using stemming or a similar technique.

**Findings**:

* Both **Google** and **Yahoo** seem to use a more advanced form of word reduction called **lemmatization** rather than basic stemming. Lemmatization goes beyond simply chopping off endings; it looks at the context and meaning of a word.
* For example, searching for "cool" and "coolest" may not yield identical results but will still return relevant content because the search engine understands that "cool" and "coolest" share the same root meaning, rather than just chopping off suffixes mechanically like stemming does.

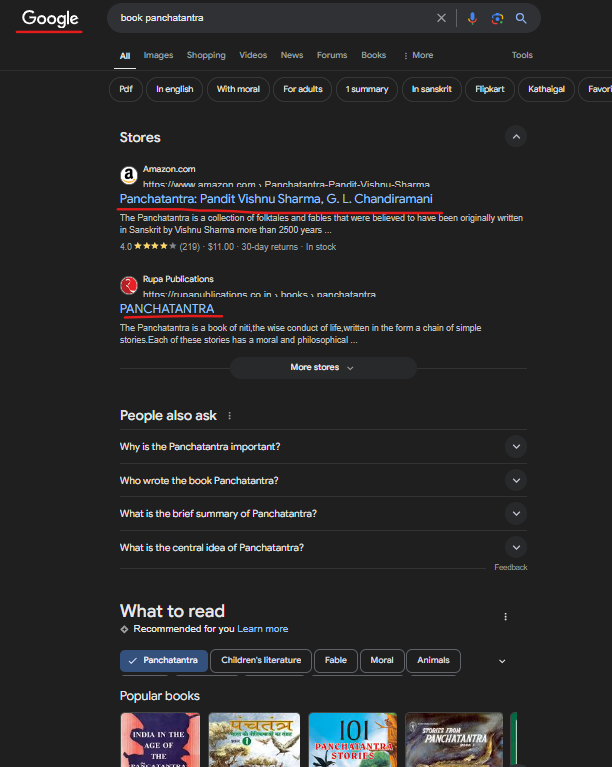
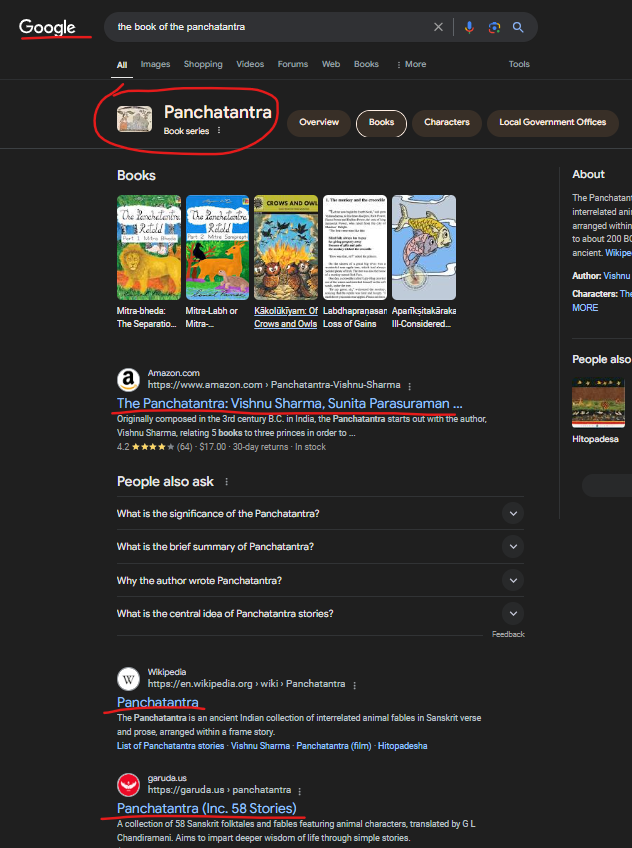
**Conclusion**:  
While both search engines seem to perform some sort of term normalization, they do not strictly use stemming. Instead, they utilize more sophisticated techniques like lemmatization, which provides results based on the meaning of the word rather than just its root form.

#### Do both search engines filter stop words?

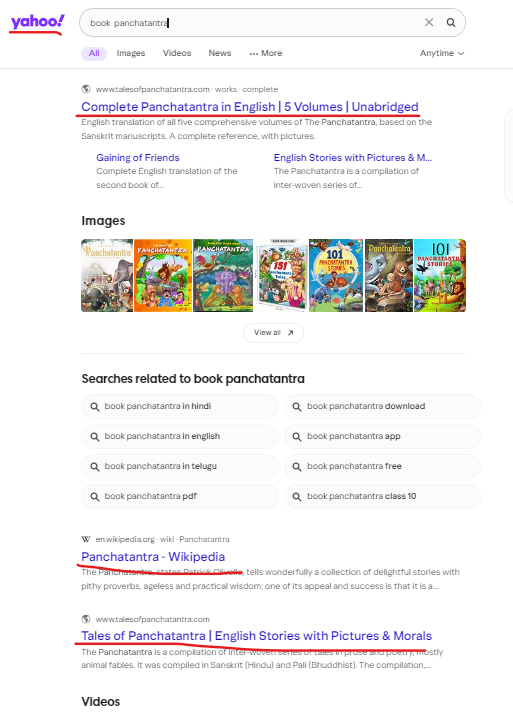
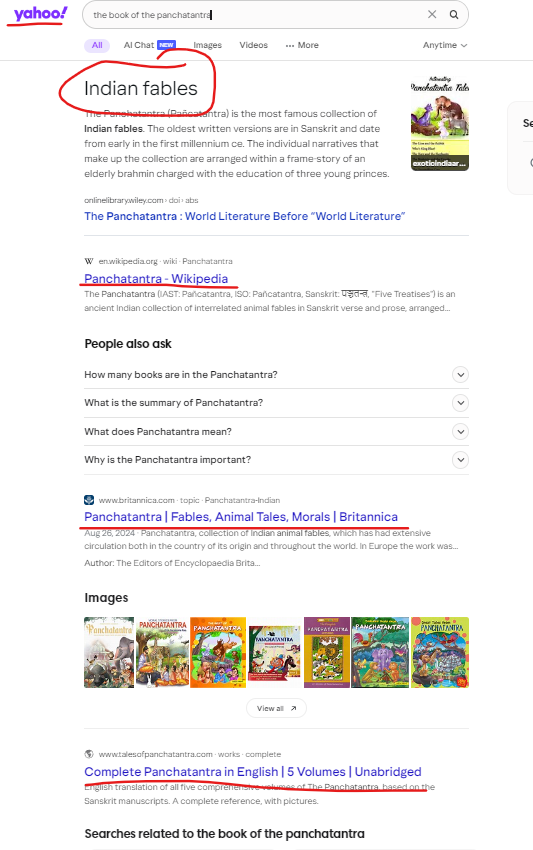
* **Stop words** are common words such as "the," "is," "at," "in," and "of" that typically don’t add significant meaning to search queries. Search engines like **Google, Yahoo, Bing, etc.** often ignore these words to speed up searches and focus on the most important terms.

**Testing Method**:

1. Performing searches with stop words and without stop words on both **Google** and **Yahoo**
   1. “the book of the panchatantra" vs "history world".
      1. Google –



* + 1. Yahoo -



1. I compared the results to see whether the inclusion of stop words affected the ranking or content of the results.
2. The ranking did differ by a small margin, by prioritising on rank 1 the Amazon(ecommerce website) or the short summary of the book.

**Results**:

* Both **Google** and **Yahoo** appear to ignore stop words in most cases. For example, searching for " **the book of the panchatantra**" yields the same results as searching for "**book panchatantra**"
* However, there are exceptions. For very specific searches, like exact phrase matches (using quotation marks), stop words do seem to matter. For example, searching for “**The Book**” (with quotes) yields different results than searching for ”**Book**” alone.

**Conclusion**:  
Yes, both search engines filter out stop words in general searches, unless you're doing a very specific search or using quotes for exact matches. In most scenarios, stop words don't affect the results, allowing the engines to focus on the key terms that matter most to the search.

# Question 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%)

#### What is the precision of the system on this search?

The precision of the system for this search is approximately **46.67%**. This indicates that about **46.67%** of the retrieved documents were relevant.

#### What is the recall of the system on this search?

The recall for this search is **17.5%**, which means that the system only found about of the relevant documents available in the collection, leaving a significant portion undiscovered.