Problem 1 (What comes first). Consider a repository of 1000 documents of which 28 are relevant to a user query. There are two search engines A and B. Search engine A returns 10 documents and search engine B returns 20 documents. An expert says that 7 of the 10 documents returned by A are relevant while 5 of the 20 returned by B are relevant.

(a) Calculate the precision, recall and accuracy metrics of the search engines.

(b) Which search engine is the winner in terms of its i) accuracy and ii) recall?

(c) Accuracy is not the best metric for IR evaluation. Briefly justify this statement based on the previous results.

(d) Let the order of relevant documents in the corpus be 1 (most relevant), 2, ..., 28 (least relevant). We mark the documents retrieved by search engines A and B, by relevance and X if not relevant and get the following results.

A: 1 X X 4 2 X 3 6 7 5

B: 1 3 2 4 5 X .. (13 irrelevant documents) .. X

Which search engine do you think is a better performer for the top 5 results. Why?

Problem 2 (Et tu, Brute?). Postings or inverted list increase the efficiency since instead of scanning entire documents word for word, we look up for the query terms and return the linked list of document IDs that contain the term in constant time. Assume that you have created an inverted index on the Shakespeare’s Collected Works. The terms Brutus, Cæsar and Calpurnia have posting lists of sizes x, y and z, respectively.

a What is the time complexity for the querying the keywords "Brutus" AND "Calpurnia"? When is this achieved?

b What is the time complexity for querying "Brutus" OR "Calpurnia"?

c Make the following queries on Google (keeping uppercases intact) and report the number of estimated search results.

(a) Caesar

(b) Brutus

(c) Caesar AND Brutus

(d) Caesar OR Brutus

(e) Brutus AND Caesar

(i) Do the numbers in the first four queries follow Boolean logic? Briefly justify.

(ii) Do the numbers in the third and fifth query follow Boolean logic? Briefly discuss, with an example, why word order should affect search results.
Problem 3 (To grep or not to grep). The season of Triwizard Tournament is back and there is a stir in both Wizarding and Muggle worlds. With the challenge of the unprecedented tasks impending on the Champions, they need all hands on deck. As a Muggle powered by the knowledge of IR, you are hired as a consultant to the Hogwarts School to help its champions complete their tasks.

Consider the following users and their queries. Your job is to choose a method of retrieval from the ones taught in the lecture (grep, boolean, tf-idf, ..), to be used for finding relevant documents. Give one line justifying your choice. Assume that all data exists only as text documents. Any component of the IR pipeline necessary for the final retrieval must be mentioned.

(a) **User**: Harry Potter (a competing champion in the Triwizard tournament)
**Background**: Harry wants to know all instances of deaths that might have occurred in the previous tournaments. As a champion he has access to the documents recording the events of past tournaments by year and Harry would want to retrieve all documents containing the the following keywords.
**Query**: "Triwizard Cup" and "death"

(b) **User**: Hermione Granger (A bright witch who is also Harry’s best friend and supporter)
**Background**: Hermione knows that the first task involves dragons and wants to read up on all charms and curses that can be used on dragons. The Hogwarts library has millions of documents and thousands on dragons alone. Scanning each document serially would take days. She has but a few hours and some important keywords.
**Query**: "charm", "curse", "dragon"

(c) **User**: Ronald Weasly (Harry’s best friend)
**Background**: Ron finds out the four dragon species that have been brought in for the first task. He knows that the library has an index of the dragon species and wants documents with exact mentions of any of the dragon species.
**Query**: "Hungarian Horntail" or "Swedish Short-Snout" or "Welsh Green" or "Chinese Fireball"

(d) **User**: Albus Dumbledore (headmaster of Hogwarts)
**Background**: Professor Dumbledore suspects foul play by organizers, since Harry being an underage wizard should never have been selected as a Champion. He wants to look into any past instances of involvement of organizers in manipulating results in the tournament.
**Query**: "Rigging the Goblet of fire in Triwizard tournament"