Name:

1. **Text laws and pre-processing**.

   - ❑ Stemming has no effect on an Information Retrieval System, as long as it is done on both documents and queries.
     FALSE: it has effect on size of the index, time it takes to compute answers, and the answers themselves.
   - ❑ Heap's law studies the relationship between length of documents and their vocabulary size.
   - ❑ Zipf's law studies the relationship between number of occurrences of words and their length.
     FALSE: not their length but their rank when sorting according to frequency.
   - ❑ When we plot rank vs. frequency of words in human-generated text in a log-log scale it is not uncommon to observe a linear dependence.
   - ❑ When we plot rank vs. frequency of words in human-generated text it is not uncommon to observe an exponential decay of the frequencies.
     FALSE typically we observe a heavy tail.
   - ❑ Zipf's law is a law and therefore it is always true, even for artificially generated texts.
     FALSE definitely not true for artificially generated texts.
   - ❑ Heap's and Zipf's law are essentially the same in that they relate the same aspects of text.
     FALSE, they related different aspects.
   - ❑ Given a text, one can use linear regression techniques to estimate α, even though this parameter is in the exponent of the rank variable.
   - ❑ ElasticSearch is a NoSQL/document database with the capability of indexing and searching text documents.
   - ❑ Scrapy is a distributed document database for developing web crawlers and extracting information from web pages.
     FALSE not a database, just a python software library.

2. **IR Models**.

   - ❑ The Vector model takes into account the frequency of words in documents.
   - ❑ The Boolean model takes into account the order of words in documents.
     FALSE, order is ignored.
   - ❑ The Vector model takes into account the order of words in documents.
     FALSE, order is ignored.
   - ❑ In the Boolean model, it is important to return answers sorted by their relevance with respect to a given query.
     FALSE in the Boolean model a document is either fully relevant or not relevant, and nothing in between so we cannot sort documents in an answer.
   - ❑ In the Vector model, documents are represented using vectors of non-negative real numbers.
   - ❑ In the Vector model using tf-idf weights, if a term t appears more times in document d than in document d', then its weight in d will always be higher than in d'.
     FALSE not true since we normalize by the max frequency of a word in a document.
   - ❑ The norm of a tf-idf vector of any document is always positive and bounded by +1.
     FALSE the norm can certainly be higher than 1, there are plenty of examples in the material for the course.
   - ❑ The cosine similarity between two documents in a corpus can be negative in case the documents are very dissimilar.
     FALSE vectors do not contain negative values and therefore cosine similarity cannot be negative.

- ❑ The length of tf-idf vectors depends on the length of the documents they represent.
  FALSE the length of tf-idf vectors is fixed for all documents (it is the size of the vocabulary).
- ❑ If two documents have cosine similarity of 1, it means that they are the same document.
  FALSE just permute all the words in the document.

3. **Implementation**.

- ❑ Storing the document-term frequency matrix is necessary in order to compute query answers efficiently.
  FALSE, we can use an inverted index.
- ❑ A large part of the query-answering time is spent bringing posting lists from disks to RAM.
- ❑ In a unary compression scheme, the length of encoding x is proportional to the value of x.
- ❑ In Elias-Gamma code, the length of encoding x is proportional to the value of x.
  FALSE it is proportional to the logarithm of x.
- ❑ Unary code is useful for encoding frequencies, since their distribution is biased towards small numbers.
- ❑ Query optimization is the process by which one finds the best queries for a given retrieval task.
  FALSE is the process by which one finds the best execution plan for a given query.
- ❑ Gap compression in combination with a fixed-length binary encoding scheme for document identifiers drastically reduces the size of the inverted index.
  FALSE Gap compression is only useful with variable-length schemes.
- ❑ If we use unary encoding to compress the frequencies in posting lists, then the size of the inverted index (in bits) is roughly equal to the length of the corpus.
  FALSE we still need to encode document ids; **however** this question could be understood like the size of the inverted index that encodes frequencies has the size denoted, which is TRUE so I accepted both.
- ❑ The Elias-Gamma code for the number 4 has length 4.
  FALSE the Elias-Gamma code for the number 4 is 00100 (length 5).
- ❑ Compressing 10 natural numbers using a unary encoding scheme, needs $10 * log_2(10)$ bits.
  FALSE the length depends on the value of the numbers themselves.

4. **Evaluation and Relevance Feedback**.

- ❑ It is trivially easy to optimize recall in an Information Retrieval system.
- ❑ It is very hard to optimize precision in an Information Retrieval system.
  FALSE: return empty answers.
- ❑ We typically find a balance between recall and precision by playing with the size of the answer.
- ❑ The rank-precision curve decreases monotonically.
  FALSE it may increase as we move as we increase the size of answer (or rank).
- ❑ The rank-recall curve increases monotonically.
- ❑ In general, we should always optimize precision over recall because it is important to present relevant documents to users.
  FALSE it depends on the application.
- ❑ In Rocchio's rule, the weight of existing terms in the original query can never decrease.
  FALSE, if non-relevant documents contain those terms it may go down (provided $\gamma > 0$)
- ❑ Relevance feedback is typically used to optimize precision.
  FALSE it typically increases recall at the expense of precision.
- ❑ Relevance feedback is a technique that uses user's feedback to (potentially) improve on user's initial queries.

- ❑ In web search, precision matters much more than recall, so the extra computation time and user patience required by relevance feedback may not be productive

5. **Web Search**.

- ❑ Crawling is the process by which search engines obtain the content and structure of the web graph.
- ❑ Take a star-shaped graph with n nodes, with all edges pointing from the central node to the outside $n-1$ nodes. Then, the pagerank of the central node is $\frac{1-\lambda}{n}$.
  FALSE, it will be higher because we should take into account the extra pagerank from the outer nodes (we add the "edges" from outer nodes to every other node).
- ❑ In the graph from the previous question, all nodes have the same pagerank independently of λ.
  FALSE the central node has different pagerank from outer nodes.
- ❑ The number of neighbors of a node in a graph determines its pagerank.
  FALSE the number of neighbors is not enough to fully determine the pagerank.
- ❑ In a complete graph, the pagerank of nodes changes as a function of λ.
  FALSE in this case, the pagerank is $1/n$ independently of λ.
- ❑ In PageRank, the power method is guaranteed to converge for all values of λ, including $\lambda = 1$.
  FALSE, if $\lambda = 1$ and we have a cyclic graph, for example, then the power method does not converge.
- ❑ PageRank is an algorithm that uses content and structure of web pages to determine the relevance of a page.
  FALSE content is not used at all.
- ❑ The hub value of a node is determined by the hub values of neighboring nodes.
  FALSE, the hub value is determined by authority values of neighboring nodes.
- ❑ The pagerank value of a node is determined by the pagerank values of neighboring nodes.
- ❑ In HITS, the hub and authority values are computed for a relevant subset of the web graph only.