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.
   - ❏ 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.
   - ❏ 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.
   - ❏ Zipf's law is a law and therefore it is always true, even for artificially generated texts.
   - ❏ Heap's and Zipf's law are essentially the same in that they relate the same aspects of text.
   - ❏ 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

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.
   - ❏ The Vector model takes into account the order of words in documents.
   - ❏ In the Boolean model, it is important to return answers sorted by their relevance with respect to a given query.
   - ❏ 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'.
   - ❏ The norm of a tf-idf vector of any document is always positive and bounded by $+1$.
   - ❏ The cosine similarity between two documents in a corpus can be negative in case the documents are very dissimilar.
   - ❏ The length of tf-idf vectors depends on the length of the documents they represent.
   - ❏ If two documents have cosine similarity of 1, it means that they are the same document.

3. **Implementation**.

- ❏ Storing the document-term frequency matrix is necessary in order to compute query answers efficiently.
- ❏ 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.
- ❏ 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.
- ❏ Gap compression in combination with a fixed-length binary encoding scheme for document identifiers drastically reduces the size of the inverted index.
- ❏ 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.
- ❏ The Elias-Gamma code for the number 4 has length 4.
- ❏ Compressing 10 natural numbers using a unary encoding scheme, needs $10 * log_2(10)$ bits.

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.
- ❏ We typically find a balance between recall and precision by playing with the size of the answer.
- ❏ The rank-precision curve decreases monotonically.
- ❏ 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.
- ❏ In Rocchio's rule, the weight of existing terms in the original query can never decrease.
- ❏ Relevance feedback is typically used to optimize 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**.

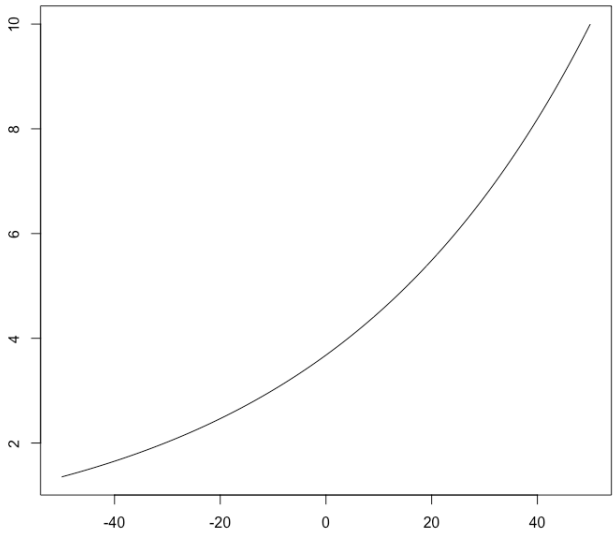
- ❏ Crawling is the process by which search engines obtain the content and structure of the web graph.
- ❏ Take a star-shaped graph with n nodes, with all edges pointing from the central node to the outside $n-1$ nodes. Then, the pagerank of the central node is $\frac{1-\lambda}{n}$.
- ❏ In the graph from the previous question, all nodes have the same pagerank independently of λ.
- ❏ The number of neighbors of a node in a graph determines its pagerank.
- ❏ In a complete graph, the pagerank of nodes changes as a function of λ.
- ❏ In PageRank, the power method is guaranteed to converge for all values of λ, including $\lambda = 1$.
- ❏ PageRank is an algorithm that uses content and structure of web pages to determine the relevance of a page.
- ❏ The hub value of a node is determined by the hub 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.

CAIM (Primer parcial - Nov. 15th, 2018)

Instructions:

- ❏ tick **clearly** the claims that you think are **true** with a $\surd$
- ❏ tick **clearly** the claims that you think are **false** with a $\times$
- ❏ if you want to "withdraw" an already ticked box, black it out as ■ (it will count now as unanswered)
- ❏ all questions are equally weighted (the headings define **blocks** of **ten** questions each)
- ❏ there is no obligation to answer individual questions, but at least half (**five**) questions in each block must be answered
- ❏ individual question grading: correct answers count $+1$ point, incorrect answers count -1 point; no answer counts 0 points (there are 50 questions = 50 points maximum)
- ❏ letting S be the number of points, the overall grade is obtained as

$$f(S) = 10 \exp\left(\frac{S}{50} - 1\right)$$



- ❏ time: 2h