**Homework assignment Set #1**

**Instructions:** This document is generated with Microsoft Word MacOS version 16.16.6. You may use a different version, as long as you can read the problems and input answers. Type the answers into empty space below the problems, create additional pages if needed. Please create the figures using your preferred plotting tool, e.g., Excel, PyPlot, and paste the them to where is suitable. Please do not make any changes to the original problem! You may insert equations using the Insert->Equation module if it is available, but this is not obligated.

After finishing the problems, please rename the file by appending your first name after the current file name. For example, set1-jian.docx, and upload it to the box folder I shared with you by **noon on Tuesday, February 12, 2019.**

Please honor the **academic integrity** in the syllabus. Violators, if confirmed, will be graded as 0.

**Problem 1:** (1) Give the Big O complexity (asymptotic complexity) of the growth of work of each expression below; (2) Order the following as to increasing complexity as a function of the work (all unspecified terms such as , , etc. are positive to be determined constants); (3) Label each as reasonable/unreasonable and practical/impractical for scaling. For full credit show your reasoning. Use additional space below the table if needed.

|  |  |  |  |
| --- | --- | --- | --- |
| Expression |  | Order (highest to lowest) | Scaling |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Problem 2:** In this exercise, you will familiarize yourself with standard methods for evaluating information retrieval systems with an emphasis on precision and recall.

1. Suppose that an IR system contains only 1000 documents. A query is known to generate 27 relevant documents as listed below:

{d1, d5, d7, d10, d88, d151, d200, d211, d250, d300, d399, d401, d405, d450, d473, d500, d501, d530, d545, d590, d600, d735, d700, d720, d800, d888, d900}.

Two different IR systems are used to retrieve ranked documents for this query. Each system only returns the top 10 ranked documents in order of ranking. Systems 1 and 2 each retrieves documents one at a time in the following order with all 10 documents eventually returned:

System 1: d122, d211, d150, d88, d37, d1, d501, d800, d201, d5.

System 2: d10, d700, d6, d250, d88, d600, d59, d422, d500, d7.

R

P

Answer the following and show your work:

1. Plot the Precision and the Recall graphs for each system as a function of the number of documents returned (for 1 document returned, 2 documents returned, etc).
2. Plot the Precision versus Recall for Systems 1 and 2 using these query results as a function of the number of documents returned. Note that is the value of precision and recall for the first document, for the 2 documents.
3. Which IR system is better? Justify your answer.

**Problem 3:** What can be measured by a search engine? Precision or recall or both? Why?