2019 Ph.D. Program Qualify Examination – Information Retrieval

1. (25%) Please explain what is the *Zipf Distribution* for indexing scheme. What is the linear and log scale?

2. (25%) Please explain the concept of "TFIDF" for an information retrieval system in terms of word frequency and word rank.

3. (25%) Please describe the detailed procedure of how to obtain the 11-point *precision* and *recall rates* curve. Given you have a test data set consists of *n* documents for the purpose of information retrieval, and illustrate how to draw the curve from the retrieving results. Also, explain the *sensitivity* and *specificity* in terms of positive/negative results.

4. (25%) What is the "high recall and high precision" strategy? Please explain in detail how to possibly achieve it during design an IR system?