Web Information Retrieval and Mining Study Pre-Quiz

1. (70 points) For each of the following statements, determine whether it is correct or not. If you don’t know what it says, please leave it blank.
   (1). (T) F-measure is used to evaluate the performance of an IR system.
   (2). (F) C4.5 is a clustering algorithm.
   (3). (T) False-positive and false-negative all mean the wrong answers.
   (4). (F) TKC effect would occur within one Web page.
   (5). (F) HITS is a query-independent link analysis algorithm.
   (6). (F) The IDF value depends on the term frequency and the document frequency.
   (7). (F) Inverted index is a fast technique to get terms in a document.
   (8). (T) Number of indexed pages in Google is more than 8 billion now.
   (9). (F) Signature file is a reversible technique to encrypt documents.
   (10). (T) SALSA is a link analysis algorithm.
   (11). (F) The precision and recall ratios are important for all IR systems.
   (12). (F) Stop words are used to delimit the word segments in a sentence.
   (13). (T) We can get those pages linking to a specific page by a single query in Google.
   (14). (F) K-means is a classification algorithm and K-NN is a clustering algorithm.

2. (30 points) For the following researchers, please list their related work from the listed items.
   (1). (e) Claude Elwood Shannon
   (2). (b) Jon. M. Kleinberg
   (3). (f) Chia-Hui Chang
   (4). (l) Tim Berners-Lee
   (5). (h) Soumen Chakrabarti
   (6). (i) Sergey Brin and Larry Page

   (a) PLSA, (b) HITS, (c) HMMs, (d) Power-law, (e) Information theory, (f) IEPAD, (g) k-NN, (h) CLEVER, (i) PageRank, (j) SALSA, (k) SVM, (l) WWW