Week 1 Homework

Ben Arancibia

August 29, 2015

Exercise 1.2.1 : Using the information from Section 1.2.3, what would be the number of suspected pairs if the following changes were made to the data (and all other numbers remained as they were in that section)?

(a) The number of days of observation was raised to 2000.

From section 1.2.3 - the original formula: $5 * 10^{17} * 5 * 10^5 * 10^{-18} = 250000$

The approximate $5 * 10^5$ would increase since it is now 2000, choose 2 which is $2 * 10^6$.

New formula is now: $5 * 10^{17} * 2 * 10^6 * 10^{-18} = 1000000$

(b) The number of people observed was raised to 2 billion (and there were therefore 200,000 hotels).

Increasing the number of people observed to 2 billion changes the probability of visiting the same hotel. Before, that was 10^{-9}. With 200,000 hotels, the probability is now $\frac{.0001}{2*10^9}$ which is $5 * 10^{-10}$. The chance that two people will visit the same hotel on two different given days is now $5 * 10^{-20}$.

The increase also effects the number of pairs of people visiting the hotel on the same day.

This also affects the the number of pairs of people, which is now approximately nˆ2/2 where n = $2 * 10^9$, which evaluates to 4*10ˆ18 / 2 = $2^{10}18$.

Plugging in the new values, the new formula is:

$2 * 10^{18} * 5 * 10^5 * 5 * 10^{-20} = 50000$

(c) We only reported a pair as suspect if they were at the same hotel at the same time on three different days.

The calculations or the three different days means that 1000 choose 2 needs to change to 1000 choose 3. This approximately is $1.7 * 10^9$.

The new formula is:

$5 * 10^{17} * 1.7 * 10^9 * 10^{-18} = 8.5 * 10^8 = 850000000$

Exercise 1.3.2: Suppose there is a repository of ten million documents, and word w appears in 320 of them. In a particular document d, the maximum number of occurrences of a word is 15. Approximately what is the TF.IDF score for w if that word appears (a) once (b) five times?

Given: $N = 10^7$ $n_i = 320$. $max_k f_{kj} = 15$.

IDF score is $IDF_i = log_2(N/n_i) = log2(10^7/320) = 14.931$. Only the TF score changes.

(a) What is the TF.IDF score for w if that word appears once:

$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}} = 1/15$

$TFIDF = TF_{ij} * IDF_i = \frac{14.931}{15} = 0.9954$

(b) What is the TF.IDF score for w if that word appears five times:

$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}} = \frac{5}{15} = \frac{1}{3}$

$TFIDF = TF_{ij} * IDF_i = \frac{14.931}{3} = 4.977$

Exercise 1.3.5 : Use the Taylor expansion of e^x to compute, to three decimal places: (a) $e^1/10$ (b) $e(1/10)$ (c) e^2.

(a) The taylor series expansion of $e^{1/10}$:

$$e^{1/10} = 1 + (1/10) + \frac{(1/10)^2}{2} + \frac{(1/10)^3}{6} + ...$$

Computes to 1.105

(b) The taylor serious of $e^-(1/10)$ is the reciprocal of (a).

Compute to .905. The taylor expansion has alternating signs since the initial value is negative.

(c) The taylor series expansion of e^2:

$$e^2 = 1 + (2) + \frac{(2)^2}{2} + \frac{(2)^3}{6} + \frac{(2)^4}{24} + \frac{(2)^5}{120} + \frac{(2)^6}{720} + + \frac{(2)^7}{7!} + \frac{(2)^8}{8!}$$
$$= 1 + 2 + 2 + 1.33333 + 0.6666667 + 0.2666667 + 0.08888889 + 0.02539683 + 0.006349206$$

Computes to 7.387.