Programming and Classification

List 2

Marek Klonowski

(To solve some problems from this list you will need some knowledge presented during lecture # 2 and # 3 as well as some background in probability theory.)

IV. More about probability and combinatorics

1. What is the number of words of length k over the alphabet $\{a, b, c, d\}$?
2. What is the number of words of length k over the alphabet $\{a, b, c, d\}$ without letter a ?
3. What is the number of words of length $k > 3$ over alphabet $\{a, b, c, d\}$ without letter b starting with $'aaa'$?
4. We toss a symmetric coin till we get a tail for the first time. Let X be the random variable denoting the number of trials before getting the first tail.
5. What is the number of dictionaries (mappings) form a set A into set B?
   (a) Compute $\Pr[X > 5]$.
   (b) Compute $\Pr[X \text{ is even}]$.
   (c) Compute $E(X)$.
   (d) Let t_1, t_2 be natural numbers. Prove that

$$\Pr[X > t_1 + t_2 | X > t_1] = \Pr[X \geq t_2] .$$

V. Simple text processing methods

Let us consider a corpus of $N = 3$ texts:

$text_1$ =`We keep our marriage classy by texting each other instead of shouting through walls while we are home.`
$text_2$ =`Marriage teaches you a lot about yourself. For instance, I have learned that I do not need to use so many paper towels, and they are expensive.`
$text_3$ =`The best way to remember your wife birthday is to forget it once .`

1. Compute $TF_{\text{classy},1}$, $TF_{\text{it},2}$, IDF_{are}.
2. Compute $TF.IDF(\text{are}, 1)$, $TF.IDF(\text{are}, 2)$, $TF.IDF(\text{do}, 3)$.
3. When $TF.IDF(\text{are}, 1) = 0$? Consider all cases. Try to provide some interpretation.