

# Assignment 1: Linear Algebra and Probability

Submission: Tuesday August 25th

Maximum 2 students per group

Prof. Fabio A. González  
Machine Learning - 2015-II  
Maestría en Ing. de Sistemas y Computación

1. Do the tutorial “Getting Started With Python For Data Science” (<https://www.kaggle.com/wiki/GettingStartedWithPythonForDataScience>).
2. Let  $D = \{d_1, \dots, d_n\}$  be a set of documents and  $T = \{t_1, \dots, t_m\}$  a set of terms (words). Let  $TD = (TD_{i,j})_{i=1\dots m, j=1\dots n}$  be a matrix such that  $TD_{i,j}$  corresponds to the number of times the term  $t_i$  appears in the document  $d_j$ . Also, let  $l_i$  be the length, number of characters, of term  $t_i$ , and let  $L = (l_1, \dots, l_m)$  be a column vector. Finally, assume a process where a document  $d_j$  is randomly chosen with uniform probability and then a term  $t_i$ , present in  $d_j$ , is randomly chosen with a probability proportional to the frequency of  $t_i$  in  $d_j$ .

For all the following expressions you must provide:

- a mathematical expression to calculate it that includes  $TD$ ,  $L$ , constants (scalars, vectors or matrices) and linear algebra operations
- a expression in Numpy (<http://www.scipy.org>) that, when evaluated, generates the requested matrix, vector or scalar (the expression must be a linear algebra expression that does not involve control structures such as for, while etc.)
- the result of evaluating the expression, assuming:

$$TD = \begin{bmatrix} 2 & 3 & 0 & 3 & 7 \\ 0 & 5 & 5 & 0 & 3 \\ 5 & 0 & 7 & 3 & 3 \\ 3 & 1 & 0 & 9 & 9 \\ 0 & 0 & 7 & 1 & 3 \\ 6 & 9 & 4 & 6 & 0 \end{bmatrix} \quad L = \begin{bmatrix} 5 \\ 2 \\ 3 \\ 6 \\ 4 \\ 3 \end{bmatrix}$$

- (a) Matrix  $P(T, D)$  (each position of the matrix,  $P(T, D)_{i,j}$ , corresponds to the joint probability of term  $t_i$  and document  $d_j$ ,  $P(t_i, d_j)$ )
- (b) Matrix  $P(T|D)$
- (c) Matrix  $P(D|T)$
- (d) Vector  $P(D)$
- (e) Vector  $P(T)$
- (f)  $E[l]$  (the expected value of the random variable  $l$  corresponding to the length of a randomly chosen term)
- (g)  $\text{Var}(l)$  (the variance of  $l$ )

3. The assignment must be submitted as a IPython notebook through the following Dropbox file request, before midnight of the deadline date. The file must be named as `m1-assign1-unalusername.ipnb`, where `unalusername` is the user name assigned by the university of one of the members of the group.