LINGI2263 : COMPUTATIONAL LINGUISTICS

Group 2 : Project 3

Authors:
Crochelet Martin (2236-10-00)
Baugnies Benjamin (6020-10-00)

Professor:
Pierre Dupont
Cédric Fairon

2013 - 2014

1 TF-IDF

2 The underpinnings of Log-Entropy

2.1 Why is there a '+1' in the logarithm function ?

It is simply because the logarithm function is defined on the domain of the strictly positive real: it is possible for tf_{ij} to be equal to 0 simply because of it's definition: it is possible the i^{th} document does not contain the j^{th} therm. Hence, the '+1' allows us to shift the domain of tf_{ij} from [0, $||document_i||$] to [1, $||document_i|| + 1$] on which the logarithm function is entirely defined.

2.2 What is the point to apply the logarithm function to the term frequency instead of plugging it directly in the formula ?

This allows us to diminish the relative importance of the high frequency words versus the middle ones. Indeed, the logarithm function is defined so that, compared with the linear function suggested, it keeps the middle frequencies importance while really lowering the high frequencies. This effect allows us to take into account the fact that some words tends to appear in every document while not being decisive for de description of the document (words such as "the", "a", "b", etc.). Moreover, some other words only appear a few times in the document in question while being really important for the definition/description of it; the importance of theses words is better taken into account by the logarithm function than by the linear one.

2.3 Could explain intuitively the mechanics of the global weight ? And why is there a division by log n? (Hint: What does the entropy of a distribution represents?)

First, let us remark that the global weight can easily be written as:

$$g_{ij} = 1 - \frac{-\Sigma_i \, p_{ij} \cdot \log p_{ij}}{\log n} \tag{1}$$

With this particular notation, the resemblance with formula of the entropy for the distribution p_{ij} is quite hard to miss ($\Sigma_i \, p_{ij} \cdot \log p_{ij}$). This entropy measures the uncertainty of the location of j^{th} in the i^{th} document. Now we focus on the division by the $\log n$ which is simple too: the maximum value of this entropy is actually $\log n$ meaning that we project the domain of the entropy of the distribution onto a domain bounded by 0 and 1. Then, we inverse the sense of the domain by the operation of subtraction: this allows us to define a greater weight for the terms that more certain (with less uncertainty) to be present in the document. Moreover, this also allows us to give a greater weight to the terms that a specific to this document by lowering the one of the terms that are more likely to appear in each document.