Guillaume Lemaitre

Data Mining Learning Machine
Coursework C

Find similarities of documents using Jaccard coefficient

We present two simple methods allowing to find the similarities between two documents using Jaccard coefficient. The Jaccard coefficient is defined by:

$$\frac{A \cap B}{A \cup B}$$

Where A and B are two respectively a vector of each document.

1. **Method based on the fact that document is a set of words:**

This method is a very simple and not very robust method to find similarities two documents. We assume that a document is a set of words. To use Jaccard coefficient, we need to define a vector for each document. Hence, we put every different words of one document inside a vector. For instance, if the document is “Elephants are not pink. Elephants are gray.” the vector that we must create is the following: A = {Elephants, are, not, pink, gray}. So, we create a vector for each document and apply the Jaccard coefficient. However, the main problem of this method is that we don’t find the sense of the document. And we consider that irrelevant words like “be” or “have”, have the same weight that relevant words specific to the text. The second method can improve this aspect.

2. **Method based on lexical fields:**

The second will be a method where we create several vectors representing lexical fields of a document. Then, we could apply Jaccard coefficient between each vectors and find if a vector of a document have a correspondence with a vector of another document. Basically, the meaning of this method is to understand the subject of each document and try to know if the subject of each document speaks about the same subject. The problem of this method is to compute lexical fields. In fact, it is difficult to construct a vector like that. Maybe, we can apply a training to choose on each lexical field belong each word.