Detection of/between similarity of documents with hashing

Roger Vilaseca Darné, Xavier Lacasa Curto and Xavier Martín Ballesteros
Algorithms

1st December 2018

1 Introduction

2 Jaccard Index

The Jaccard Index, also known as Intersection Over Union (IOU), calculates the percentage of similarity between two sets.

For any pair of sets S and T, the Jaccard Index is defined as:

$$J(S,T) = \frac{|S \cap T|}{|S \cup T|} \tag{1}$$

We can easily deduce that the more common words, the bigger the Jaccard Index, which means that it is more probable that one set is a duplicate of the other.

Example 2.1 *In Figure we see two sets S and T. There are X elements in the intersection and Y in their union. Thus, J(S, T) = X.*

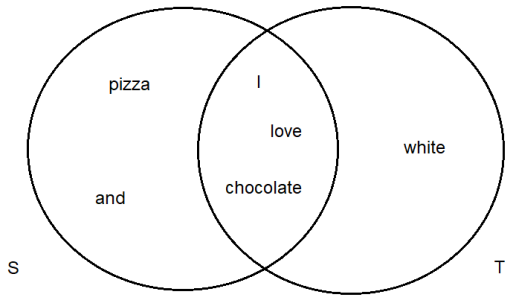

Figure 2.1: Two sets with Jaccard Index 3/6.

3 Shingling

Mas texto.

4 Minhashing

Mas texto.

5 Locality Sensitive Hashing (LSH)

Mas texto.

5.1 Referencies

https://towardsdatascience.com/understanding-locality-sensitive-hashing-49f6d1f6134
https://santhoshhari.github.io/Locality-Sensitive-Hashing/
https://www.youtube.com/watch?v=96WOGPUgMfw
https://www.youtube.com/watch?v=_1D35bN95Go
https://medium.com/engineering-brainly/locality-sensitive-hashing-explained-304e
http://www.mit.edu/~andoni/LSH/
http://infolab.stanford.edu/~ullman/mmds/ch3.pdf

References

[1] Author, *Title*, Editor, (year)