Data Mining Assignment 1

Mauro Pungo,Viktoria Sartor

November 2021

1 Description

2 Build Instructions

Syntax: python homework.py [ARG_FLAG1] [ARG_VALUE1]
For further help just run *python homework1.py -h* Run script *homework1.py* in the same directory were the text-directory is stored.

Argument parser:

```
--text1 choices: 'dogs_wiki.txt', 'cats_wiki.txt', 'lorem_1.rtf',
lorem_2.rtf', 'church_1.txt', 'church_2.txt'

--text2 choices: 'dogs_wiki.txt', 'cats_wiki.txt', 'lorem_1.rtf',
'lorem_2.rtf', 'church_1.txt', 'church_2.txt'

--shingle_len default: 5

--permutations default 100

All flags with double hyphens!
```

3 Classes

3.1 Shingling

Given a document, this function deals with the representation of a text document as sets of length k characters sequence, where k is a parameter.

3.2 CompareSets

This class implements methods to compute the Jaccard similarity and distance

3.2.1 Jaccard Similarity

This method takes as arguments two documents, represented as a sets of k-shingles and computes their Jaccard Similarity (JS), given by the number of overlapping elements on the 2 sets and the number of total unique elements on the 2 sets. Mathematically we have $JS(A, B) = \frac{\#\{A \cap B\}}{\#\{A \cup B\}}$

3.2.2 Jaccard Distance

The Jaccard distance (JD) of 2 sets can be seen as the "dissimilarity" between both sets. Therefore it is easily defined in terms of the Jaccard similarity by:

$$JD(A, B) = 1 - JS(A, B) = 1 - \frac{\#\{A \cap B\}}{\#\{A \cup B\}}$$

3.3 MinHashing

Mention that the variance between runs can be decreased by increasing the number of permutations,thus sacrificing computational speed.

3.4 CompareSignatures

Given 2 signature vectors of size n of 2 documents, this class computes the similarity between the 2 documents by computing the fraction of similar hash values.

3.5 LSH

4 Texts

We used two different lorem upson texts with each 100 words. We also used parts of the german wikipedia articles about the protestant and the catholic church. The two last texts are full wikipedia articles about cats and dogs. The sources of the texts can be found in the text files respectively.

5 Results

With our default values we were not able to find similar documents using the larger text files. But the topics of the documents were not really similar, so the results are not surprising. When we compare one document to itself we get a Jaccard similarity and a signature similarity of 1.0 and a Jaccard distance of 0.0. When we used simple strings as inputs the minhashing found correct similarities.