Data Science and Big Data

Summer Term 2017

5 - Document Deduplication

electronic submission: June 21, 2017, 14:25 (strict)

Sergey Kuptsov (`sergeykuptsovde@gmail.com`)

1. Implementation + Dataset [40 points]

Create a set $\{D_1, \ldots, D_{100}\}$ of documents generated as follows:

- Let D_1 be a random string of length 1000 bytes.
- Let D_i $(1 < i \leq 100)$ be a string obtained from D_1 by replacing k_i random characters in D_1 by random values between $0, \ldots, 255$, and by swapping l_i random character pairs in D_1. The parameters k_i and l_i should increase with i, i.e., with increasing i we add more and more amount of noise to D_1.

Implement the algorithm computing the sketch matrix M_s for $\{D_1, \ldots, D_{100}\}$ as presented in the lecture. Use characters as shingles and Rabin's fingerprints for representing q-shingles with $N = 16$ (cf. Slide 11). You can use any irreducible polynomial $P(x)$ fixed in advance (i.e., you don't need to implement an algorithm generating random irreducible polynomials). To get an irreducible polynomial, you can use e.g. the website

`http://zenfact.sourceforge.net/PIPS/polyformind.html`

(Set "Characteristic of the field" to 2 and "First extension" to 16.)

2. Presentation of your results [20 points]

For at least 10 different values of q, calculate the Jaccard similarity between D_1 and D_i in M for $i = 2, \ldots, 100$, as well as their similarity in M_s obtained by minhashing (see, also, Slide 24). Use 100 minhash functions (i.e., M_s will have 100 rows). Present your results in the exercise class.

Some technical remarks:

- Make sure to comment the code!
- Programming languages should be one of the following: Python (recommended), C++, C, or Java. If you want to use any other language please contact your tutor!

- Please send sources + compiled version with subject *Big Data Exercise 5* by

  June 21st, 14:25

  (i.e., before lecture start) to your tutor. In the email please give (i) your **group number**, (ii) the **names** of the group members who contributed to the solution, as well as (iii) the **programming language** you used.