

Detecció de similitud de documents amb hashing*

GRAU A CURS 2016-2017

Departament de Ciències de la Computació
alg@cs.upc.edu

Resum

Aquesta pràctica té com objectiu una validació experimental de l'efectivitat de diferents algorismes de hash per a la detecció de similitud de documents.

La pràctica es farà en grups de 3 persones (excepcionalment 2, sota autorització expressa). La composició dels grups s'haurà de comunicar a alg@cs.upc.edu abans del 20 de Octubre de 2016.

El lliurament de la pràctica es farà en línia via Racó, teniu temps fins las 23:59 hores del dia 15 de Novembre de 2016. Alguns grups poden ser convocats per a una entrevista personal (data per decidir) la setmana del 19 de desembre amb prova interactiva. Altres grups poden rebre preguntes per e-mail. És obligatori que a l'entrevista (si cal) estiguin presents tots els membres del grup.

I. OBJECTIUS

L'objectiu d'aquesta pràctica és per una part analitzar l'efectivitat de la detecció de similitud de documents de text en funció de la precisió de la representació del text i per una altra, implementar un algorisme de detecció de documents similars basats en hash. Per això us proposem que considereu representacions dels documents basades en l'ús de k -shingles (o k -grams) i de signatures minhash per tal d'avaluar l'efectivitat de la mesura de similitud computada. Una vegada determinats els paràmetres adients per a una col·lecció, volem un algorisme per determinar els documents mes similars d'un donat.

Una descripció exacta de les definicions i els mètodes la podeu trobar al capítol 3 del llibre *Mining of Massive Datasets* [1]. En concret la representació d'un document per un conjunt de k -shingles a la secció 3.2 i la representació d'un document per una signatura minhash amb t components obtinguda utilitzant t funcions de hash a la secció 3.3.2. Una versió electrònica del capítol la podeu trobar a <http://infolab.stanford.edu/~ullman/mmds/ch3.pdf>.

Us demanem que implementeu algorismes per:

- Obtenir la similitud de Jaccard de dos documents representats per conjunts de k -shingles. Recordeu que el grau de similitud de Jaccard de dos conjunts A i B es defineix com

$$Jsim(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

- Obtenir una aproximació del grau de similitud de Jaccard a través d'una representació via signatures minhash basades en t funcions de hash. La mesura de similitud de dos signatures a i b amb t components es defineix com:

$$sim(a, b) = \frac{1}{t} \sum_{i=1}^t \begin{cases} 1 & \text{if } a_i = b_i \\ 0 & \text{if } a_i \neq b_i \end{cases}.$$

*La versió més actualitzada d'aquest document, així com qualsevol material addicional relacionat es publicarà al Racó.

- Implementar un algorisme de Locality-Sensitive Hashing (LSH) basat en signatures minhash.

L'objectiu és per una part veure experimentalment la dependència entre precisió i temps de còmput en funció dels valors del paràmetres k i t per diferents col·leccions de documents i famílies de funcions de hash i per una altra, analitzar experimentalment la qualitat de les solucions computades per l'algorisme de LSH.

II. ENTREGA

El nivell de sofisticació i esforç dedicat a la pràctica és opcional i es tindrà en compte a l'hora d'avaluar-la. En la versió més senzilla (suficient per tenir bona nota si està acompanyada d'un bon disseny d'experiments) implementeu programes en C++ per als problemes proposats seguint la versió més simple del capítol 3 de [1] en particular l'algorisme de LSH per signatures minhash descrit a la secció 3.4. Versions més sofisticades de la Pràctica inclouran la implementació de algorismes no bàsics o de més d'un algorisme de LSH o de variacions de l'algorisme bàsic.

Tingueu en compte que haureu de mesurar el temps dels algorismes. A més, pot ser haureu de fer un seguiment de diversos comptadors que permetin quantificar el tipus i la quantitat de treball que el programa fa i la qualitat de la mesura/solució computada.

III. DADES

Aquest document és intencionadament vague. Per tant, a més d'analitzar i experimentar amb diferents funcions de hash i versions d'algorismes, haureu de documentar tant les decisions preses com el disseny d'experiments que els hi donen suport i la selecció de conjunt de dades. Si feu servir un repositori públic, cal que proporcioneu l'adreça del repositori així com (si cal) les modificacions i/o simplificacions sobre els conjunts de dades triats.

Com a únic requisit us demanem que un dels conjunts de dades sigui format per 20 documents que haureu d'obtenir mitjançant permutacions aleatòries d'un document de text amb 50 paraules.

IV. QUÈ CAL LLIURAR

Cal lliurar una carpeta comprimida (.zip) que contingui:

- Una breu descripció del què heu implementat, les funcions de hash que heu fet servir, les proves que heu fet i la comparació dels resultats que heu obtingut. També és interessant que indiqueu idees que heu provat encara que no hagin donat bons resultats. El document en format PDF ha d'incloure les referències adients.
- Una carpeta amb totes les fonts necessàries per compilar i executar la vostra pràctica. S'han d'incloure les instruccions per a la compilació i execució, així com els conjunts de dades utilitzats o els programes per generar-los aleatòriament si cal.

REFERÈNCIES

- [1] Jure Leskovec, Anand Rajaraman and Jeffrey D. Ullman *Mining of Massive Datasets* Cambridge University Press