Eerste deeltoets DAR

27 mei 2016

11:00 - 13:00 Educ Megaron

- Vermeld op elk vel je naam en studentnummer.
- Toon bij het inleveren je collegekaart.
- Schrijf en formuleer duidelijk.
- Bij elke vraag wordt verwacht dat je laat zien hoe je aan het antwoord komt (tenzij anders wordt vermeld).
- Je mag een spiekbriefje (A4) raadplegen, maar verder niets.
- TA = Threshold Algorithm
  NRA = No Random Access Algorithm
- **Succes!**

1 MapReduce (10 punten)

Een *bag* is een collectie die duplicaten kent. De minus op bags lijkt op die van de sets, maar houdt rekening met duplicaatfrequenties. Voorbeeld:

$$\{e, c, a, b, c, b, a, e\} - \{d, e, b, a, e, e\} = \{a, b, c, c\}$$

De representatie van bags is vrij van volgorde en groepering, dus de uitkomst had net zo goed $\{c, b, a, c\}$ kunnen zijn.

Schrijf pseudocode voor Map en Reduce om de bag minus te implementeren. Voor elke Map worker staat een input file klaar die de volgende structuur heeft.

$$e, c, a, b, c, b, a, e, \#, d, e, b, a, e, e$$

Elke Map worker werkt op een gedeelte van de linkeroperand en op een gedeelte van de rechteroperand. Het speciale symbool # scheidt de elementen van de linker operand van die van de rechter operand.

Geef een korte toelichting bij je code. Uiteraard is performance zeer belangrijk.

2 Top-k (10 punten)

De domeinen van attributen A, B zijn niet-negatieve gehele getallen. De doelfunctie $F = A + B$ moet gemaximaliseerd worden.

OID	**A**
5	200
3	160
2	120
1	120
4	80
...	...

OID	**B**
2	200
5	160
1	160
3	40
4	40
...	...

Beide kolommen zijn gesorteerd. Op de OID van de linkerkolom (A) is een zoekindex beschikbaar, maar niet op de rechter.

(i) Welk algoritme kunnen we hier toepassen voor top-k berekeningen? TA, NRA, of kun je een eigen variant bedenken?

(ii) Bereken twee ronden van het gekozen algoritme. Geef aan welke waarden de variabelen Max-A, Max-B en Threshold na elke ronde hebben. Geef ook aan hoe de buffer er na elke ronde uitziet. Geef tenslotte na elke ronde aan welke top-k berekend is.

(iii) Stel nu dat de te maximaliseren doelfunctie $F = A - B$ is. Zie je nog steeds mogelijkheden voor toepassing van een top-k algoritme? Licht kort toe.

3 Google Pagerank (8 punten)

Ter herinnering:

$$G = \alpha S + (1 - \alpha)T$$
$$S = H + \frac{1}{n} e a^T$$
$$T = \frac{1}{n} e e^T$$

G is geen sparse matrix. Hoe kunnen we toch sparse matrix technieken gebruiken voor het fix point algoritme?

4 Sequence alignment (12 punten)

Hier zie je een fragment van een BLOSUM-matrix.

	D	E	K	V	I	L
D	8	2	-1	-4	-4	-4
E	2	6	1	-3	-4	-3
K	-1	1	6	-3	-3	-3
V	-4	-3	-3	5	4	1
I	-4	-4	-3	4	5	2
L	-4	-3	-3	1	2	5

De gap penalty is $-gd$, waarin g de lengte van de gap is en $d = 8$.

(i) Bereken met behulp van Smith-Waterman een optimale local match tussen de strings EVIL en KILL. Geef de complete dynamic programming matrix.

(ii) Iemand komt op het idee om Smith-Waterman te implementeren met MapReduce. Zie je mogelijkheden of zie je vooral problemen? Licht toe (maximaal 100 woorden).

In de meest eenvoudige vorm zou je de volgende heuristiek voor sequence alignment kunnen hanteren: op elke plaats waar een 3-gram match is tussen de query string en een string in je database, ga je proberen om via een lokaal expansie-algoritme een goede match vast te stellen.

(iii) Beschrijf twee uitbreidingen op het hierboven beschreven principe die in Protein BLAST gebruikt worden om de precision en de recall te veranderen. Geef voor elk van de methoden aan wat het effect is op precision en recall.

(iv) Bedenk zelf twee variaties die in de context van Protein BLAST toegepast zouden kunnen worden om precision en recall te veranderen.

Einde