Интеллектуальные информационные системы

MapReduce TF-IDF

Кафедра управления и интеллектуальных технологий НИУ «МЭИ»

2023 г.

TF-IDF

Term frequency – inverse document frequency

$$x_j^{(i)} = f_{ij} \log\left(\frac{N}{N_i}\right)$$

f_{ij} – частота термина i в документе j

N – количество документов

N_i - количество документов, в которых встречается термин i

TF-IDF

Требуется 3 MapReduce задачи:

1) Подсчитать частоту термина i в документе j (tf)
2) Подсчитать, в скольких документах i-й термин встречается (idf)
3) Перемножить tf на idf

TF-IDF. Job 1. Mapper

Sample Input:
1: jingle bells
2: jingle bells
2: jingle all the way

Sample output:

jingle#1	1
bells#1	1
jingle#2	1
bells#2	1
jingle#2	1
all#2	1
the#2	1
way#2	1

```
Method Map(docid id, doc d):
      for each term t in doc d:
            Emit ((term t, docid id), count 1)
```

TF-IDF. Job 1. Reducer

Sample input:

all#2	1
bells#1	1
bells#2	1
jingle#1	1
jingle#2	1
jingle#2	1
the#2	1
way#2	1

Sample output:

all	2	1
bells	1	1
bells	2	1
jingle	1	1
jingle	2	2
the	2	1
way	2	1

```
Method Reduce ((term t, docid id), counts [c1, c2,..]):
      tf= 0
      for each c in counts [c1, c2,..] :
            tf += c
      Emit (term t, (docid id, tf))
```

TF-IDF. Job 2. Mapper

Sample input:

all	2	1
bells	1	1
bells	2	1
jingle	1	1
jingle	2	2
the	2	1
way	2	1

Sample output:

all	2	1	1
bells	1	1	1
bells	2	1	1
jingle	1	1	1
jingle	2	2	1
the	2	1	1
way	2	1	1

```
Method Map((term t, docid id), tf):
          Emit (term t, (docid id, tf, 1))
```

TF-IDF. Job 2. Reducer

Sample input:

all	2	1	1
bells	1	1	1
bells	2	1	1
jingle	1	1	1
jingle	2	2	1
the	2	1	1
way	2	1	1

Sample output:

all#2	1	1
bells#1	1	2
bells#2	1	2
jingle#1	1	2
jingle#2	2	2
the#2	1	1
way#2	1	1

```
Method Reduce(term t, (docid id, tf, 1)):
      words, values =[],[]
      for line in (term t, docid id, tf, 1):
            words = words.append[line.t]
            values = values.append[(line.t, line.id, line.tf)]
      for val in values:
            Emit ((val.t, val.id), (val.tf, words.count(val.t)))
```

TF-IDF. Job 3. Mapper

Sample input:

all#2	1	1
bells#1	1	2
bells#2	1	2
jingle#1	1	2
jingle#2	2	2
the#2	1	1
way#2	1	1

```
Method Map ((term t, docid id), (tf, doc_count dc))):
       Emit ((term t, docid id), tfidf tf*log(N/dc))
```