Distributed TFIDF

Andrew Sinclair
Kim Paterson
Bradley Zeller

What is TFIDF?

Term Frequency - Inverse Document Frequency is a measure of a word's importance to a document in a corpus.

$$\mathrm{tf}(t,d) = 0.5 + \frac{0.5 \times \mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}} \qquad \mathrm{idf}(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

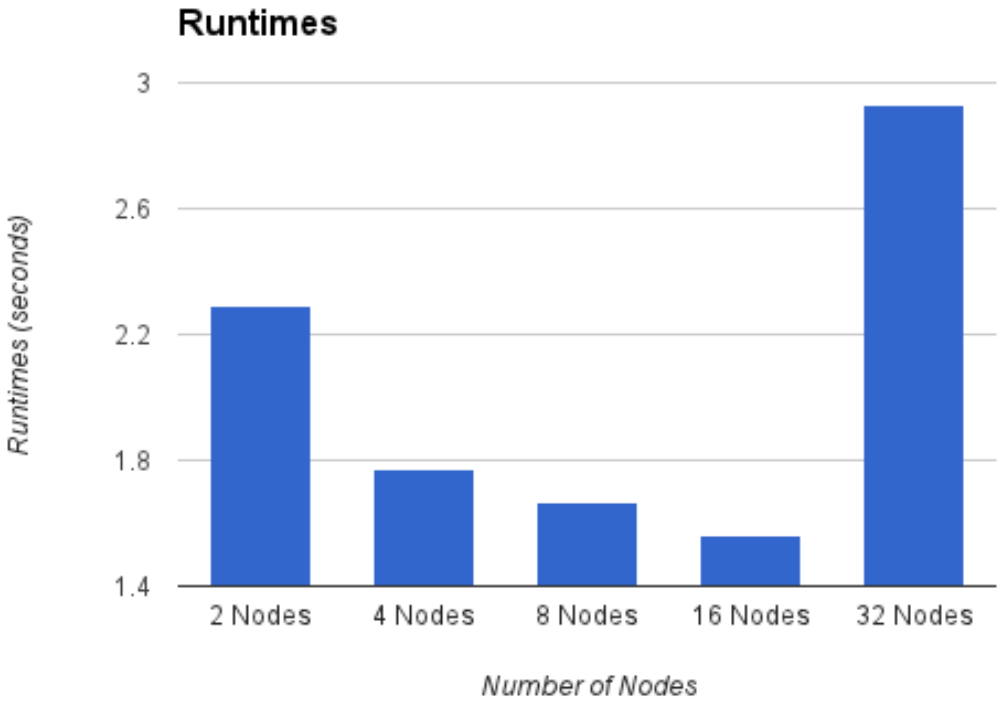
How We Distributed it

- Divide files between nodes
- Compute Term Frequency for each file
- Send document counts to master
- Master Broadcasts global document counts
- Calculate normalized TF
- Calculate TF*IDF for each word in each file
- Output results for each file

Experiment

- 62 documents
- 2, 4, 8, 16, 32 nodes
- Average of 3 runs

Performance (62 Documents)



Results

The Bible	Hamlet	Moby Dick	ca36
GOD,0.505747 Levites,0.505323 cubits,0.504692 commandments, 0.504183 Zion,0.503965 Jeremiah,0.503916 Manasseh,0.503892 Joab,0.503868 goeth,0.503746	Ham,0.606055 Qu,0.556407 Laer,0.554655 Ophe,0.551152 Pol,0.545022 Rosin,0.539767 Horatio,0.53714 Polon,0.535388 Clo,0.527506 Ophelia,0.526631	Stubb,0.516106 Queequeg,0.515942 Starbuck,0.512868 Pequod,0.51155 Sperm,0.509519 Flask,0.507818 Nantucket,0.506775 Moby,0.506665 whalemen,0.506006 Pip,0.505842	monopoly,0.699728 electrical,0.687629 commerce,0.67553 commodity,0.67553 giants,0.67553 monopolies,0.67553 abuses,0.66948 aimed,0.66948 allegedly,0.66948 argue,0.66948

Interesting….

ca 36	more…...
a,0.0 an,0.0 and,0.0 as,0.0 at,0.0 be,0.0 by,0.0 for,0.0 from,0.0 has,0.0 have,0.0 in,0.0 is,0.0 it,0.0	of,0.0 on,0.0 one,0.0 out,0.0 s,0.0 that,0.0 the,0.0 to,0.0 was,0.0 which,0.0 who,0.0 will,0.0 with,0.0