Problem Set 2: Semantic Similarity

Renae Fisher

CS 4903: Text Mining – Professor Mackey

July 26, 2019

Methodology

BuildTermContextMatrix

Time complexity	$O(DT \log V) + O(DT \cdot W^2) + O(V^2)$
Space complexity	$O(V) + O(V^2)$

The method `buildTermContextMatrix` relies on `getVocab` to find the number of distinct terms. The size of `vocab` is used as the dimensions of the $|V| \times |V|$ term-context matrix. The method `getVocab` also builds an array of terms, which the program later uses to map terms to the indices of the term-context matrix. The method `getVocab` first uses a `TreeSet` to build an ordered set of V distinct terms. It takes $O(DT \log V)$ time to add the T terms in D documents to the ordered set. After it's done reading all files, the method transfers the contents of the `TreeSet` to an `ArrayList`. It takes $O(V)$ time to fill the `ArrayList`. Altogether, the method has a runtime of $O(DT \log V + V)$, or $O(DT \log V)$. It has a space complexity of $O(V)$, because it only stores the distinct words it identifies.

The method `buildTermContextMatrix` first uses $O(V^2)$ space for the term-context matrix. It also takes $O(V)$ space to temporarily store the total frequencies for each term, which `weightTerms` uses to calculate PPMI. Next, `buildTermContextMatrix` starts to iterate across the files in the input directory. As it reads the tokens in a document, it builds a segment of size W, the window size. This task takes $O(W)$ space total. When the segment is complete, the method passes this segment and the prior one to `countTerms`. The method `countTerms` uses two loops to find the distinct pairs of contextual terms within the two segments. The outer loop in `countTerms` runs from a to w, and the inner loop runs from zero to $w - a$. Its total runtime is

$O(W^2)$. Together, the two methods take $O(DT \cdot W^2)$ time to create the term-context matrix. After the term-context matrix is complete, `weightTerms` takes the matrix and the aggregated frequencies in `sum`. It uses two loops to calculate weights for each cell in the term-context matrix. This takes $O(V^2)$ time.

CalculateSimilarity

Time complexity	$O(V)$
Space complexity	$O(3)$

The method `calculateSimularity` uses a single for loop and three variables to calculate the cosine simularity for two words. This takes $O(V)$ time.

GetContext

Time complexity	$O(V^2 \cdot \log V)$
Space complexity	$O(V) + O(k)$

The method `getContext` uses a single for loop to iterate over the entire vocabulary V. For each term, it uses `calculateSimularity` to find the consine similarity between the two words. Then, it stores an object with the row number and cosine similarity in a priority queue. This action takes $O(\log V)$ time and $O(V)$ space. After `getContext` examines every word in the vocabulary, it uses a loop to store the top k words in the priority queue.

Results

The results below demonstrate the output of the application at a window size of four, eight, and sixteen for given queries.

CalculateSimularity

Data, Computer	
real	1m01.855s 1m34.686s 2m31.227s
user	1m09.082s 1m42.089s 2m38.800s
sys	0m26.240s 0m27.483s 0m26.011s
4	0.1173
8	0.1095
16	0.1160

Data, Dog	
real	1m02.029s 1m29.595s 2m31.415s
user	1m09.463s 1m36.940s 2m38.664s
sys	0m26.487s 0m26.451s 0m26.443s
4	0.0
8	0.0
16	0.0

Data, Pencil	
real	0m58.257s 1m29.073s 2m32.061s
user	1m05.531s 1m36.252s 2m38.923s
sys	0m26.708s 0m26.621s 0m27.515s
4	0.0
8	0.0
16	0.0

Hot, Dog	
real	1m01.973s 1m29.131s 2m39.099s
user	1m09.499s 1m36.824s 2m46.906s
sys	0m26.802s 0m25.979s 0m26.133s
4	0.0156
8	0.0113
16	0.0092

GetContext

Computer	
real	1m02.276s 1m02.276s 2m36.153s
user	1m10.174s 1m10.174s 2m43.326s
sys	0m25.738s 0m25.738s 0m26.321s
4	lahor pakistan karachi islamaba scienc embed outsourc hadoop parallel etichet
8	islamaba lahor karachi pakistan scienc embed etichet parallel mooc neural
16	islamaba lahor karachi pakistan scienc parallel onthoud credenti embed whitepap

Data	
real	1m02.553s 1m38.405s 2m36.050s
user	1m10.140s 1m46.046s 2m43.234s
sys	0m26.243s 0m26.559s 0m26.772s
4	lahor pakistan karachi islamaba hadoop globalhe sentri wareh mine hatenabl
8	islamaba lahor karachi pakistan hadoop globalhe mine sentri analyt big
16	lahor islamaba karachi pakistan hadoop globalhe mine big hatenabl analyt

Pencil	
real	1m02.337s 1m38.154s 2m44.197s
user	1m10.161s 1m45.576s 2m52.020s
sys	0m25.955s 0m27.353s 0m25.786s
4	sharpen snail eyelin mascara lipstick crayon eras brow sketch eyeshado
8	sharpen snail lipstick mascara eyelin brow eyeshado lip crayon eyebrow
16	sharpen snail lipstick mascara eyelin eyeshado lip brow crayon bronzer

Dog	
real	1m02.736s 1m34.042s 2m46.206s
user	1m10.271s 1m41.399s 2m53.791s
sys	0m26.405s 0m26.754s 0m26.209s
4	puppi pet terrier breed leash hound kennel chow collar breeder
8	pet puppi terrier breed leash hound chow kennel beagl breeder
16	pet breed puppi terrier leash hound kennel flea chow beagl