Information Retrieval
Homework 1

Ondřej Měkota

November 26, 2019

Overview

- Run-0 – nothing interesting
- Run-[1-2] – search over 18k combinations of hyperparameters
- Document vector representation as sparse matrix

Sparse matrix

- Incidence matrix would be too large in dense form.
- For $\sim 500k$ words, $\sim 80k$ documents, single precision $\rightarrow$ 160 GB of raw data
- Lot of elements are zero $\rightarrow$ store only non-zeros
- **Solution 1:** store list of indices and data A[i,j] = x
- Memory efficient but slow row and columns slicing and multiplication

Sparse matrix 2

Solution 2

- Store the data row by row.
- Keep three arrays: *data*, *indices* with column index of each element, and *indptr*: `indptr[i] = #non-zero elements in rows [0, i)`
- Then n-th row is `data[indptr[n]:indptr[n+1]]`
- Fast multiplication, hence fast computation of cosine similarity.

Tables with results[1]

Run	Czech	English
Run-0	0.0567	0.0455
Run-1	0.3020	0.3516
Run-2	0.3156	0.3843

Table: map

Run	Czech	English
Run-0	0.0640	0.0720
Run-1	0.3480	0.3960
Run-2	0.3480	0.4280

Table: P_10

[1] `https://ui.neptune.ml/pixelneo/retrieval/experiments`