Carnegie Mellon University
10-709 Fall09: Reading the Web
Prof. Tom Mitchell

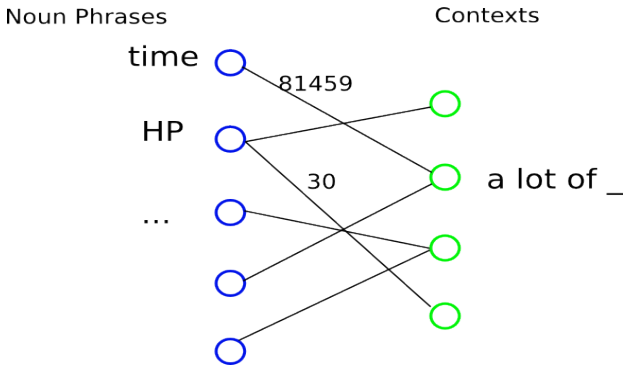
The NPIC500 Dataset

Description

The NPIC500[1] dataset was produced from a crawl of around 200 million webpages as part of the RTW project at *NELL*, CMU. The clauses in these pages were segmented using OpenNLP into noun-phrases (NP) and contexts (C), without keeping track of which document each NP or C belongs to. Instead, the number of times a certain NP and C occur together is recorded. The records were thresholded at 500 individual occurrences; i.e. if a NP or C did not occur (by itself) 500 times or more in the whole dataset, it is removed from the records and you don't see any of its co-occurrence counts.

Bipartite Graph View

A good way to think of the data in this dataset is a bipartite graph, where the NPs are on one side and the Cs are on the other, and the edge exists between a NP and a C when these two co-occur at least once. The weight of the edge is the number of co-occurrences.



Files

NPIC500 consists of 3 files:

- `nps.txt`: a text file listing of all NPs, each NP on a separate line. The line number in this file is used as the NP's Id in the file matrix.txt
- `contexts.txt`: similar to nps.txt, but listing contexts
- `matrix.txt`: a tab-separated text file of NP-C co-occurrence counts, where each line (entry) is of the form <npid> <cid> <coocc>, where all 3 values are integers.
    - This file is actually organized as a valid CCS (Compressed Column Storage) Matlab sparse matrix that can be loaded with `spconvert()` (see Getting Started below).
    - The uncompressed file size is 264 MB

Size

There are 88M distinct[2] NPs and 99M distinct Contexts in total, and 20M[3] co-occurrence counts (i.e.

1 Noun-Phrase in Context
2 Distinct as textual representations only. “IBM”, “IBM Incorp”, “International Business Machines” are different NPs
3 `wc -l matrix.txt`

matrix entries) in total. Hence, it's about 0.22% loaded. Matlab[4] should load this matrix with modest RAM requirements (1GB should give fine performance).

Getting Started

You can find with the dataset some useful Matlab utility functions to read the files. Each one exists in a separate file, and is briefly documented in its header.

An easy way to get started is:

```
$ less nps.txt
history
programming
all types
poetry
digital art
HP
...

$ less contexts.txt
...
A degree in _
A steady stream of _
Cars for _
Nothing in _
Paintings of _
Please keep this in _
Some members of _
...

$ less matrix.txt
...
62      15      2
1608    15      6
1609    16      2
1610    16      6
1611    16      3
1612    16      2
1613    16      2
294     16      9
1614    16      2
1       16      9
...
```

The above lines in the matrix say that

- NP (62) "*a variety*" and Context (15) "*Surrounded by* _" co-occurred 2 times (I.e. the clause "*Surrounded by a variety*" occurred 2 times.
- NP (1) "*history*" and Context (16) "*The burden of* _" co-occurred 9 times
- and so on

4 As well as GNU Octave

In Matlab

```
% Load the matrix as triples; gives a 20M x 3 matrix
T = load("/path/to/matrix.txt");

% see how big T is
size(T)

% convert to 88K x 99K NP-C matrix; columns 1 and 2 in T become the row and
% column indices, and col 3 the value of the entry in M
M = spconvert(T);
[m,n] = size(M);

% x should get the value 2
x = M(62,15);

% do all your fancy stuff
% ...

% what's the highest co-occurrence count of them all[5]?
C = T(:,3);
[row] = find(C==max(C));
T(row,:)
```

5 *"a lot of time"*