Experiments with Markov chains and Hansard

Danny Robson

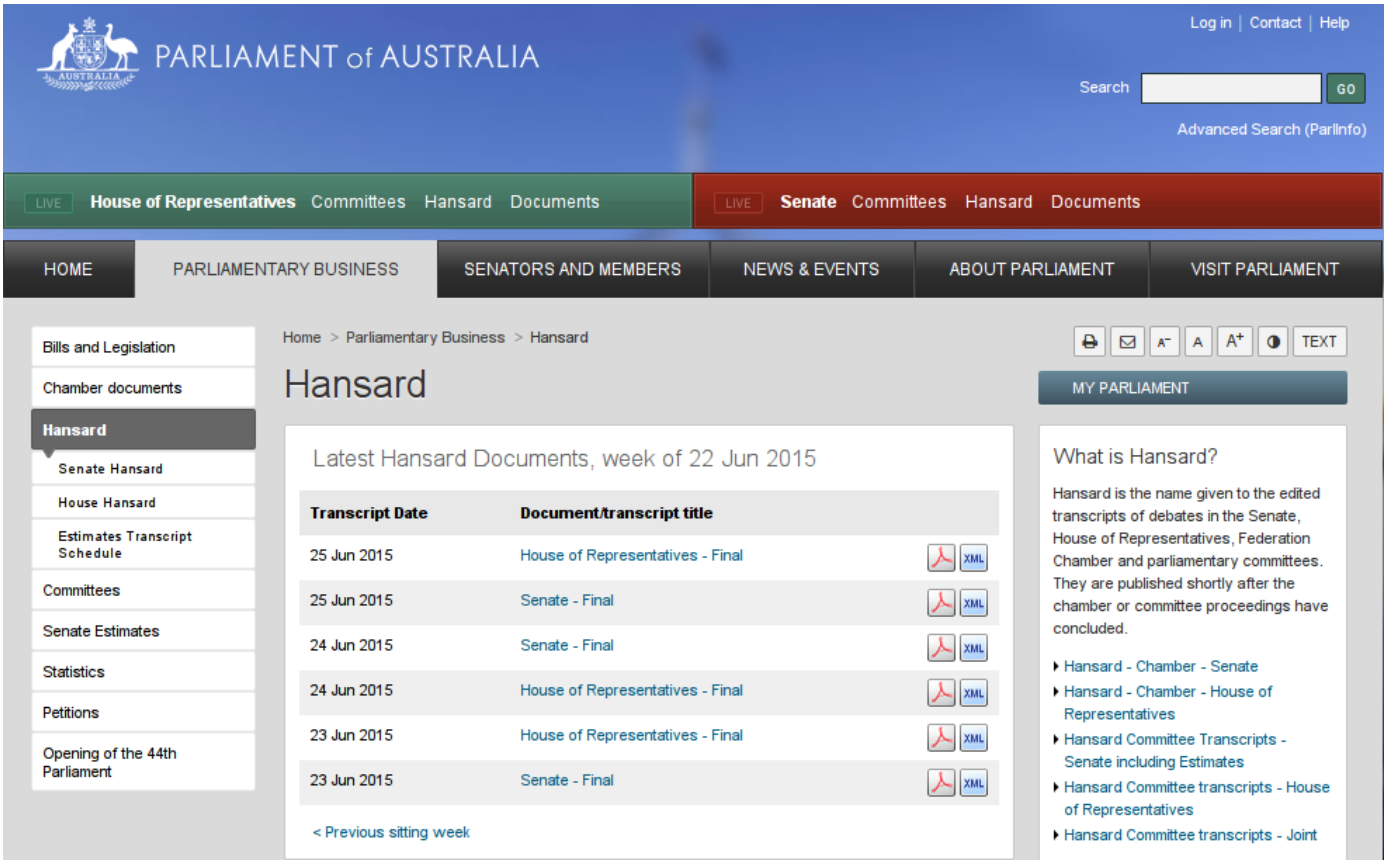
about

- Experiments motivated by
  - repetition in parliamentary debates
  - XKCD Robot9000/Signal IRC channel
  - *extremely* limited machine learning experience
- Purely for my own amusement

Hansard9000

1. detect repeated phrases
2. assign uniqueness scores per member
3. ...
4. profit?

Corpus



1. scrape the hansard website
2. save the XML versions
3. examine XML and cry

Convert to text

- text is easy to deal with, let's discard most/all markup
- XPath to the rescue!

```
for i in data/*
do
    xmlstarlet sel -t -v
"//talk.text/body/p" < "$i" >>
./text;
done
```

Repetition detection

- knock up some quick Python for repetition
  1. read each line
  2. tokenise into n-grams
  3. write counts and visualise later

Repetition output

```
19 ('stop', 'the', 'boats')
39 ('to', 'climate', 'change')
915 ('the', 'carbon', 'tax')
1426 ('the', 'prime', 'minister')
2861 ('the', 'member', 'for')
```

- Hmm, somewhat amusing… But not what I'd hoped.
- Turns out parliament is highly formulaic.

months pass...

New Motivation

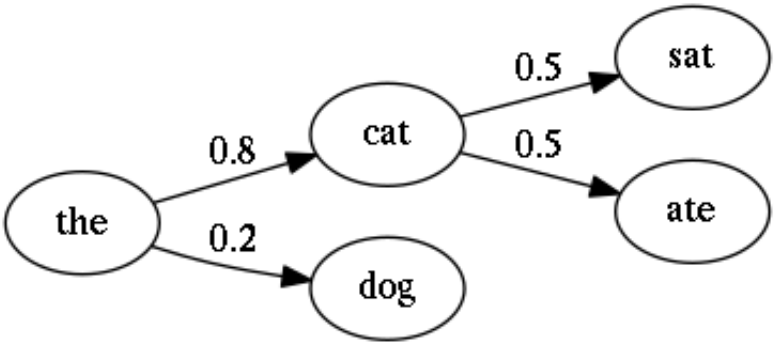
I'm going to do a quick talk about how I installed Linux on chromebook. If anyone else can do a talk, even a lightening talk let me know.

See you guys tomorrow night at The Dan.

from
Mick

Markov Chains

- Machine learning seems to be popular right now, how hard can it be…
- ‘That spam text generator’



- Better results with more history, but I’m lazy.

Development

1. analyse
    a. tokenise the text file
    b. record the successor for each token
2. synthesise
    a. pick a start token
    b. print the token
    c. randomly pick a successor
    d. repeat

Output

```
Stop Live cattle are essentially
self-medication for centenary
national trade negotiating peace
process. We already sorted
yourselves out going on? The third
country. He raised throughout a day
—and new criteria to detract unduly
constrained. However, my travel.
$945.08 for bad days, 54 per
second, total family business, when
Treasury's economic challenges
which 224 boats usually around its
omission from research funding
contributes 1.4 million lost 200
people? Why wouldn't know those
sorts of Legal Affairs, to pride
and trust placed upon as Treasurer,
he missed it.
```

- Not bad for 30 minutes development at midnight Sunday.
  - The most heinous code I've written in many years.

- However it’s clearly nonsense text.

Future

- Use more history
- Better utilise the XML information
  - Find/derive schemas
- Finish Hansard9000 scoring system