Making the Brown Corpus Readable

Here's a simple example of developing a command line removing the tags from the Brown corpus files.

In [1]:
```
!ls brown/. | head
```

```
CONTENTS
README
ca01
ca02
ca03
ca04
ca05
ca06
ca07
ca08
```

We should probably look at the README for the definition of the tag file, but let's just figure this out.

In [2]:
```
!head brown/README
```

```
BROWN CORPUS

A Standard Corpus of Present-Day Edited American
English, for use with Digital Computers.

by W. N. Francis and H. Kucera (1964)
Department of Linguistics, Brown University
Providence, Rhode Island, USA

Revised 1971, Revised and Amplified 1979
```

Here's the first 10 lines from the file `brown/ca07`.

In [3]:
```
!sed 10q brown/ca07
```

```
	Resentment/nn welled/vbd up/rp yesterday/nr among/in Democratic/jj-tl
district/nn leaders/nns and/cc some/dti county/nn leaders/nns at/in reports/nns
that/cs Mayor/nn-tl Wagner/np had/hvd decided/vbn to/to seek/vb a/at third/od
term/nn with/in Paul/np R./np Screvane/np and/cc Abraham/np D./np Beame/np as/cs
running/vbg mates/nns ./.

	At/in the/at same/ap time/nn reaction/nn among/in anti-organization/jj
Democratic/jj-tl leaders/nns and/cc in/in the/at Liberal/jj-tl party/nn to/in
the/at Mayor's/nn$-tl reported/vbn plan/nn was/bedz generally/rb favorable/jj ./.

	Some/dti anti-organization/jj Democrats/nps saw/vbd in/in the/at
program/nn an/at opportunity/nn to/to end/vb the/at bitter/jj internal/jj
fight/nn within/in the/at Democratic/jj-tl party/nn that/wps has/hvz been/ben
going/vbg on/rp for/in the/at last/ap three/cd years/nns ./.
```

The main thing is that every word or punctuation is followed by a `/something`. We can remove that with a simple regular expression. Well, it's not quite so simple...

- We want to replace `/`, but that's already the regular expression delimiter, so we need to escape it: `\/`
- the g is needed because we want to replace all occurrences

In [7]:
```
!sed 's/\/[^ ]*//g;10q' brown/ca07
```

```
	Resentment welled up yesterday among Democratic district leaders and
some county leaders at reports that Mayor Wagner had decided to seek a third
term with Paul R. Screvane and Abraham D. Beame as running mates .

	At the same time reaction among anti-organization Democratic leaders and
in the Liberal party to the Mayor's reported plan was generally favorable .

	Some anti-organization Democrats saw in the program an opportunity to
end the bitter internal fight within the Democratic party that has been going on
for the last three years .
```

Let's now clean up the whitespace at the beginning of the line. `\t` is a shorthand for the tab character.

In [8]:
```
!sed 's/\/[^ ]*//g;s/^[ \t]*//;10q' brown/ca07
```

```
Resentment welled up yesterday among Democratic district leaders and some county
leaders at reports that Mayor Wagner had decided to seek a third term with Paul
R. Screvane and Abraham D. Beame as running mates .


At the same time reaction among anti-organization Democratic leaders and in the
Liberal party to the Mayor's reported plan was generally favorable .


Some anti-organization Democrats saw in the program an opportunity to end the
bitter internal fight within the Democratic party that has been going on for the
last three years .
```

There are a lot of blank lines; the `cat -s` (squeeze) command gets rid of them.

In [9]:
```
!sed 's/\/[^ ]*//g;s/^[ \t]*//;10q' brown/ca07 | cat -s
```

```
Resentment welled up yesterday among Democratic district leaders and some county
leaders at reports that Mayor Wagner had decided to seek a third term with Paul
R. Screvane and Abraham D. Beame as running mates .

At the same time reaction among anti-organization Democratic leaders and in the
Liberal party to the Mayor's reported plan was generally favorable .

Some anti-organization Democrats saw in the program an opportunity to end the
bitter internal fight within the Democratic party that has been going on for the
last three years .
```

Now we still have a problem with extra spaces before punctuation. We can fix that with another regular expression. This one contains *grouping* `\(...\)` and a backwards reference to the group `\1`

In [13]:
```
!sed 's/\/[^ ]*//g;s/^[ \t]*//;s/ \([.,]\)/\1/;10q' brown/ca07 | cat -s
```

```
Resentment welled up yesterday among Democratic district leaders and some county
leaders at reports that Mayor Wagner had decided to seek a third term with Paul
R. Screvane and Abraham D. Beame as running mates.

At the same time reaction among anti-organization Democratic leaders and in the
Liberal party to the Mayor's reported plan was generally favorable.

Some anti-organization Democrats saw in the program an opportunity to end the
bitter internal fight within the Democratic party that has been going on for the
last three years.
```

Finally, let's wrap the long lines back around. The `fmt` command is handy for that.

In [14]:
```
!sed 's/\/[^ ]*//g;s/^[ \t]*//;s/ \([.,]\)/\1/;10q' brown/ca07 | cat -s | fmt
```

```
Resentment welled up yesterday among Democratic district leaders and
some county leaders at reports that Mayor Wagner had decided to seek a
third term with Paul R. Screvane and Abraham D. Beame as running mates.

At the same time reaction among anti-organization Democratic leaders and
in the Liberal party to the Mayor's reported plan was generally favorable.

Some anti-organization Democrats saw in the program an opportunity to
end the bitter internal fight within the Democratic party that has been
going on for the last three years.
```