F2

Words that are present in both groups but have a considerably different frequency count

- equalize corpus sizes (now: compareSizeOfCorpora)

- choose a threshold to determine 'considerable difference' (now: freq1 > 3 x freq2)

- get rid of very low counts (now: freq > 5)

- determine how much more frequent the word is in each case

- features: prevalentL, prevalentR

- prevalentL + prevalentR = 1

F3

Words that belong only to one of the groups

choose the meaningful words manually?

keep words only with relatively high frequency?

use all and see how it reflects on accuracy? (now)

- get rid of singletons (now)

F4

Names: frequency of proper names

- need to select names (started in f4\_marked\_words\_\*)

F5

foxnews, fox + prefix/suffix fox

F6

'democrat party' / 'Democratic national convention'