1. Words that are the same in both groups but have a considerably different frequency count

need to: equalize corpus sizes

2. Words that belong only to one of the groups

need to: choose the meaningful words manually / keep words only with relatively high frequency / use all and see how it reflects on accuracy

3. Names: frequency of proper names (need to select names)

4. foxnews, fox + prefix/suffix fox

5. 'democrat party' / 'Democratic national convention'