TUNE YOUR BROWN CLUSTERING, PLEASE

LEON DERCZYNSKI, SEAN CHESTER, KENNETH S. BØGH

INTRO

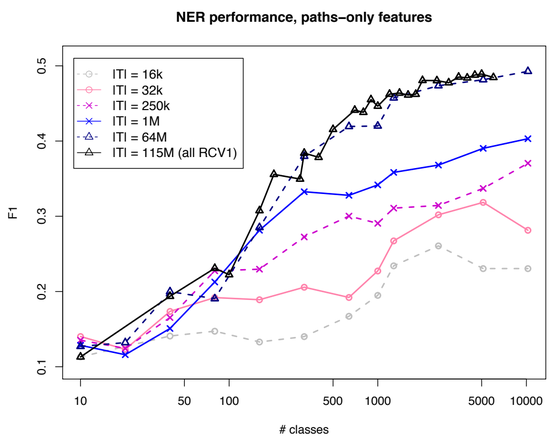
Brown clustering takes a corpus and outputs c clusters of word types

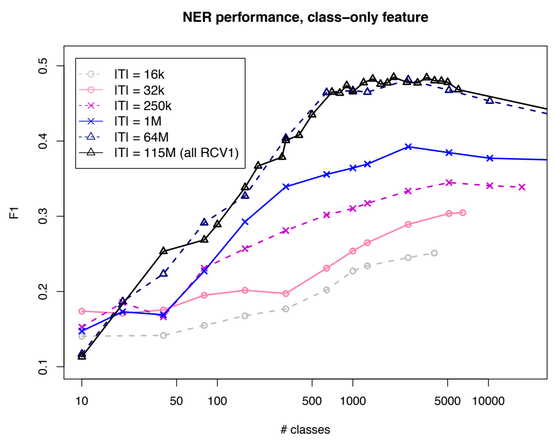
These clusters are placed in a tree

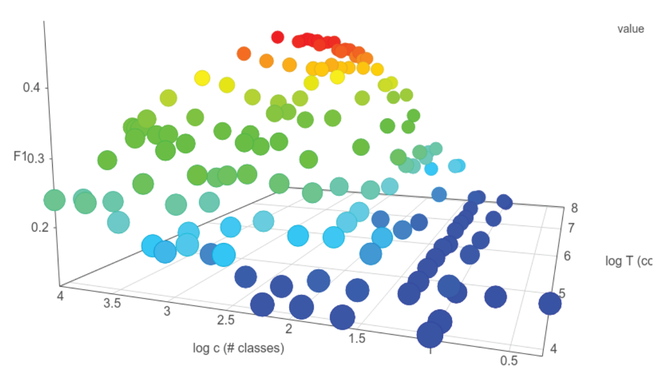
THIS LEARNS A REPRESENTATION OF THE TEXT, WHICH IS MUCH MORE POWERFUL THAN E.G. BAG-OF-WORDS

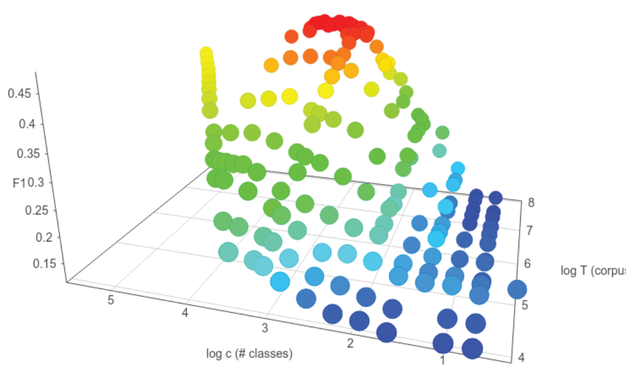
But how do we know how many clusters to generate?
Too few, and too many, are both harmful

LET'S EVALUATE BROWN CLUSTERS, USING NAMED ENTITY RECOGNITION AS THE TASK:









PRACTICAL ADVICE:

- AVOID DEFAULT VALUES OF C
- Getting a big corpus is more helpful than generating many clusters – usually!
- If you care more about tree information, make c high
- If you care more about how words cluster together, don't make c very high
- Try random search for c, weighted away from very low and high values of c
- To save time, start your parameter search using some of our pre-generated clusterings