Project #1: Scalable Document Classification

Team Alias - Shubhi, Priyanka, Narita

Approach

- Multinomial Naive Bayes
- Features: Word Counts conditioned on class
- Smoothing: Laplace (add-1)
- Stopwords
- Referred: https://web.stanford.edu/class/cs124/lec/naivebayes.pdf
- Custom Stopwords:
    - Antoine Blanchard, Understanding and customizing stopword lists for enhanced patent mapping, World Patent Information, Elsevier, 2007, 29 (4), pp.308. <10.1016/j.wpi.2007.02.002>

    - Zipf's Law: The frequency of any word is inversely proportional to its rank. Luhn (1958) suggested that both extremely common and extremely uncommon words were

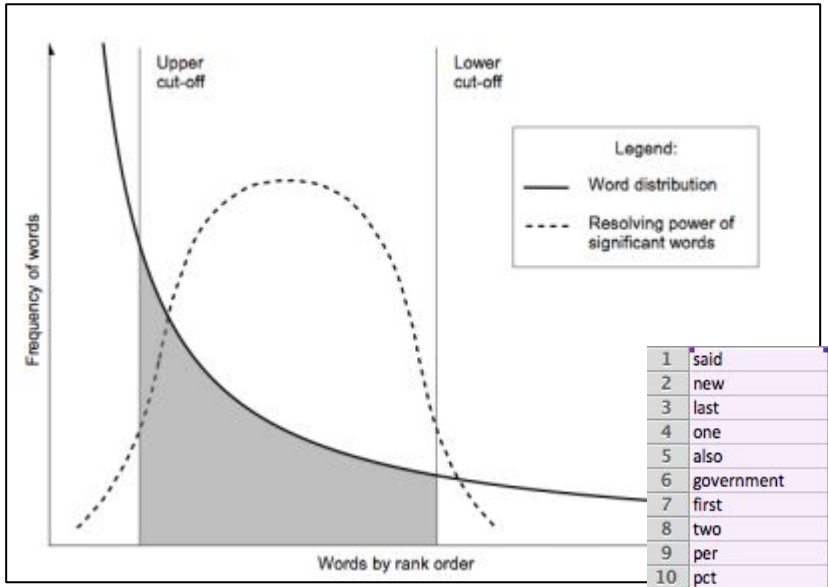

Fig. Plot of word distribution

Reference: Antoine Blanchard, Understanding and customizing stopword lists for enhanced patent mapping, World Patent Information, Elsevier, 2007, 29 (4), pp.308.

1	said	2252940
2	new	388471
3	last	296240
4	one	295703
5	also	275403
6	government	265978
7	first	248186
8	two	245574
9	per	201201
10	pct	198194
11	prices	182603
12	sales	175733
13	group	175427
14	monday	173206
15	told	172623
16	friday	169889
17	trade	168790
18	price	165388
19	three	164030
20	years	160694
21	minister	152502
22	stock	150920
23	president	149033
24	state	146294
25	rate	144896
26	months	142382