Contents

1 Problem 3

Note: I was having issues with using the entire $\text{masc}_{\text{tagged}}$ corpus, so I opted to use only the category 'blog', which seemed to work for whatever reason

Words Total: Unigrams:33319 Bigrams:31556 Trigrams:29815

POS Total: Unigrams:33319 Bigrams:31556 Trigrams:29815

	Train types	Test Types	Not in tr	pct not in tr	over all pct
Unigrams	5655	1202	493	41%	14%
$\text{POS}_{\text{Unigrams}}$	49	46	1	2.1%	0%
Bigrams	19394	2686	2023	75.3%	63%
$\text{POS}_{\text{Bigrams}}$	963	503	32	6.36%	1%
Trigrams	24275	2866	2735	95.42%	90%
$\text{POS}_{\text{Trigrams}}$	5467	1561	349	22.3%	11%

a) It seems like training an NLP using just words (espeically word unigrams) is a poor idea. This is because word unigrams are really just a measure of the vocabulary used by a text. Obviously things like singeltons are incredibly useful when trying to determine authorship or category membership, but for general lingusitic competancy (e.g. prediction) word unigrams do not seem to be very useful. Indeed, even though the training set is many times larger than the testing set, 41% of unigram types in the testing set were not present in the training set. That being said, in terms of words, compared to bigrams and trigrams, the testing and training sets showed the most overlap with unigrams

b)The bigger n is (assuming n represents the size of the corpus) the more likely you are to see word types repeat. This means that as n increases there will be more overlap in word type between training and testing sets. HoweverNevertheless, as n increases, the number of novel words introduced into the corpus will increase as well, however I believe the number of types will level out with sufficiently large ns (given the limited number of words in English).

c) The POS unigrams are similar to pos words in that the training and testing sets share the most unigrams. They differ in the particulars: the training and testing set share nearly 98% of pos unigram types vs 59% word

unigrams. This is probably because the universe of POS is much smaller than the universe of words.