South Park Methodology

Kaylin Walker

February 2, 2016

Intro

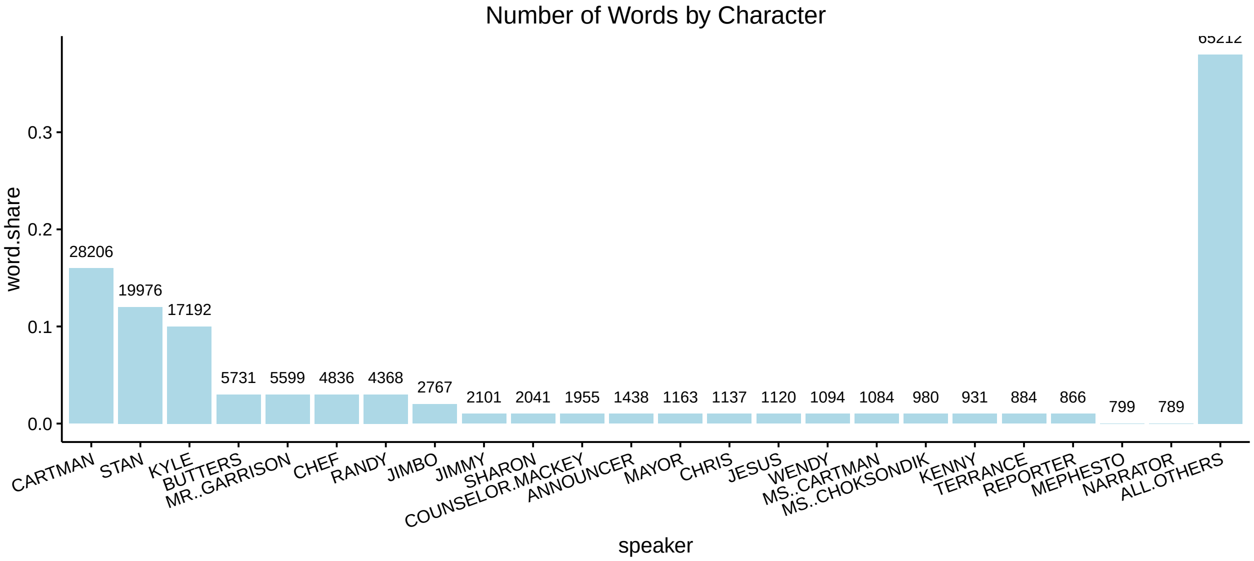
Method & Summary Statistics

Transcripts of the first 132 episodes of South Park (season 1 through season 9, episode 7) were scraped from unstructured text on the Internet Movie Script Database using the XML package. The series has run for 19 seasons and 267 episodes, but to my knowledge the remaining transcripts are not available online.

I was able to assign a speaker to each line by splitting the html at `<b>` tags that contained only uppercase text and were shorter than 40 characters. Each line was followed by a blank line, so I used their index to create a starting and stopping point for text to be attributed to a speaker. From there, I used the tm package to pre-process the text (to lowercase, remove punctuation, numbers and whitespace; remove stopwords for unigrams and bigrams, but left them in for tri-, 4- and 5-grams) and form a corpus, which contained more than 18,000 unique words spoken more than 211,000 times. Reducing the sparsity brought that down to about 2,300 words spoken 172,000 times.

ngram.size	total	unique
1	172269	2349
2	140679	3026
3	19313	662
4	4925	360
5	1334	87

24 characters with the most words were retained, and the remaining 1781 speakers combined into one "all others" category so as not to lose the text.



Log Likelihood

Each corpus was analyze to determine the most characteristic words for each speaker. Frequent and characteristic words are not the same thing - otherwise words like “I”, “school”, and “you” would rise to the top instead of unique words and phrases like “professor chaos”, “hippies” and “you killed kenny.”

Log likelihood was used to measure the unique-ness of the n-grams by character. Log likelihood compares the occurrence of a word in a particular corpus (the body of a character’s speech) to its occurrence in another corpus (all of the remaining South Park text) to determine if it shows up more or less likely that expected.

The **chi-square test (χ^2)**, or goodness-of-fit test, can be used to compare the occurrence of a word across corpora.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O = observed frequency and E = expected frequency.

However, flaws have been identified: invalidity at low frequencies (Dunning, 1993) and over-emphasis of common words (Kilgariff, 1996). Dunning was able to show that the **log-likelihood statistic** was accurate even at low frequencies:

$$2 \sum O_i * ln(\frac{O_i}{E_i})$$

Which can be computed from the contingency table below as $2 * ((a * log(\frac{a}{E1}) + (b * log(\frac{b}{E2}))$, where E1 = $(a + c) * \frac{(a+b)}{(c+d)}$, and E2 = $(b + d) * \frac{(a+b)}{(c+d)}$.

Table 2: Basic Framework

Group	Corpus.One	Corpus.Two	Total
Word	a	b	a+b
Not Word	c	d	c+d
Total	a+c	b+d	N=a+b+c+d

Table 3: An Example

Group	Cartmans.Text	Remaining.Text	Total
‘hippies’	36	5	41
Not ‘hippies’	28170	144058	172228
Total	28206	144063	172269

Computed:
E1 = 28206 * (41/172269) = 6.71 E2 = 144063 * (41/172269) = 34.28
LL = 2 * [36 * log(36/6.71) + 5 * log(5/34.28)] = 101.7

Based on the overall ratio of the word “hippies” in the text, 41/172269 = 0.00023, we would expect to see hippies in Cartman’s speech 6.71 times and in the remaining text 34.28 times. The log likelihood value of 101.7 is significant far beyond even the 0.01% level.

Table 4: Log Likelihood Significance Levels

Level	Critical.Value	P.Value
5%	3.84	0.05
1%	6.63	0.01
0.1%	10.83	0.001
0.01%	15.13	.0001

Significance of Log Likelihood Scores

5.64% (1657/29405) of ngrams were found to be significantly characteristic for a ceratin character.

References

Dunning, T. (1993) *Accurate Methods for the Statistics of Surprise and Coincidence.* Computational Linguistics, 19, 1, March 1993, pp. 61-74.
Kilgarriff. A. (1996) *Why chi-square doesn't work, and an improved LOB-Brown comparison.* ALLC-ACH Conference, June 1996, Bergen, Norway.