HW6: language model ranking

Part 4: impressionistic evaluation

The following results are based on a probabilistic 6-gram language model of the English editions of the WMT News Crawl. The 2008 data was used as training, and the 2009 as test sets.

The scores were calculated using entropy, which provides the average amount of (bits of) information needed to represent an event based on the probability distribution of a random variable. The events, in this case, are text strings.

The entropy scores were analyzed based on the bit of relevant information contained in the text strings. Low entropy is associated with unsurprising events, while higher entropy is a sign of surprising events.

The lowest entropy

The lowest entropy scores are linked to low interest sentences, such as boilerplates. Boilerplates can be described as reusable chunks of sentences, with little changes to the original[1]. In this context, they are "canned" texts, automatically generated to compose the media web pages - such as navigation bars, page headers, link lists, and some advertisements. They are not original, but formulaic.

We can see logo links, dates or subpages information, linke "WASHINGTON, May 8 /PRNewswire-FirstCall/ –".

These types of sentences were mostly common in scores from 0.8 to 1.36.

Table 1: Low entropy

Entropy	Sentence
0.800239	Logo: http://www.newscom.com/cgi-bin/prnh/20080519/RNCLOGO
0.820111	WASHINGTON, May 8 /PRNewswire-FirstCall/ --
0.832520	NEW YORK, July 16 /PRNewswire-FirstCall/ --
0.836913	WASHINGTON, March 19 /PRNewswire-USNewswire/ --
0.840588	CINCINNATI, April 29 /PRNewswire-FirstCall/ --
0.841357	CINCINNATI, April 23 /PRNewswire-FirstCall/ --
0.849048	Hillary Rodham Clinton.

[1] Wikipedia: https://en.wikipedia.org/wiki/Boilerplate_text

0.850457	21 /PRNewswire-USNewswire/ --
0.854003	21 /PRNewswire-FirstCall/ --
0.856434	12 /PRNewswire-FirstCall/ --

The highest entropy

As mentioned in the homework instructions, higher entropy *"are often of low-quality, and may even reflect errors in data collection or text encoding"*. In this case, we found random characters and atypical natural language sentences, such as "dU\n" and "d\n".

These types of sentences were mostly common in scores greater than 4.

Table 1: High entropy

Entropy	**Sentence**
24.257831	"dU\n
25.540480	"d\n
25.847289	"h\n
25.997927	"a\n
26.036764	"l\n
27.777818	"c\n
28.270156	"x\n
29.024469	"i\n
30.382851	"v\n
31.569306	"j\n

Overall entropy

As mentioned in the homework, *"one can use bits-per-character ranking to select sentences of medium entropy for further annotation"*. In this analysis, the meaningful information was found in medium scores from 1.37 to 3.9.

Entropy level	**Entropy range**	**Sentence description**
Lower entropy	0.8 - 1.36	Boilerplate of little interest
Medium entropy	1.37 - 3.9	Meaningful information
Highest entropy	> 4	Low-quality texts