UNIVERSITY OF SURREY

Readability for author profiling?

Notebook for PAN at CLEF 2013

Lee Gillam, University of Surrey

UNIVERSITY OF SURREY

Performances on the English portion of the test data

Submission	Accuracy			Adult			Predator			Runtime
	Total	Gender	Age	Gender	Age	Both	Gender	Age	Both	(incl. Spanish)
meina13	0.3894	0.5921	0.6491	6	8	6	72	41	41	383821541
pastor13	0.3813	0.5690	0.6572	1	8	0	72	32	32	2298561
mechti13	0.3677	0.5816	0.5897	2	6	2	52	29	20	1018000000
santosh13	0.3508	0.5652	0.6408	9	9	9	69	32	29	17511633
yong13	0.3488	0.5671	0.6098	6	1	1	28	30	17	577144695
ladra13	0.3420	0.5608	0.6118	9	9	9	72	33	33	1729618
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
gillam13	0.3268	0.5410	0.6031	1	4	0	72	30	30	615347
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
haro13	0.3114	0.5456	0.5966	0	8	0	69	44	41	9559554
aditya13	0.2843	0.5000	0.6055	0	0	0	72	40	40	3734665
hidalgo13	0.2840	0.5000	0.5679	0	0	0	72	40	40	3241899
farias13	0.2816	0.5671	0.5061	4	2	1	55	34	26	24558035
jankowska13	0.2814	0.5381	0.4738	1	0	0	72	44	44	16761536
flekova13	0.2785	0.5343	0.5287	4	4	4	61	39	34	18476373
weren13	0.2564	0.5044	0.5099	1	0	0	71	40	39	11684955
ramirez13	0.2471	0.4781	0.5415	9	0	0	12	40	9	64350734
jimenez13	0.2450	0.4998	0.4885	6	2	1	27	31	14	3940310
moreau13	0.2395	0.4941	0.4824	4	4	2	33	39	19	448406705
baseline	0.1650	0.5000	0.3333	–	–	–	–	–	–	–
patra13	0.1574	0.5683	0.2895	5	4	1	55	17	12	22914419
cagnina13	0.0741	0.5040	0.1234	4	7	4	24	9	8	855252000

Performances on the Spanish portion of the test data

Submission	Accuracy			Runtime
	Total	Gender	Age	(incl. English)
santosh13	0.4208	0.6473	0.6430	17511633
pastor13	0.4158	0.6299	0.6558	2298561
haro13	0.3897	0.6165	0.6219	9559554
flekova13	0.3683	0.6103	0.5966	18476373
ladra13	0.3523	0.6138	0.5727	1729618
jimenez13	0.3145	0.5627	0.5429	3940310
kern13	0.3134	0.5706	0.5375	18285830
yong13	0.3120	0.5468	0.5705	577144695
ramirez13	0.2934	0.5116	0.5651	64350734
aditya13	0.2824	0.5000	0.5643	3734665
jankowska13	0.2592	0.5846	0.4276	16761536
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
gillam13	0.2543	0.4784	0.5377	615347
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
weren13	0.2463	0.5362	0.4615	11684955
cagnina13	0.2339	0.5516	0.4148	855252000
hidalgo13	0.2000	0.5000	0.4000	3241899
farias13	0.1757	0.4982	0.3554	24558035
baseline	0.1650	0.5000	0.3333	–
ayala13	0.1638	0.5526	0.2915	23612726
mechti13	0.0287	0.5455	0.0512	1018000000

NIVERSITY OF
SURREY

UNIVERSITY OF SURREY

“Scientific” foundations?

- We know that text readability measures have been correlated with age (e.g. http://www.cs.surrey.ac.uk/BIMA/People/L.Gillam/downloads/publications/2010.LNCS-readability.pdf)
- But what of gender?
  - “Previous research has shown that women talk almost three times as much as men. In fact, an average woman notches up 20,000 words in a day, which is about 13,000 more than the average man.”
    - http://www.scienceworldreport.com/articles/5073/20130220/why-women-talk-more-men-language-protein.htm
  - But: “Large studies have been conducted on sex differences in verbal abilities within the normal population, and a careful reading of the results suggests that differences in language proficiency do not exist”. Wallentin, M. (2009) “Putative sex differences in verbal abilities and language cortex: A critical review”. Brain and Language 108(3): 175-183.

“Scientific” foundations?

- So for author profiling, can we

1. measure simple features of readability and see if age can be inferred?

2. see if there’s a trace of increased word use merely in sentence lengths?

- *And if the latter works, let others draw whatever conclusions they wish.*

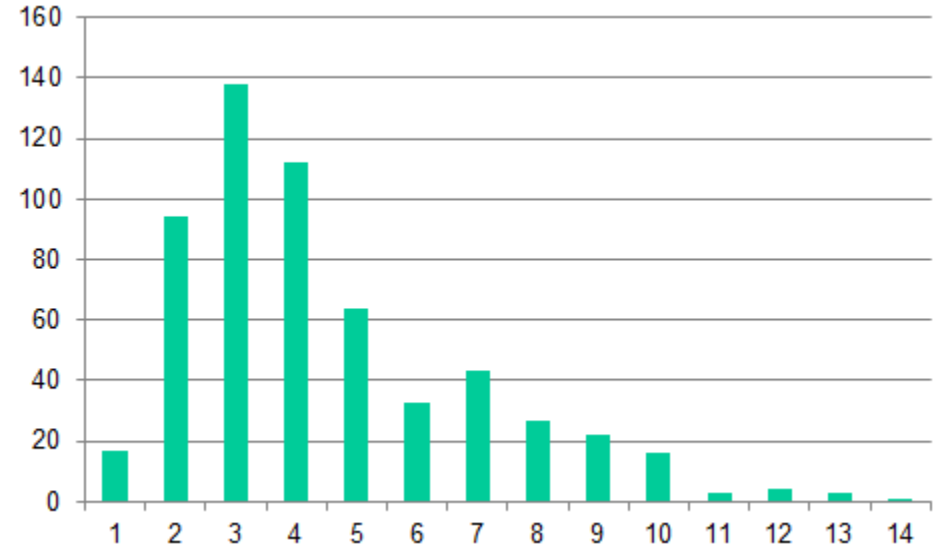
“Scientific” foundations?

- The best known readability measures already encode these for us, so lets break them out:

	Flesch	Kincaid	Fog Index	SMOG	ARI	Dale-Chall	Fry
Sentence count	✓	✓	✓	✓	✓	✓	✓
Word count	✓	✓	✓		✓	✓	
Characters count					✓		
Syllables count	✓	✓					✓
Polysyllable words count (more than three syllables)			✓	✓			
List of easy words						✓	
Scale	0-100	US Grade Level	US Grade Level	US Grade Level	US Grade Level	US Grade Level	US Grade Level

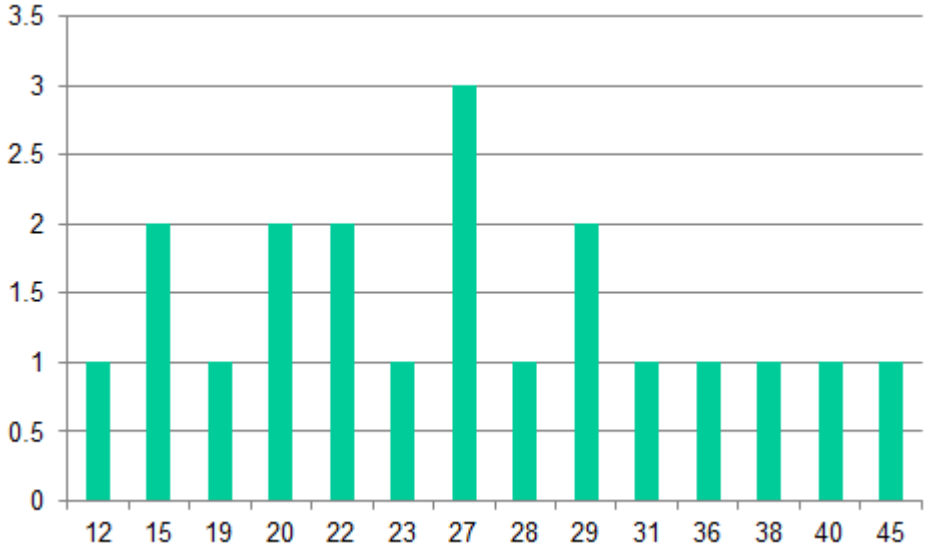
Approach

Word lengths



Sentence lengths

- Ignore if < 35 characters
- Fudged for speed by chars/6



d50d5110d7db800410a47f004b6e92cc_en_20s_male.xml

Approach

- d50d5110d7db800410a47f004b6e92cc_en_20s_male.xml
  - Length-ordered, word at 50% is of length 4.
  - Length-ordered, sentence at 50% is of length 27.
- But these alone don't account for distributions, and in particular a tendency towards longer words and sentences
  - Word at 90% length 7
  - Sentence at 90% length 38
  - So, two values per file (– does the 'readability' tell us anything?):
    - 7+4 = 11 (average + n std devs did not appeal)
    - 27+38 = 65

Approach

- Too many datapoints to interpret manually – so throw at a decision tree and look for compactness (ability to generalise). Weka, J48.

```
J48 pruned tree
------------------

word <= 10: 20s (7673.0/3974.0)
word > 10
|   sentence <= 108: 30s (19334.0/7365.0)
|   sentence > 108
|   |   word <= 11: 20s (45.0/14.0)
|   |   word > 11: 30s (206.0/92.0)

Number of Leaves  :     4

Size of the tree :      7


Time taken to build model: 8.58 seconds
```

```
J48 pruned tree
------------------

wordlength <= 4
|   wordlength <= 3
|   |   sentlength <= 10: male (166.0/60.0)
|   |   sentlength > 10: female (124.0/57.0)
|   wordlength > 3: male (9605.0/4405.0)
wordlength > 4
|   sentlength <= 12
|   |   sentlength <= 9: male (3832.0/1747.0)
|   |   sentlength > 9
|   |   |   wordlength <= 6
|   |   |   |   wordlength <= 5: male (3067.0/1500.0)
|   |   |   |   wordlength > 5: female (2149.0/1065.0)
|   |   |   wordlength > 6
|   |   |   |   sentlength <= 10: female (245.0/111.0)
|   |   |   |   sentlength > 10: male (692.0/296.0)
|   sentlength > 12
|   |   sentlength <= 35
|   |   |   wordlength <= 7
|   |   |   |   sentlength <= 14
|   |   |   |   |   wordlength <= 6: female (5211.0/2526.0)
|   |   |   |   |   wordlength > 6: male (755.0/357.0)
|   |   |   |   sentlength > 14: female (39450.0/18560.0)
|   |   |   wordlength > 7: male (1016.0/482.0)
|   |   sentlength > 35: male (1814.0/842.0)

Number of Leaves  :     13

Size of the tree :      25


Time taken to build model: 38.92 seconds
```

Age ‘easier’ than gender?

Gender: on samples

Age_gender?

UNIVERSITY OF SURREY

```
word <= 10
| sent <= 45
| | word <= 8
| | | word <= 0: 30s_male (14363.0/10626.0)
| | | word > 0
| | | | word <= 6
| | | | | word <= 5: 20s_male (5.0/1.0)
| | | | | word > 5
| | | | | | sent <= 15
| | | | | | | sent <= 11: 20s_female (9.0/6.0)
| | | | | | | sent > 11
| | | | | | | | sent <= 13: 20s_male (22.0/13.0)
| | | | | | | | sent > 13: 30s_male (26.0/18.0)
| | | | | | | sent > 15: 20s_female (27.0/14.0)
| | | | word > 6
| | | | | sent <= 27
| | | | | | sent <= 25
| | | | | | | word <= 7
| | | | | | | | sent <= 14: 30s_male (273.0/182.0)
| | | | | | | | sent > 14
| | | | | | | | | sent <= 17: 20s_male (122.0/87.0)
| | | | | | | | | sent > 17
| | | | | | | | | | sent <= 19: 30s_female (79.0/53.0)
| | | | | | | | | | sent > 19: 20s_male (96.0/66.0)
| | | | | | | | word > 7
| | | | | | | | | sent <= 23
| | | | | | | | | | sent <= 14: 20s_male (241.0/156.0)
| | | | | | | | | | sent > 14: 30s_male (601.0/422.0)
| | | | | | | | | sent > 23: 20s_male (68.0/44.0)
| | | | | | | sent > 25: 30s_male (86.0/55.0)
| | | | | sent > 27
| | | | | | word <= 7
| | | | | | | sent <= 41
| | | | | | | | sent <= 33
| | | | | | | | | sent <= 29: 20s_female (17.0/11.0)
| | | | | | | | | sent > 29: 30s_male (39.0/18.0)
| | | | | | | | sent > 33: 20s_female (17.0/8.0)
| | | | | | | sent > 41: 20s_male (8.0/4.0)
| | | | | | word > 7
| | | | | | | sent <= 42: 20s_female (275.0/195.0)
| | | | | | | sent > 42: 20s_male (16.0/9.0)
| | word > 8
| | | word <= 9
| | | | sent <= 30
| | | | | sent <= 11: 20s_male (26.0/14.0)
| | | | | sent > 11
| | | | | | sent <= 14: 30s_female (384.0/284.0)
| | | | | | sent > 14
| | | | | | | sent <= 19
| | | | | | | | sent <= 18
| | | | | | | | | sent <= 15: 20s_male (15.0/10.0)
| | | | | | | | | sent > 15: 30s_male (496.0/365.0)
| | | | | | | | sent > 18: 20s_male (16.0/10.0)
| | | | | | | sent > 19
| | | | | | | | sent <= 21: 30s_female (250.0/182.0)
| | | | | | | | sent > 21: 20s_male (670.0/468.0)
| | | | sent > 30
| | | | | sent <= 43
| | | | | | sent <= 39
| | | | | | | sent <= 36
| | | | | | | | sent <= 35: 20s_male (220.0/155.0)
| | | | | | | | sent > 35: 30s_male (94.0/67.0)
| | | | | | | sent > 36: 20s_female (80.0/58.0)
| | | | | | sent > 39
| | | | | | | sent <= 42
| | | | | | | | sent <= 40: 20s_male (43.0/25.0)
| | | | | | | | sent > 40: 30s_male (38.0/26.0)
| | | | | | | sent > 42: 20s_male (7.0/3.0)
| | | | | sent > 43: 30s_male (35.0/24.0)
| | | | word > 9
| | | | | sent <= 14
| | | | | | sent <= 11: 20s_male (22.0/13.0)
| | | | | | sent > 11: 30s_male (355.0/256.0)
| | | | | sent > 14
| | | | | | sent <= 28
| | | | | | | sent <= 18
| | | | | | | | sent <= 16: 30s_female (196.0/142.0)
| | | | | | | | sent > 16: 30s_male (256.0/182.0)
| | | | | | | sent > 18
| | | | | | | | sent <= 20: 20s_male (309.0/226.0)
| | | | | | | | sent > 20
| | | | | | | | | sent <= 25
| | | | | | | | | | sent <= 21: 30s_female (44.0/32.0)
| | | | | | | | | | sent > 21: 20s_male (537.0/387.0)
| | | | | | | | | sent > 25
| | | | | | | | | | sent <= 27: 30s_male (285.0/211.0)
| | | | | | | | | | sent > 27: 20s_male (200.0/140.0)
| | | | | | sent > 28
| | | | | | | sent <= 29: 30s_male (58.0/38.0)
| | | | | | | sent > 29
| | | | | | | | sent <= 31: 20s_male (160.0/114.0)
| | | | | | | | sent > 31
| | | | | | | | | sent <= 34: 30s_female (285.0/203.0)
| | | | | | | | | sent > 34: 30s_male (708.0/508.0)
| | sent > 45
| | | sent <= 187
| | | | word <= 7
| | | | | word <= 6: 20s_male (6.0/3.0)
| | | | | word > 6
| | | | | | sent <= 128: 30s_male (30.0/13.0)
| | | | | | sent > 128: 10s_female (2.0/1.0)
| | | word > 7
| | | | sent <= 75: 20s_male (974.0/653.0)
| | | | sent > 75
| | | | | word <= 9
| | | | | | word <= 8: 20s_male (65.0/47.0)
| | | | | | word > 8
| | | | | | | sent <= 102
| | | | | | | | sent <= 76: 20s_male (6.0/3.0)
| | | | | | | | sent > 76
| | | | | | | | | sent <= 89
| | | | | | | | | | sent <= 82: 30s_male (23.0/14.0)
| | | | | | | | | | sent > 82: 30s_female (23.0/13.0)
| | | | | | | | | sent > 89: 20s_male (27.0/16.0)
| | | | | | | sent > 102
| | | | | | | | sent <= 110
| | | | | | | | | sent <= 103: 10s_male (3.0/2.0)
| | | | | | | | | sent > 103: 30s_female (11.0/3.0)
| | | | | | | | sent > 110
| | | | | | | | | sent <= 160: 30s_male (43.0/32.0)
| | | | | | | | | sent > 160
| | | | | | | | | | sent <= 175
| | | | | | | | | | | sent <= 165: 30s_female (2.0)
| | | | | | | | | | | sent > 165: 20s_female (4.0/1.0)
| | | | | | | | | | sent > 175: 30s_female (5.0/2.0)
| | | | | word > 9
| | | | | | sent <= 112
| | | | | | | sent <= 93: 20s_male (93.0/59.0)
| | | | | | | sent > 93
| | | | | | | | sent <= 94: 30s_female (2.0)
| | | | | | | | sent > 94: 30s_male (53.0/39.0)
| | | | | | sent > 112
| | | | | | | sent <= 117: 20s_female (6.0/3.0)
| | | | | | | sent > 117: 20s_male (60.0/35.0)
| | sent > 187: 20s_male (143.0/76.0)
word > 10
| sent <= 28
| | word <= 20: 30s_male (7715.0/5436.0)
| | word > 20
| | | word <= 21
| | | | sent <= 19
| | | | | sent <= 14: 20s_female (3.0/1.0)
| | | | | sent > 14: 30s_male (2.0/1.0)
| | | | sent > 19: 30s_male (67.0/1.0)
| | | word > 21
| | | | sent <= 23
| | | | | sent <= 19
| | | | | | word <= 24
| | | | | | | word <= 22: 20s_male (3.0/1.0)
| | | | | | | word > 22: 20s_female (4.0/2.0)
| | | | | | word > 24: 20s_male (4.0/1.0)
| | | | sent > 19
| | | | | word <= 24: 30s_female (3.0/1.0)
| | | | | word > 24: 20s_female (3.0/1.0)
| | | | sent > 23: 20s_male (8.0/4.0)
| sent > 28
| | sent <= 110
| | | word <= 11
| | | | sent <= 72
| | | | | sent <= 45
| | | | | | sent <= 36
| | | | | | | sent <= 34
| | | | | | | | sent <= 33: 30s_male (1053.0/736.0)
| | | | | | | | sent > 33: 30s_female (276.0/180.0)
| | | | | | | sent > 34: 30s_male (523.0/359.0)
| | | | | | sent > 36
| | | | | | | sent <= 39
| | | | | | | | sent <= 37: 30s_male (224.0/154.0)
| | | | | | | | sent > 37: 30s_female (519.0/367.0)
| | | | | | | sent > 39: 30s_male (1350.0/932.0)
| | | | | sent > 45
| | | | | | sent <= 52: 30s_female (1117.0/752.0)
| | | | | | sent > 52: 30s_male (1210.0/843.0)
| | | | sent > 72
| | | | | sent <= 95
| | | | | | sent <= 74: 30s_male (40.0/26.0)
| | | | | | sent > 74: 30s_female (248.0/183.0)
| | | | | sent > 95
| | | | | | sent <= 106
| | | | | | | sent <= 99
| | | | | | | | sent <= 98: 20s_male (27.0/13.0)
| | | | | | | | sent > 98: 10s_male (2.0/1.0)
| | | | | | | sent > 99: 30s_female (18.0/10.0)
| | | | | | sent > 106
| | | | | | | sent <= 108: 30s_male (8.0/4.0)
| | | | | | | sent > 108: 20s_male (3.0/1.0)
| | | word > 11
| | | | word <= 14
| | | | | sent <= 38
| | | | | | word <= 12: 30s_female (4493.0/2964.0)
| | | | | | word > 12
| | | | | | | sent <= 32
| | | | | | | | word <= 13
| | | | | | | | | sent <= 30: 30s_female (319.0/214.0)
| | | | | | | | | sent > 30: 30s_male (456.0/311.0)
| | | | | | | | word > 13: 30s_male (340.0/231.0)
| | | | | | | sent > 32
| | | | | | | | sent <= 33
| | | | | | | | | word <= 13: 30s_female (230.0/155.0
| | | | | | | | | word > 13: 30s_male (71.0/44.0)
| | | | | | | | sent > 33
| | | | | | | | | word <= 13: 30s_male (1765.0/1184.(
| | | | | | | | | word > 13: 30s_female (666.0/431.0)
| | | | sent > 38
| | | | | sent <= 77: 30s_female (32225.0/21271.0)
| | | | | sent > 77
| | | | | | sent <= 78
| | | | | | | word <= 12: 30s_female (86.0/51.0)
| | | | | | | word > 12: 30s_male (126.0/79.0)
| | | | | | sent > 78
| | | | | | | sent <= 79: 30s_female (128.0/81.0)
| | | | | | | sent > 79
| | | | | | | | word <= 12
| | | | | | | | | sent <= 80: 30s_male (46.0/26.0)
| | | | | | | | | sent > 80
| | | | | | | | | | sent <= 100
| | | | | | | | | | | sent <= 96
| | | | | | | | | | | | sent <= 81: 30s_female (56.0/36.0
| | | | | | | | | | | | sent > 81: 30s_male (426.0/293.0)
| | | | | | | | | | | sent > 96: 30s_female (42.0/26.0)
| | | | | | | | | | sent > 100: 30s_male (80.0/50.0)
| | | | | | | | | word > 12
| | | | | | | | | | sent <= 86
| | | | | | | | | | | sent <= 83: 30s_male (298.0/194.0)
| | | | | | | | | | | sent > 83
| | | | | | | | | | | | sent <= 84: 30s_female (63.0/41.0)
| | | | | | | | | | | | sent > 84
| | | | | | | | | | | | | sent <= 85: 30s_male (58.0/30.0)
| | | | | | | | | | | | | sent > 85: 30s_female (80.0/60.0)
| | | | | | | | | | sent > 86: 30s_female (603.0/396.0)
| | | word > 14
| | | | word <= 16
| | | | | sent <= 44
| | | | | | word <= 15
| | | | | | | sent <= 36
| | | | | | | | sent <= 31: 30s_female (71.0/47.0)
| | | | | | | | sent > 31: 30s_male (248.0/160.0)
| | | | | | | sent > 36
| | | | | | | | sent <= 38: 30s_female (94.0/62.0)
| | | | | | | | sent > 38: 30s_male (238.0/164.0)
| | | | | | word > 15
| | | | | | | sent <= 39
| | | | | | | | sent <= 30: 30s_female (26.0/15.0)
| | | | | | | | sent > 30: 30s_male (118.0/80.0)
| | | | | | | sent > 39
| | | | | | | | sent <= 40: 20s_male (30.0/16.0)
| | | | | | | | sent > 40: 30s_male (58.0/36.0)
| | | | | sent > 44
| | | | | | sent <= 66
| | | | | | | sent <= 64: 30s_female (1020.0/674.0)
| | | | | | | sent > 64
| | | | | | | | sent <= 65: 30s_male (33.0/19.0)
| | | | | | | | sent > 65: 30s_female (38.0/24.0)
| | | | | | sent > 66
| | | | | | | sent <= 70
| | | | | | | | sent <= 69
| | | | | | | | | sent <= 67: 30s_male (31.0/18.0)
| | | | | | | | | sent > 67: 30s_female (55.0/28.0)
| | | | | | | | sent > 69: 20s_female (34.0/22.0)
| | | | | | | sent > 70
| | | | | | | | sent <= 104
| | | | | | | | | sent <= 74: 30s_male (83.0/47.0)
| | | | | | | | | sent > 74
| | | | | | | | | | sent <= 93: 30s_female (175.0/105.0)
| | | | | | | | | | sent > 93
| | | | | | | | | | | sent <= 102
| | | | | | | | | | | | sent <= 99: 20s_female (28.0/20.0
| | | | | | | | | | | | sent > 99
| | | | | | | | | | | | | sent <= 100: 20s_male (7.0/3.0)
| | | | | | | | | | | | | sent > 100: 30s_male (5.0/1.0)
| | | | | | | | | | | | sent > 102: 20s_male (7.0/5.0)
| | | | | | | | | | sent > 104
| | | | | | | | | | | sent <= 107
| | | | | | | | | | | | sent <= 106: 20s_female (6.0/3.0)
| | | | | | | | | | | | sent > 106: 10s_male (5.0/3.0)
| | | | | | | | | | | | sent > 107: 30s_male (10.0/4.0)
| | | | | word > 16
| | | | | | word <= 17
| | | | | | | sent <= 93
| | | | | | | | sent <= 59
| | | | | | | | | sent <= 53
| | | | | | | | | | sent <= 50
| | | | | | | | | | | sent <= 40: 30s_male (62.
| | | | | | | | | | | sent > 40: 30s_female (39
| | | | | | | | | | sent > 50: 30s_male (4.0/1.0
| | | | | | | | | sent > 53
| | | | | | | | | | sent <= 55: 30s_female (6.0/
| | | | | | | | | | sent > 55
| | | | | | | | | | | sent <= 57: 20s_male (6.0
| | | | | | | | | | | sent > 57: 10s_male (2.0/1
| | | | | | | | | sent > 59
| | | | | | | | | | sent <= 61: 30s_male (7.0/4.0)
| | | | | | | | | | sent > 61
| | | | | | | | | | | sent <= 64
| | | | | | | | | | | | sent <= 63: 30s_female (5
| | | | | | | | | | | | sent > 63: 10s_female (2.(
| | | | | | | | | | sent > 64
| | | | | | | | | | | sent <= 84
| | | | | | | | | | | | sent <= 77
| | | | | | | | | | | | | sent <= 69
| | | | | | | | | | | | | | sent <= 67: 30s_m
| | | | | | | | | | | | | | sent > 67: 20s_fem
| | | | | | | | | | | | | sent > 69
| | | | | | | | | | | | | | sent <= 76: 30s_fe
| | | | | | | | | | | | | | | sent > 76: 30s_ma
| | | | | | | | | | | | | | sent > 77: 20s_male (4.
| | | | | | | | | | | | | sent > 84: 30s_female (5.(
| | | | | | | | | sent > 93
| | | | | | | | | | sent <= 101: 10s_male (3.0/2.0)
| | | | | | | | | | sent > 101: 10s_female (2.0)
| | | | | | word > 17
| | | | | | | sent <= 38
| | | | | | | | word <= 20
| | | | | | | | | word <= 19
| | | | | | | | | | word <= 18
| | | | | | | | | | | sent <= 33: 20s_female (1
| | | | | | | | | | | sent > 33
| | | | | | | | | | | | sent <= 37: 30s_male (3
| | | | | | | | | | | | sent > 37: 20s_male (6.
| | | | | | | | | | | word > 18: 20s_male (10.0/4
| | | | | | | | | | word > 19: 20s_female (2.0/1.0
| | | | | | | | | word > 20: 20s_male (15.0/10.0)
| | | | | | | | sent > 38
| | | | | | | | | word <= 19
| | | | | | | | | | word <= 18
| | | | | | | | | | | sent <= 42: 30s_male (5.0/2.
| | | | | | | | | | | sent > 42: 30s_female (26.0/
| | | | | | | | | | word > 18
| | | | | | | | | | | sent <= 43: 30s_female (3.0)
| | | | | | | | | | | sent > 43
| | | | | | | | | | | | sent <= 58: 30s_male (6.0
| | | | | | | | | | | | sent > 58
| | | | | | | | | | | | | sent <= 61: 30s_female
| | | | | | | | | | | | | sent > 61: 30s_male (5.
| | | | | | | | | word > 19: 30s_female (8.0/3.0)
| | sent > 110
| | | word <= 11
| | | | sent <= 121
| | | | | sent <= 115
| | | | | | sent <= 114
| | | | | | | sent <= 113: 20s_male (7.0/4.0)
| | | | | | | sent > 113: 20s_female (2.0)
| | | | | | | sent > 114: 20s_male (2.0/1.0)
| | | | | sent > 115
| | | | | | sent <= 117: 30s_male (5.0/3.0)
| | | | | | sent > 117: 30s_female (12.0/7.0)
| | | | | sent > 121
| | | | | | sent <= 142
| | | | | | | sent <= 133: 20s_male (20.0/11.0)
| | | | | | | sent > 133
| | | | | | | | sent <= 137: 20s_female (5.0)
| | | | | | | | sent > 137: 20s_male (5.0/2.0)
| | | | | | sent > 142: 20s_male (79.0/53.0)
| | | word > 11
| | | | sent <= 578
| | | | | word <= 13
| | | | | | word <= 12
| | | | | | | sent <= 173
| | | | | | | | sent <= 153
| | | | | | | | | sent <= 113
| | | | | | | | | | sent <= 112
| | | | | | | | | | | sent <= 111: 30s_female (2.0)
| | | | | | | | | | | sent > 111: 20s_male (4.0/2.0)
| | | | | | | | | | sent > 112: 30s_female (6.0/3.0)
| | | | | | | | | | sent > 113: 30s_male (126.0/83.0)
| | | | | | | | | sent > 153
| | | | | | | | | | sent <= 167
| | | | | | | | | | | sent <= 156: 30s_female (4.0/2.0)
| | | | | | | | | | | sent > 156: 20s_female (10.0/6.0)
| | | | | | | | | | sent > 167: 30s_female (3.0/1.0)
| | | | | | | | | sent > 173: 30s_male (74.0/49.0)
| | | | | | word > 12
| | | | | | | sent <= 305
| | | | | | | | sent <= 115
| | | | | | | | | sent <= 113: 30s_male (18.0/11.0)
| | | | | | | | | sent > 113: 20s_male (14.0/7.0)
| | | | | | | | sent > 115
| | | | | | | | | sent <= 120
| | | | | | | | | | sent <= 116: 20s_female (4.0/1.0)
| | | | | | | | | | sent > 116: 30s_male (18.0/11.0)
| | | | | | | | | sent > 120
| | | | | | | | | | sent <= 236: 30s_female (91.0/59.0)
| | | | | | | | | | sent > 236
| | | | | | | | | | | sent <= 282
| | | | | | | | | | | | sent <= 250: 30s_female (4.0/2.0)
| | | | | | | | | | | | sent > 250: 20s_female (4.0)
| | | | | | | | | | | sent > 282: 30s_female (3.0/1.0)
| | | | | | | | | sent > 305
| | | | | | | | | | sent <= 376
| | | | | | | | | | | sent <= 315: 10s_male (2.0)
| | | | | | | | | | | sent > 315: 30s_female (7.0/3.0)
| | | | | | | | | | sent > 376
| | | | | | | | | | | sent <= 520: 10s_male (4.0/1.0)
| | | | | | | | | | | sent > 520: 20s_male (2.0/1.0)
| | | | word > 13
| | | | | word <= 27
| | | | | | word <= 16
| | | | | | | word <= 14
| | | | | | | | sent <= 130
| | | | | | | | | sent <= 120
| | | | | | | | | | sent <= 118
| | | | | | | | | | | sent <= 115
| | | | | | | | | | | | sent <= 114
| | | | | | | | | | | | | sent <= 113: 30s_female
(8.0/4.0)
| | | | | | | | | | | | | sent > 113: 20s_male (5.0/3.0)
| | | | | | | | | | | | | sent > 114: 30s_male (3.0/1.0)
| | | | | | | | | | | | | sent > 115: 20s_female (11.0/7.0)
| | | | | | | | | | | | sent > 118: 30s_male (4.0/1.0)
| | | | | | | | | | sent > 120
| | | | | | | | | | | sent <= 128
| | | | | | | | | | | | sent <= 121: 20s_male (3.0/1.0)
| | | | | | | | | | | | sent > 121
...
```

Number of Leaves : 220

Size of the tree : 439

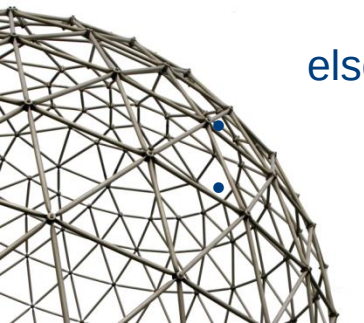
Final rules

- **AGE:**

```
if( word <= 10):  return "20s"
else: if( sentence <= 108):  return "30s"
        else: if( word <= 11): return "20s"
                else: return "30s"
```

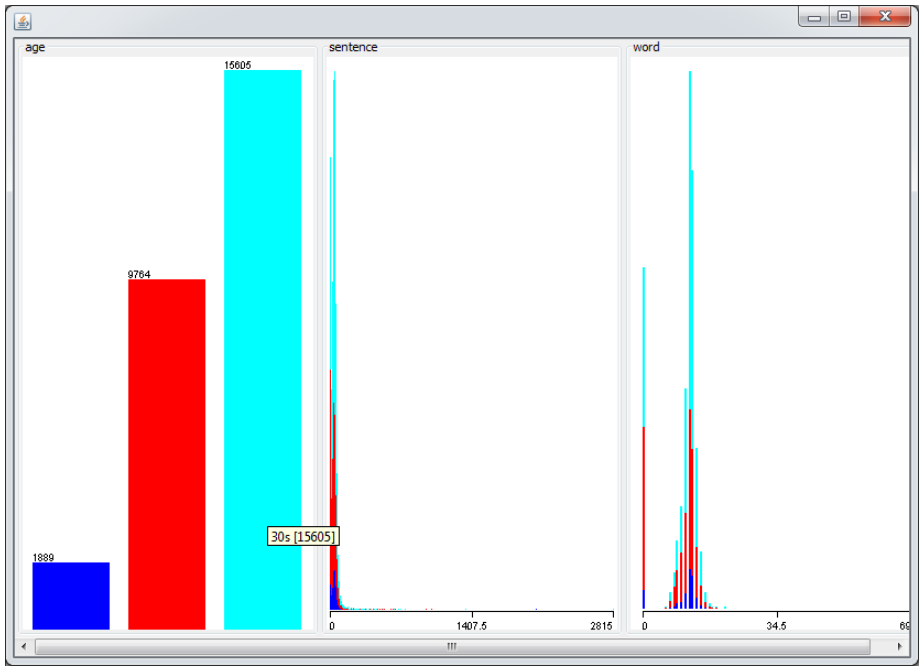
- **GENDER:**

```
if( sentence <= 28):
        if( word <= 18): return "male"
        else:
                if( sentence < 17):
                        if( word <= 21): return "female"
                        else: return "male"
                else:
                        return "male"
else:
                if(word <= 11): return "male"
                else: return "female"
```



Caveats

- Little effort put in to deriving results – hadn't noticed Spanish texts to start with – just wanted to see if this simple approach did anything.
- Approach works quite quickly (after all, it isn't doing much!)
- Should really do sentence lengths 'properly'.
- Many parameter values could be tested; different values to encompass distribution.
- And:



Caveats

- No 10s!

```
J48 pruned tree
------------------

word <= 10: 20s (7673.0/3974.0)
word > 10
|   sentence <= 108: 30s (19334.0/7365.0)
|   sentence > 108
|   |   word <= 11: 20s (45.0/14.0)
|   |   word > 11: 30s (206.0/92.0)

Number of Leaves  :     4

Size of the tree :      7


Time taken to build model: 8.58 seconds
```

  - Correctly Classified Instances 15802 57.972 %
  - Incorrectly Classified Instances 11456 42.028 %

  - Small proportion labelled 10s – so, ‘guesses’ towards 20s/30s.

- Test set proportions for 10s, 20s, 30s?

UNIVERSITY OF SURREY

Performances on the English portion of the test data

Submission	Accuracy			Adult			Predator			Runtime
	Total	Gender	Age	Gender	Age	Both	Gender	Age	Both	(incl. Spanish)
meina13	0.3894	0.5921	0.6491	6	8	6	72	41	41	383821541
pastor13	[illegible]	0.5690	0.6572	1	8	0	72	32	32	2298561
mechti13	0.3677	0.5816	0.5897	2	6	2	52	29	20	1018000000
santosh13	0.3508	0.5652	0.6408	9	9	9	69	32	29	17511633
yong13	0.3488	0.5671	0.6098	6	1	1	28	30	17	577144695
ladra13	0.3420	0.5608	0.6118	9	9	9	72	33	33	1729618
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
gillam13	0.3268	0.5410	0.6031	1	4	0	72	30	30	615347
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
haro13	0.3114	0.5456	0.5966	0	8	0	69	44	41	9559554
aditya13	0.2843	0.5000	0.6055	0	0	0	72	40	40	3734665
hidalgo13	0.2840	0.5000	0.5679	0	0	0	72	40	40	3241899
farias13	0.2816	0.5671	0.5061	4	2	1	55	34	26	24558035
jankowska13	0.2814	0.5381	0.4738	1	0	0	72	44	44	16761536
flekova13	0.2785	0.5343	0.5287	4	4	4	61	39	34	18476373
weren13	0.2564	0.5044	0.5099	1	0	0	71	40	39	11684955
ramirez13	0.2471	0.4781	0.5415	9	0	0	12	40	9	64350734
jimenez13	0.2450	0.4998	0.4885	6	2	1	27	31	14	3940310
moreau13	0.2395	0.4941	0.4824	4	4	2	33	39	19	448406705
baseline	0.1650	0.5000	0.3333	–	–	–	–	–	–	–
[illegible]	[illegible]	[illegible]	[illegible]	5	4	1	55	17	12	22914419
cagnina13	0.0741	0.5040	0.1234	4	7	4	24	9	8	855252000

Performances on the Spanish portion of the test data

Submission	Accuracy			Runtime
	Total	Gender	Age	(incl. English)
santosh13	0.4208	0.6473	0.6430	17511633
pastor13	0.4158	0.6299	0.6558	2298561
haro13	0.3897	0.6165	0.6219	9559554
flekova13	0.3683	0.6103	0.5966	18476373
ladra13	0.3523	0.6138	0.5727	1729618
jimenez13	0.3145	0.5627	0.5429	3940310
kern13	0.3134	0.5706	0.5375	18285830
yong13	0.3120	0.5468	0.5705	577144695
ramirez13	0.2934	0.5116	0.5651	64350734
aditya13	0.2824	0.5000	0.5643	3734665
jankowska13	0.2592	0.5846	0.4276	16761536
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
gillam13	0.2543	0.4784	0.5377	615347
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
weren13	0.2463	0.5362	0.4615	11684955
cagnina13	0.2339	0.5516	0.4148	855252000
hidalgo13	0.2000	0.5000	0.4000	3241899
farias13	0.1757	0.4982	0.3554	24558035
baseline	0.1650	0.5000	0.3333	–
[illegible]	[illegible]	[illegible]	[illegible]	23612726
mechti13	0.0287	0.5455	0.0512	1018000000

What influence data bias?

Thank you

Questions?

L.Gillam@surrey.ac.uk

The work presented here has been supported in part by the TSB (IPCRESS) and in the recent past by EPSRC, JISC, TSB (KTP), amongst others.