Outline

Comparing the distributions

We compare the variations of distributions for the n most frequent words

, de . la à et que le il qu' l' un d' les qui une en pas ne des dans était pour n' du ce se s' est

Need a distance or a similarity measure between the word rankings

$$\text{rank-distance}(d_a, d_b) = \Sigma_w |r_a(w) - r_b(w)|$$

Other (normalized) measures are available:
Spearman correlation measure $\rho \in [-1, 1]$, Kandall coefficient τ

$$\rho = 1 - \frac{6\Sigma_w (r_a(w) - r_b(w))^2}{n(n^2 - 1)}$$

Distance matrix

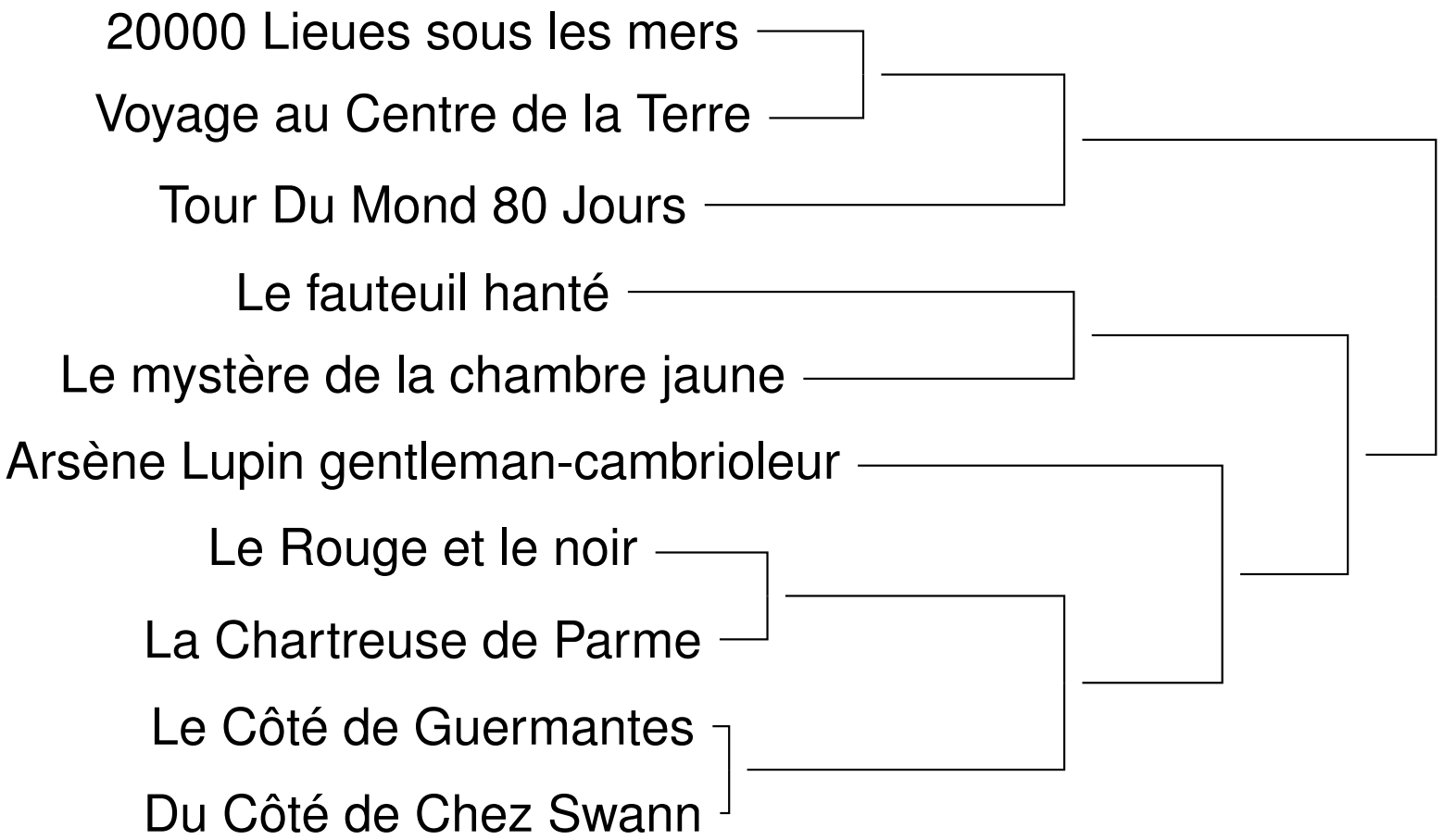
Rank-distance matrix for $n = 50$

> perl ./rankdis.pl *.voc

	Du Côté de Chez ...	La Chartreuse ...	Le mystère de ...	Le fauteuil hanté	Arsène Lupin ...	Tour Du Mond 80 ...	Voyage au Centre ...	20000 Lieues ...	Le Rouge et le ...	Le Côté de Guerma
Du Côté de Chez ...	0	62	106	92	84	108	120	118	68	32
La Chartreuse ...		0	100	92	84	78	100	90	36	66
Le mystère de ...			0	68	100	122	136	122	100	112
Le fauteuil hanté				0	76	108	134	122	88	100
Arsène Lupin ...					0	84	88	88	84	82
Tour Du Mond 80 ...						0	72	62	86	112
Voyage au Centre ...							0	46	104	102
20000 Lieues ...								0	98	102
Le Rouge et le ...									0	72
Le Côté de Guermantes										0

Regroupement (50)

, de . la à et que le il qu' l' un d' les qui une en pas ne des dans était pour n' du ce se s' est



Alternative

Instead of comparing the ranks, we directly compare the distributions of frequencies (seen as probabilities)

Use of Kullback-Leibler divergence KL

$$KL(p \parallel q) = \Sigma_x p(x) \log \frac{p(x)}{q(x)}$$

KL almost a distance:

- $KL(p \parallel p) \geq 0$
- $KL(p \parallel q) = 0 \Longleftrightarrow p = q$
- but $KL(p \parallel q) \neq KL(q \parallel p)$

Recall: entropy $H(p) = -\Sigma_x p(x) \log p(x)$

KL divergence more and more applied
examples found for Language Identification and Authorship Attribution