3

Erreurs

sur

LA CHD REINERT

lucie.loubere@readtext.fr
www.readtext.fr

Lucie Loubère
Data Analyste - Formatrice et recueils & analyses de données

CC BY NC ND

Erreur 1 : On classe des ~~Mots~~ des segments de textes

La classification mise au point par Max Reinert consiste à segmenter les textes puis les classer en fonction de la présence ou absence des mêmes mots. **C’est donc les segments que nous classons**

**** *id_2 *nom_science_and_the_policies_of_governments *an_1963 *typ_corps

foreword this report was prepared by the ad hoc advisory group on science policy which i appointed late in 1961 to advise me on
community

as well as on the general orientation and long term objectives of o e cd activities in this field the group consisted of scientists and

but as individuals knowledgeable and experienced in matters relating to the interaction of science and policy their report therefo

i have read the report with great interest and believe that it represents an important step forward in the elaboration of national a

private organisations and individuals concerned with the relations of science and policy i consider that the o e cd has been fortun

i urge the public and scientific authorities of the oecd_member_countries of the o e cd to encourage wide distribution of ihis docu

introduction chapter 1 scope and purpose of the report previous reports surveys of science and technolo gy on the international s

a t o and unesco 1 the oeec and n a t o reports urged increased national efforts to strengthen science and expand international sc

the unesco report provided valuable information on the state of the physical and biological sciences and on opportunities inherent
r_and_d activities

and to the consequent need for deliberate and consistent policies for their advancement and management 1 dana wilgress co oper

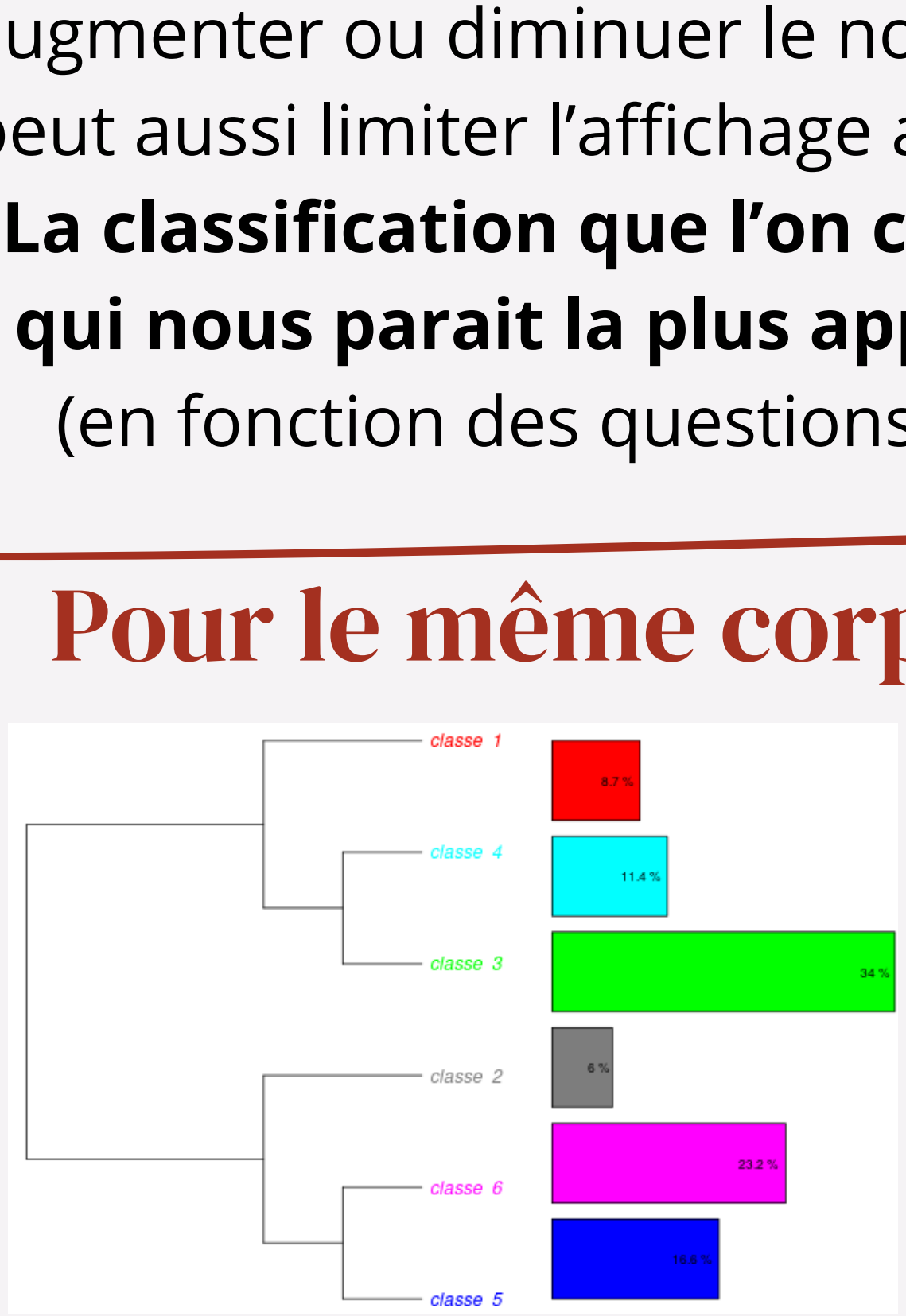
Exemple d’un corpus en couleur… chaque segment est coloré selon son appartenance à une classe

by Lucie Loubere readtext.fr

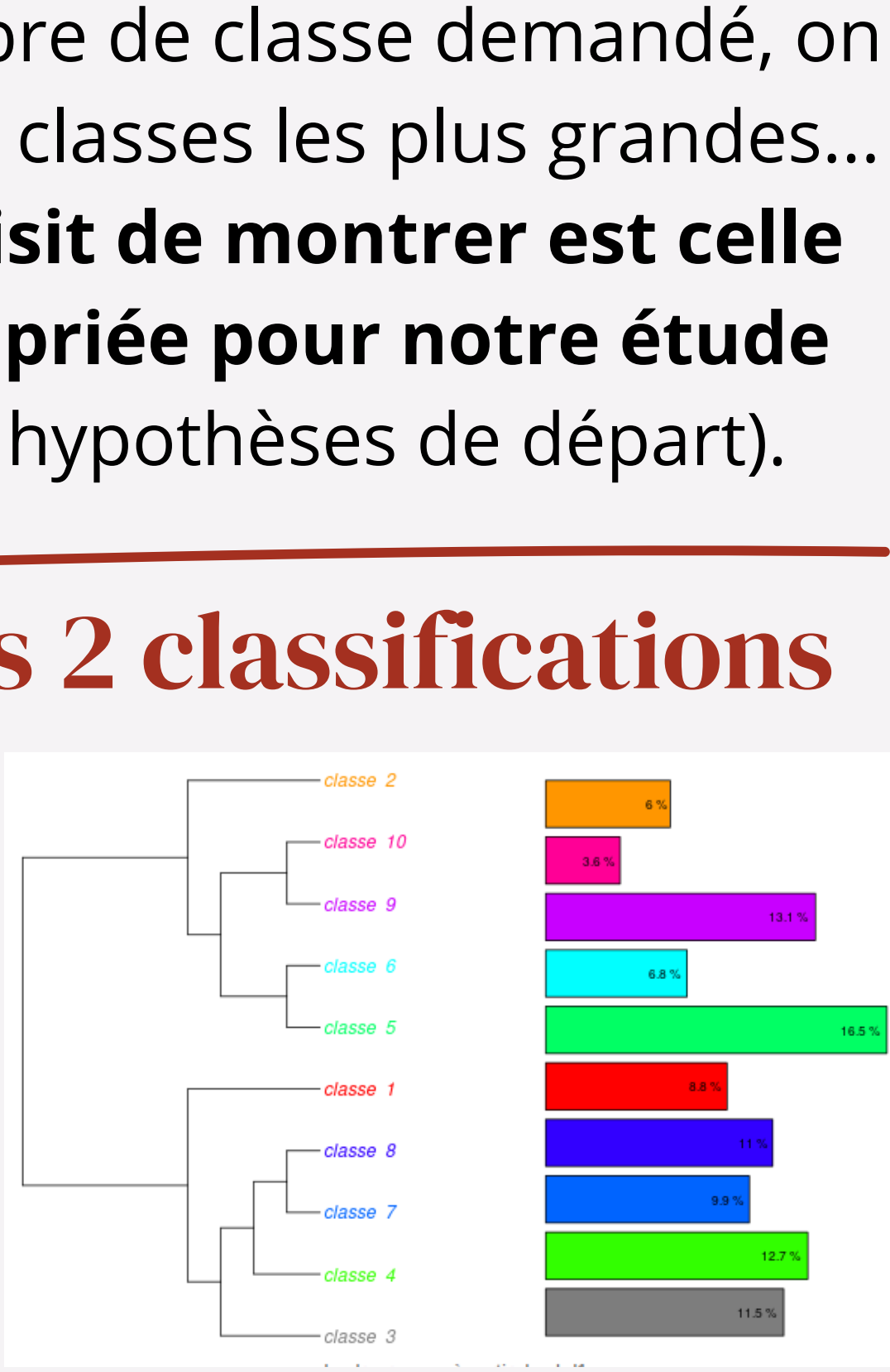
Erreur 2 : Il y a 8 classes dans ce ~~corpus~~ cette classification

Le nombre de classes est défini par l'utilisateur : on peut augmenter ou diminuer le nombre de classe demandé, on peut aussi limiter l'affichage aux classes les plus grandes...
La classification que l'on choisit de montrer est celle qui nous parait la plus appropriée pour notre étude
(en fonction des questions et hypothèses de départ).

Pour le même corpus 2 classifications



classification en 6 classes
17647 segments classés sur 17675 soit 99,84%.



classification en 10 classes
17522 segments classés sur 17675 soit 99,13%.

by Lucie Loubere readtext.fr

Erreur 3 : Les mots les plus hauts dans les profils de classe sont ceux les plus ~~fréquents~~ spécifiques

Les profils des classes sont constitués des mots surreprésentés dans les segments de la classe. Ils sont affichés par ordre décroissant de leur chi2... Ça veut dire que **plus un mot est spécifique de la classe plus il sera haut dans la liste**. En gros un mot qui n'apparaît que 10 fois mais à chaque fois dans la même classe sera tout en haut alors que celui qui apparaît 1000 fois mais de façon homogène ne sera même pas dans les profils.

1 Classe 1	2 Classe 2	3 Classe 3	4 Classe 4	5 Classe 5	6 Classe 6	7 Cla
5032/44401	2228/44401	5081/44401	1196/44401	1771/44401	4846/44401	3786
11.33%	5.02%	11.44%	2.69%	3.99%	10.91%	8.53

num	eff. s.t.	eff. total	pourcentage	chi2	Type	forme
0	706	1101	64.12	3131.074	nom	cent
1	633	950	66.63	2954.158	nom	total
2	522	704	74.15	2808.835	nom	table
3	1001	2233	44.83	2625.041	nr	usa
4	541	1065	50.8	1691.255	nom	expenditure

Ici, le mot USA apparaît dans 1001 segments de cette classe mais n'est qu'en 3me position, alors que le mot total est en 2me position alors qu'il n'apparaît que dans 633 segments de cette même classe

by Lucie Loubere readtext.fr