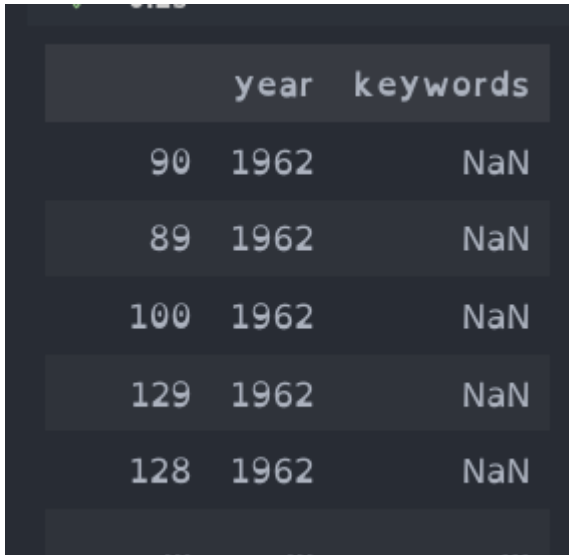


# Rendu TP2

---

## 1. Top 10 des mots-clefs

En analysant les données, j'ai vu qu'il n'y a pas de mots-clefs pour certains articles. J'ai décidé de ne pas prendre en compte les articles sans mots-clefs.



	year	keywords
90	1962	NaN
89	1962	NaN
100	1962	NaN
129	1962	NaN
128	1962	NaN

Mon programme prend trois arguments : le fichier avec les métadonnées des articles, le dossier de sortie du top par décennies (`decadeTopOutput`) et le dossier de sortie du top globale (`keywordTopOutput`).

Exemple de commande de lancement du programme :

```
yarn jar topkeywords-0.0.1.jar /user/fzanonboito/CISD/IEEEdata.csv  
decadeTopOutput keywordTopOutput
```

Mon implémentation est composée de 2 jobs Map-Reduce, le premier `TopDecade` qui utilise 2 mappers et 1 reducer. Le second `TopKeyword` qui utilise 1 mapper et 1 reducer.

Le premier job récupère les mots-clefs et fait le top pour les décennies, pour cela, j'ai le mapper et le reducer suivant :

- Le mapper `RawDataMapper` traite les données en enlevant les lignes sans mots-clefs ou sans date et retourne en clef la décennie et en valeur un mot-clef. Pour calculer la décennie, je fais simplement une division stricte par 10 pour récupérer la décennie par rapport à l'an 0 (e.g. 1998 => 199, -212 => -21), cette solution est inexacte pour les dates entre -10 et 10, car cela donne la décennie 0, mais ce cas n'arrive pas dans nos données qui commencent à partir de l'année 1962. Il peut y avoir plusieurs mappers de ce type en parallèle.
- Le reducer `DecadeReducer` compte pour chaque décennie, le nombre de fois qu'un mot-clef apparaît, puis fait le top de chaque mot-clef. La sortie de ce reducer est sous la forme d'une ligne CSV séparée par des points-virgules. Le premier élément est la décennie suivie du mot-clef puis du nombre de papiers où il apparaît dans la décennie et pour finir le top dans la décennie. Les lignes de

chaque décennie sont triés du plus fréquent au moins fréquent puis par ordre alphabétique pour les mots-clefs par contre les décennies peuvent être dans n'importe quel ordre, mais toutes les lignes d'une décennie se suivent. Il peut y avoir plusieurs reducers de ce type en parallèle et chacun produit son propre fichier. Je garde le nombre de papier où apparaît le mot-clef pour permettre le top global.

Le second job récupère les données par décennies du job précédent pour faire le top global. Pour faire ce top j'ai le mapper et le reducer suivant :

- Le mapper **DecadeMapper** vérifie que les données sont bien présentes et que le nombre de papiers est bien un entier, ensuite, il retourne en clef le mot-clef et en valeur le nombre de papiers par décennie ainsi que la décennie en question et le top dans la décennie. Il peut y avoir plusieurs mappers de ce type en parallèle.
- Le reduceur **KeywordReducer** fait la somme du nombre de papiers pour chaque mots-clefs, il regroupe aussi les données. Ensuite, il fait le top grâce à la somme totale. Ce reducer retourne les données sous la forme d'une ligne CSV séparée par des points-virgules. Cette ligne a pour éléments: le mot-clef, le top global, le nombre global de papier ou apparaît le mot-clef ainsi que chaque top et nombre de papier par décennie. J'ai décidé de mettre le top de chaque décennie dans le fichier final et donc de garder toute les lignes de tous les mots. Je pense que se choix n'impacte pas les performances mais juste la taille du fichier final. Je trouve qu'il est plus pratique de voir les tops de chaque décennie dans ce fichier, car il suffit de trier la colonne voulue. Les données sont triées du mot-clef le plus fréquent aux moins fréquent puis par ordre alphabétique pour les mots-clefs. Il peut y avoir qu'un seul reducer de ce type, c'est nécessaire pour faire le top global.

Les données finales qu'on obtient :

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1	keyword	top	totalNbPaper	topDecade196	nbDecade196	topDecade197	nbDecade197	topDecade198	nbDecade198	topDecade199	nbDecade199	topDecade200	nbDecade200	topDecade201	nbDecade201	topDecade202	nbDecade202	
2	Computational modeling	1	860790	2	6510	3	9702	1	25242	1	74892	1	139104	1	403326	1	202104	
3	Cloud computing	2	480522								3065	126	216	8494	2	347970	2	123942
4	Computer architecture	3	387114			30	2940	7	16086	5	50316	6	91728	5	157080	6	68964	
5	Hardware	4	286776	14	2646	4	9030	3	20622	8	45402	7	74340	7	98448	15	36288	
6	Servers	5	257040	1760	42	2834	42	1718	210	1966	420	949	1974	3	165816	5	88536	
7	Application software	6	243768	3	5754	2	11046	2	23730	2	67914	2	123858	215	11172	2406	294	
8	Computer science	7	221382	439	210	26	3108	18	10122	6	45654	4	101808	30	44478	50	16002	
9	Computer networks	8	218190	13	2814	7	8148	5	17892	4	54054	3	108990	114	19446	160	6846	
10	Computers	9	206640	17	2352	12	5208	16	10248	526	2100	193	9366	4	159096	41	18270	
11	Mathematical model	10	202272	30	1512	21	3486	53	4494	58	10752	31	28350	6	135702	44	17976	
12	Task analysis	11	195174	498	210	992	210	2309	126	2363	294	2723	420	9	82278	3	111636	
13	Resource management	12	179802	1731	42	151	1092	210	1848	167	5502	21	35028	13	77322	7	58968	
14	Distributed computing	13	178542	39	1302	16	4326	12	13020	7	45654	5	92106	141	16296	190	5838	
15	Edge computing	14	165396											14	76482	4	88914	
16	Algorithm design and analysis	15	157920	91	756	40	2352	21	8442	17	25494	17	41790	10	78792	2401	294	
17	Educational institutions	16	155400	86	798	46	2142	26	7182	30	15750	14	51534	12	77994			
18	Concurrent computing	17	146412	32	1470	58	1932	13	12726	3	61824	10	60858	379	6636	959	966	
19	Optimization	18	146244	94	1264	84	126	1364	294	1853	462	567	3402	8	93996	11	47880	
20	Costs	19	144522	11	3066	11	6342	11	13692	12	32802	11	56532	539	4620	28	27468	
21	Control systems	20	138558	7	4032	1	11340	4	20202	11	33600	15	46704	151	15078	146	7602	
22	Equations	21	131880	5	5208	5	8526	9	15498	14	31080	24	33180	42	38388			
23	Training	22	131586	1841	42	1197	168	840	546	3288	126	839	2226	16	71316	8	57162	
24	Testing	23	131460	20	2268	15	4578	15	11004	15	28854	12	52458	70	27090	216	5208	
25	Humans	24	127974	27	1554	23	3276	22	8190	16	27804	8	67620	111	19530			
26	Software	25	126840	489	210	50	2100	65	3990	440	2520	297	6426	11	78750	16	32844	
27	Feature extraction	26	119826			699	294	665	672	282	3654	116	13440	17	71106	21	30660	
28	Laboratories	27	118986	4	5502	6	8526	10	14784	10	35910	16	42420	221	10962	1044	882	
29	Computer simulation	28	118020	6	4326	8	7728	8	16044	9	36288	18	41496	282	8862	324	3276	
30	Parallel processing	29	117516	197	462	254	756	32	5754	13	31500	23	33894	56	32802	78	12348	
31	Conferences	30	112434	651	126	491	420	866	504	1345	756	239	7644	23	53130	10	49854	
32	Education	31	110502	275	336	170	1008	217	1806	81	8274	75	16254	24	51996	20	30828	
33	Security	32	109872			847	252	970	462	1420	714	127	12432	18	66528	23	29484	
34	Computer vision	33	108486			445	462	50	4578	19	18228	28	31206	65	27888	29	26124	
35	Energy consumption	34	105966	390	252	2321	42	1101	378	303	3444	95	14490	27	47376	13	39984	
36	Protocols	35	103194			329	630	60	4200	64	10080	36	25200	28	46914	49	16170	
37	Processor scheduling	36	102396	258	378	184	966	103	2940	50	11718	25	32214	53	33726	36	20454	
38	Analytical models	37	100002	64	966	71	1764	63	4032	46	12054	58	19194	37	39438	34	22554	
39	Monitoring	38	99498	577	168	95	1428	130	2604	220	4536	98	14322	20	62916	69	13524	
40	Visualization	39	98652	1130	84	606	378	350	1260	90	7812	93	14532	26	48300	30	24486	
41	Bandwidth	40	96348	53	1092	65	1806	96	3024	44	12474	34	25452	43	37926	56	14574	
42	Internet of Things	41	91182											41	38514	9	52686	
43	cloud computing	42	89880									1053	1764	15	73290	53	14826	
44	Delays	43	89914			689	294	2058	126	2596	210	3345	252	29	44688	12	43344	
45	Electroencephalography	44	87738	387	252	1282	126	1195	336	664	1680	78	15918	22	54348	51	15078	
46	Neural networks	45	86478					181	2058	20	17934	47	21420	140	16338	26	28728	
47	Data models	46	83622	1315	42	1052	168	1318	294	1289	798	701	2772	25	50274	24	29274	
48	Databases	47	83454			687	294	104	2898	107	7266	49	20874	33	41370	91	10752	
49	Switches	48	82992	21	1974	27	3108	56	4452	35	13734	85	15078	61	31332	70	13314	

Il est facile de trier le top d'une décennie en particulier avec un tableur, python... Par exemple avec les années **2000-2010** :

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	keyword	top	nbDecade196	nbDecade197	nbDecade198	nbDecade199	nbDecade200	nbDecade201	nbDecade202	nbDecade203	nbDecade204	nbDecade205	nbDecade206	nbDecade207	nbDecade208	nbDecade209	nbDecade210
2	Computational modeling	1	860790	2	6510	3	9702	1	25242	1	74802	1	139104	1	403326	1	202104
3	Application software	6	243768	3	5754	2	11046	2	23730	2	67914	2	123858	215	11172	2406	294
4	Computer networks	8	218190	13	2814	7	8148	5	17892	4	54054	3	108990	114	19446	160	6846
5	Computer science	7	221382	439	210	26	3108	18	10122	6	45654	4	101808	30	44478	50	16002
6	Distributed computing	13	178542	39	1302	16	4326	12	13020	7	45654	5	92106	141	16296	190	5838
7	Computer architecture	3	387114			30	2940	7	18086	5	50316	6	91728	5	157080	6	68964
8	Hardware	4	286776	14	2646	4	9030	3	20622	8	45402	7	74340	7	88448	15	36288
9	Humans	24	127974	27	1554	23	3276	22	8190	16	27804	8	67620	111	19530		
10	Grid computing	49	82698	1441	42	705	294	480	924	265	3822	9	61656	149	15246	1241	714
11	Concurrent computing	17	146412	32	1470	58	1932	13	12726	3	61824	10	60858	379	6636	959	966
12	Costs	19	144522	11	3066	11	6342	11	13692	12	32802	11	56532	539	4620	28	27468
13	Testing	23	131460	20	2268	15	4578	15	11004	15	28854	12	52458	70	27090	216	5208
14	Pervasive computing	69	67326	1652	42	1152	168	1367	294	425	2562	13	51996	233	10626	624	1638
15	Educational institutions	16	155400	86	798	46	2142	26	7182	30	15750	14	51534	12	77994		
16	Control systems	20	138558	7	4032	1	11340	4	20202	11	33600	15	46704	151	15078	146	7602
17	Laboratories	27	118986	4	5502	6	8526	10	14784	10	35910	16	42420	221	10962	1044	882
18	Algorithm design and analysis	15	157920	91	756	40	2352	21	8442	17	25494	17	41790	10	78792	2401	294
19	Computer simulation	28	118020	6	4326	8	7728	8	16044	9	36288	18	41496	282	8862	324	3276
20	Computer applications	60	74382	23	1806	28	3024	33	5712	32	15036	19	36120	277	9030	297	3654
21	Mobile computing	64	70770	1593	42	949	210	1358	294	71	9240	20	35574	91	22638	384	2772
22	Resource management	12	179802	1731	42	151	1092	210	1848	167	5502	21	35028	13	77322	7	58968
23	Internet	61	74214					1474	252	145	6006	22	34692	79	25200	133	8064
24	Parallel processing	29	117516	197	462	254	756	32	5754	13	31500	23	33894	56	32802	78	12348
25	Equations	21	131880	5	5208	5	8526	9	15498	14	31080	24	33180	42	38388		
26	Processor scheduling	36	102396	258	378	184	966	103	2940	50	11718	25	32214	53	33726	36	20454
27	Computer science education	85	59388	187	462	141	1134	107	2856	846	8946	26	31794	200	11760	436	2436
28	High performance computing	77	63966	396	252	232	798	190	1932	41	12894	27	31752	211	11382	226	4956
29	Computer vision	33	108486			445	462	50	4578	19	18228	28	31206	65	27888	29	26124
30	Delay	68	67410	51	1134	60	1890	39	5208	21	17724	29	30702	238	10500	2758	252
31	Frequency	57	75768	9	3738	9	7308	19	9954	18	22008	30	30576	1013	2184		
32	Mathematical model	10	202272	30	1512	21	3486	53	4494	58	10752	31	28350	6	135702	44	17976
33	Large-scale systems	138	44058	156	546	158	1050	88	3192	78	9030	32	26376	690	3528	2270	336
34	Performance analysis	86	58506	42	1260	41	2352	29	6006	23	16674	33	25662	560	4494	514	2058
35	Bandwidth	40	96348	53	1092	65	1806	96	3024	44	12474	34	25452	43	37926	56	14574
36	Data mining	72	65058	169	504	98	1386	289	1470	130	6342	35	25410	89	23184	163	6762
37	Protocols	35	103194			329	630	60	4200	64	10080	36	25200	28	46914	49	16170
38	Physics computing	116	48006	28	1554	42	2352	58	4410	43	12726	37	24654	971	2310		
39	Information technology	99	52920			2481	42	760	588	5082	38	24612	152	15078			
40	Robustness	80	63000			1168	168	223	1806	57	10878	39	24318	127	18144	143	7686
41	Ubiquitous computing	205	32508							1704	546	40	23520	363	7014	696	1428
42	Biology computing	188	34524	368	252	487	420	188	1932	104	7350	41	23268	1517	1302		
43	Shape	55	76986	89	798	64	1848	44	4872	25	16086	42	23100	97	22050	127	8232
44	Computer interfaces	118	47586	111	672	31	2856	48	4662	38	13020	43	22470	695	3486	1854	420
45	Mathematics	117	47712	116	672	135	1176	86	3234	51	11298	44	22302	364	6972	513	2058
46	Prototypes	78	63966	482	210	216	882	91	3150	34	14154	45	21924	172	13776	107	9870
47	Field programmable gate arrays	66	68754							163	5544	46	21630	50	34146	150	7434
48	Neural networks	45	86478					181	2058	20	17934	47	21420	140	16338	26	28728
49	Polynomials	82	61992	96	756	33	2856	25	7308	31	15540	48	20958	160	14532	21318	42

Exemple d'où visualiser le fichier de sortie :

```
hdfs dfs -head keywordTopOutput/part-r-00000
```

## 2. Ajout de nouvelles données

Pour prendre en compte l'ajout de nouveau papier publié après la première exécution, j'ai ajouté un quatrième argument. Celui-ci permet de charger les données des décennies déjà calculées.

Par exemple, avec la commande ci-dessous, on a les nouvelles données (**IEEE\_Newdata.csv**) suivies des données déjà calculées (dans **decadeTopOutput**) puis le dossier qui recevra les décennies mis à jour et toujours le dossier de sortie final :

```
yarn jar topkeywords-0.0.1.jar IEEE_Newdata.csv decadeTopOutput
decadeTopOutput_withNewData keywordTopOutput2
```

Pour que l'ajout de nouveau papier publié fonctionne, il faut ajouter un nouveau mapper (**ExistingDataDecadeMapper**) qui vas seulement charger les données déjà calculées. J'ai ajouté ce nouveau mapper au job **TopDecade** sur l'argument **decadeTopOutput** et on garde l'ancien mapper pour l'argument **IEEE\_Newdata.csv**. J'utilise un **MultipleInputs** pour avoir plusieurs mappers sur le job **TopDecade**. J'ai aussi modifié le reducer **DecadeReducer** pour qu'il ignore les décennies déjà calculées.

J'ai créé un fichier **IEEE\_Newdata.csv** avec des faux articles qui ont pour mots-clefs: **Energy consumption**, **Hardware** et **Software**. Après l'exécution, on voit que le mot-clef **Energy consumption** passe bien du top 34 au top 20 :

	A	B	C			A	B	C	
1	keyword	top	totalNbInPaper	top	1	keyword	top	totalNbInPaper	top
2	Computational modeling	1	860790		2	Computational modeling	1	658687	
3	Cloud computing	2	480522		3	Cloud computing	2	356581	
4	Computer architecture	3	387114		4	Computer architecture	3	318151	
5	Hardware	4	286776		5	Hardware	4	304787	
6	Servers	5	257040		6	Application software	5	243475	
7	Application software	6	243768		7	Computer networks	6	211345	
8	Computer science	7	221382		8	Computer science	7	205381	
9	Computer networks	8	218190		9	Computers	8	188371	
10	Computers	9	206640		10	Mathematical model	9	184297	
11	Mathematical model	10	202272		11	Distributed computing	10	172705	
12	Task analysis	11	195174		12	Servers	11	168505	
13	Resource management	12	179802		13	Algorithm design and analysis	12	157627	
14	Distributed computing	13	178542		14	Educational institutions	13	155400	
15	Edge computing	14	165396		15	Concurrent computing	14	145447	
16	Algorithm design and analysis	15	157920		16	Equations	15	131880	
17	Educational institutions	16	155400		17	Control systems	16	130957	
18	Concurrent computing	17	146412		18	Humans	17	127974	
19	Optimization	18	146244		19	Testing	18	126253	
20	Costs	19	144522		20	Resource management	19	120835	
21	Control systems	20	138558		21	Energy consumption	20	120281	
22	Equations	21	131880		22	Laboratories	21	118105	
23	Training	22	131586		23	Costs	22	117055	
24	Testing	23	131460		24	Computer simulation	23	114745	
25	Humans	24	127974		25	Software	24	113161	
26	Software	25	126840		26	Parallel processing	25	105169	
27	Feature extraction	26	119826		27	Optimization	26	98365	
28	Laboratories	27	118986		28	Feature extraction	27	89167	
29	Computer simulation	28	118020		29	Protocols	28	87025	
30	Parallel processing	29	117516		30	Monitoring	29	85975	
31	Conferences	30	112434		31	Task analysis	30	83539	
32	Education	31	110502		32	Computer vision	31	82363	
33	Security	32	109872		33	Grid computing	32	81985	
34	Computer vision	33	108486		34	Processor scheduling	33	81943	
35	Energy consumption	34	105966		35	Bandwidth	34	81775	
36	Protocols	35	103194		36	Security	35	80389	
37	Processor scheduling	36	102396		37	Education	36	79675	
38	Analytical models	37	100002		38	Analytical models	37	77449	
39	Monitoring	38	99498		39	Edge computing	38	76483	
40	Visualization	39	96852		40	Frequency	39	75768	
41	Bandwidth	40	96348		41	cloud computing	40	75055	
42	Internet of Things	41	91182		42	Mobile communication	41	74551	
43	cloud computing	42	89880		43	Training	42	74425	
44	Delays	43	88914		44	Databases	43	72703	
45	Electroencephalography	44	87738		45	Electroencephalography	44	72661	
46	Neural networks	45	86478		46	Visualization	45	72367	
47	Data models	46	83622		47	Computer applications	46	70729	
48	Databases	47	83454		48	Switches	47	69679	
49	Switches	48	82992		49	Shape	48	68755	

On voit que sur le top de la décennie 2000-2010, il n'y a pas eu de changement dans les données :

	A	B	C		L	M		A	B	C		L	M
1	keyword	top	totalNbInPaper	topDecade200	nbDecade200	topDecade200	1	keyword	top	totalNbInPaper	topDecade200	nbDecade200	topDecade200
2	Computational modeling	1	860790	1	139104	1	2	Computational modeling	1	658687	1	139104	1
3	Application software	6	243768	2	123858	2	3	Application software	5	243475	2	123858	2
4	Computer networks	8	218190	3	108990	3	4	Computer networks	6	211345	3	108990	3
5	Computer science	7	221382	4	101808	4	5	Computer science	7	205381	4	101808	4
6	Distributed computing	13	178542	5	92106	5	6	Distributed computing	10	172705	5	92106	5
7	Computer architecture	3	387114	6	91728	6	7	Computer architecture	3	318151	6	91728	6
8	Hardware	4	286776	7	74340	7	8	Hardware	4	304787	7	74340	7
9	Humans	24	127974	8	67620	8	9	Humans	17	127974	8	67620	8
10	Grid computing	49	82698	9	61656	9	10	Grid computing	32	81985	9	61656	9
11	Concurrent computing	17	146412	10	60858	10	11	Concurrent computing	14	145447	10	60858	10
12	Costs	19	144522	11	56532	11	12	Costs	22	117055	11	56532	11
13	Testing	23	131460	12	52458	12	13	Testing	18	126253	12	52458	12
14	Pervasive computing	69	67326	13	51996	13	14	Pervasive computing	52	65689	13	51996	13
15	Educational institutions	16	155400	14	51534	14	15	Educational institutions	13	155400	14	51534	14
16	Control systems	20	138558	15	46704	15	16	Control systems	16	130957	15	46704	15
17	Laboratories	27	118996	16	42420	16	17	Laboratories	21	118105	16	42420	16
18	Algorithm design and analysis	15	157920	17	41790	17	18	Algorithm design and analysis	12	157627	17	41790	17
19	Computer simulation	28	118020	18	41496	18	19	Computer simulation	23	114745	18	41496	18
20	Computer applications	60	74382	19	36120	19	20	Computer applications	46	70729	19	36120	19
21	Mobile computing	64	70770	20	35574	20	21	Mobile computing	49	67999	20	35574	20
22	Resource management	12	179802	21	35028	21	22	Resource management	19	120835	21	35028	21
23	Internet	61	74214	22	34692	22	23	Internet	51	66151	22	34692	22
24	Parallel processing	29	117516	23	33894	23	24	Parallel processing	25	105169	23	33894	23
25	Equations	21	131880	24	33180	24	25	Equations	15	131880	24	33180	24
26	Processor scheduling	36	102396	25	32214	25	26	Processor scheduling	33	81943	25	32214	25
27	Computer science education	85	59388	26	31794	26	27	Computer science education	64	56953	26	31794	26
28	High performance computing	77	63966	27	31752	27	28	High performance computing	60	59011	27	31752	27
29	Computer vision	33	108486	28	31206	28	29	Computer vision	31	82363	28	31206	28
30	Delay	68	67410	29	30702	29	30	Delay	50	67159	29	30702	29
31	Frequency	57	75768	30	30576	30	31	Frequency	39	75768	30	30576	30
32	Mathematical model	10	202272	31	28350	31	32	Mathematical model	9	184297	31	28350	31
33	Large-scale systems	138	44058	32	26376	32	33	Large-scale systems	105	43723	32	26376	32
34	Performance analysis	86	58506	33	25662	33	34	Performance analysis	66	56449	33	25662	33
35	Bandwidth	40	96348	34	25452	34	35	Bandwidth	34	81775	34	25452	34
36	Data mining	72	65058	35	25410	35	36	Data mining	61	58297	35	25410	35
37	Protocols	35	103194	36	25200	36	37	Protocols	28	87025	36	25200	36
38	Physics computing	116	48006	37	24654	37	38	Physics computing	84	48006	37	24654	37
39	Information technology	99	52920	38	24612	38	39	Information technology	98	45403	38	24612	38
40	Robustness	80	63000	39	24318	39	40	Robustness	68	55315	39	24318	39
41	Ubiquitous computing	205	32508	40	23520	40	41	Ubiquitous computing	173	31081	40	23520	40
42	Biology computing	188	34524	41	23268	41	42	Biology computing	147	34524	41	23268	41
43	Shape	55	76986	42	23100	42	43	Shape	48	68755	42	23100	42
44	Computer interfaces	118	47586	43	22470	43	44	Computer interfaces	91	47167	43	22470	43
45	Mathematics	117	47712	44	22302	44	45	Mathematics	94	45655	44	22302	44
46	Prototypes	78	63966	45	21924	45	46	Prototypes	71	54097	45	21924	45
47	Field programmable gate arrays	66	68754	46	21630	46	47	Field programmable gate arrays	58	61321	46	21630	46
48	Neural networks	45	86478	47	21420	47	48	Neural networks	63	57751	47	21420	47
49	Polynomials	82	61992	48	20958	48	49	Polynomials	56	61951	48	20958	48

Pour finir, on voit que le top de la décennie 2020-2030, a bien été mis à jour :

	A	B	C		P	Q		A	B	C		P	Q
1	keyword	top	totalNbInPaper	topDecade202	nbDecade202	topDecade202	1	keyword	top	totalNbInPaper	topDecade202	nbDecade202	topDecade202
2	Computational modeling	1	860790	1	1202104	1	2	Energy consumption	20	120281	1	54299	1
3	Cloud computing	2	480522	2	123942	2	3	Hardware	4	304787	2	54299	2
4	Task analysis	11	195174	3	111636	3	4	Software	24	113161	3	19165	3

## Mes tests

### La taille des données

J'ai calculé la taille des données auxquelles je m'attendais pour m'assurer du résultat. Avec un script python, j'ai déterminé que le nombre de mots-clefs différents est 130364. J'obtiens 130365 avec le compteur (`Reduce output records`) de mon reducer `KeywordReducer`, il y a une différence de 1 qui est du à une entête que j'écris dans le fichier de sortie.

J'ai aussi déterminé que le nombre de mots-clefs total (avec duplication) qui est 1060969. J'obtiens 1060964 avec les compteurs (`Map output records` et `Reduce input records`) qui montrent les données qui passent du mapper `DecadeMapper` au reducer `KeywordReducer`, il y a une différence de 5 mais je n'ai pas trouvé pourquoi.

### Performances

J'ai testé mon implémentation en augmentant artificiellement le nombre de papiers. Pour augmenter le nombre de papier, j'ai concaténé plusieurs fois le fichier `IEEEdata.csv` avec lui-même. Le fichier est passé de 123490 lignes à 5186580 lignes. J'ai vu les différences suivantes :

- Le premier job `TopDecade` lis beaucoup plus de données et en écrit plus. Il utilise maintenant 49 tâches mapper et toujours 1 tâche reducer.
- Les données manipulées par le mapper et transmises au reducer sont bien plus nombreuses par contre le nombre de données en sortie du reducer `DecadeReducer` reste le même, car il y a toujours le même nombre de mots liés aux mêmes décennies.



- Le temps d'exécution est de 1minute 21secondes ce qui est 1minute plus long que le temps d'exécution avec les données de base qui est 21secondes. Quand j'augmente le nombre de tâches reducer le temps d'exécution diminue, il passe à 48secondes avec 8 tâches (la différence avec les données de base est plus que de 27 secondes). Le fait d'augmenter le nombre de tâches fait peut augmenter le nombre d'octets écrit, mais augmente le nombre de fichiers. Ces augmentations ont un impact très faible sur le job suivant (**TopKeyword**), de moins d'une seconde pour 8 tâches.

```
NB_REDUCE_TASKS=8 yarn jar topkeywords-0.0.1.jar testIEEEdata.csv
decadeTopOutput keywordTopOutput
```

J'ai testé différents nombres de tâches reducer pour trouver le meilleur. Je n'ai fait qu'une seule exécution par nombre de tâches.

nombre de tâches	1	6	8	10	12	14	16	18	20	48
temps en secondes	81	54	48	38	38	36	36	35	34	39

Plus on a de tâches et plus le reducer **DecadeReducer** vas vite. La limite de ce gain de vitesse est certainement le nombre de mapper.

Le second job **TopKeyword** lis est écrit plus de données, car les valeurs sont plus grandes. Il utilise toujours 1 tâche mapper et 1 tâche reducer. Le temps d'exécution ne change pas est reste de 18secondes. Quand on définit plus de reducer au job **TopDecade** il y a plus de fichiers donc plus de mapper, mais le temps d'exécution est le même.

## Performances entre les représentations des données **IntWritalbe** et **Text**

J'ai comparais les représentations intermédiaires des données. J'ai vu une seule différence qui est le nombre d'octets écrit, 21 octets de plus avec la version **IntWritalbe**.

Conteur du nombre d'octets écrit par le job **TopDecade**. Version **Text** à gauche et version **IntWritalbe** à droite :

```
FILE: Number of bytes written=49458190 | → 3+ FILE: Number of bytes written=49458211
```

La différence qui peut exister est due au fait que l'on écrits directement les entiers en binaire et non caractères par caractères, ce qui fait que les nombres qui ont plus de 4 caractères sont écrit sur 4 octets alors qu'il pourrait rentrer largement sur 4 octets en étant sous la forme d'un **int32**. Donc avec plus de données, il pourrait y avoir un impact sur les performances, car plus d'octets seraient écrits. Par contre je n'ai pas vu de différence significative sur la performance. J'ai seulement fait 4 run (2 pour la version **Text** et 2 pour la version **IntWritalbe**).

## Les limitations

Une limitation de ma solution est que je n'utilise pas de **combiner** ce qui pourrait faire diminuer le temps d'exécution.

Pour le cas où on veut ajouter de nouvelles données sans recalculer les anciennes, ma solution charge toutes les données et copie dans le nouveau dossier de sortie, certaines décennies qui ne sont pas modifiées, ce qui peut amener à une perte de performance en lecture / écriture.