Machine learning

Requirement n°	Requirement description	Priority	State	Contributors
13	**Have an algorithm able to retrieve the topic of text files/transcripts and accounting for the noisy nature of transcripts.**	Must	✓	All
	We will create a classic IR algorithm that will retrieve the main topics of the documents in text format. For documents in video/audio format, we will use their transcripts to retrieve the topics.As transcripts are very error-prone, our model must take into account the nosy nature of the documents when retrieving the topics.			
14	**Respect GDPR and other ethical rules**	Must	✓	All
	As we use datasets in order to build the models, it is very important that the project is ethically sound and respects data and privacy rules such as GDPR.			
15	**Create a search model/algorithm**	Must	✓	Vincent
16	**Use elastic search for the search algorithm**	Must	✓	Vincent
	We will create a search algorithm that will be plugged into the web application using the API. It will allow users to search for documents. This search algorithm will be implemented using the library ElasticSearch.			
17	**Have all the libraries be open source**	Must	✓	All
	Mainly use the library from either MIT or APACHE license			
18	**Test our search and topic modelling algorithms through statistical analysis (Benchmarking)**	Must	✓	Vincent
19	**Test the search result quality through a survey**	Should	X	
20	**Test the search results on relevance-labeled data sets**	Should	✓	Vincent
21	**Improve time efficiency on the base models/algorithms**	Should	✓	Vincent
	Make the algorithms (search & topic modeling) quicker than their original counterparts by improving their computational complexity and / or finding more efficient ways (like using libraries in quicker languages) to perform certain operations.			
22	**Have a better content representation that accounts for transcription errors**	Could	X	
23	**Use deep learning /a neural model to improve the topic modeling**	Could	✓	Vincent
	In order to improve the topic modeling algorithm, we could use a state-of-the art model/algorithm that uses deep learning and takes the noisy nature of the documents into account.			
24	**Save a part of the data set to test the accuracy of the topic modeling algorithm against manually labeled data**	Could	X	
	We could save a random part of the X5GON dataset that we would label manually (give up to 5 topics for example). After that, we would test the algorithm against these in order to test its accuracy, to avoid overfitting…			
25	**Retrieve other information than the main topics from the files (tone, number of chapters…**	Won't	X	
26	**Show difficulty estimations of the documents to users (readability score)**	Won't	X	
27	**Estimate the current knowledge level of the user using their learning history**	Won't	X	