What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?

I, Zhou Wu (zjwu2), am working alone on this project on topic modelling, using Python and potentially R.

Topic modelling is an important task in the field of information retrieval. When users such as data scientists and business analysts wish to browse and search through a collection of text for items of value, they do not always know exactly what to search for or what the best approach is to find what they need, especially if the collection had not been curated.

The problem arises when the volume of work is becomes too large for humans to manually inspect. This is where topic modelling comes in: not only does it find latent topics – which serve as a form of summary of the text collection – but in the process, it also creates a model capable of finding similar documents, thereby enabling efficient search through the collection afterwards. In other words, topic modelling not only provides users with useful semantic insights, but also optimizes the subsequent task of retrieving information of value to users.

As the outcome/deliverable of this project, I will develop a program/web app that takes an existing text collection, or scrapes from a list of websites (reddit, sites hosting free ebooks, academic papers, movie reviews, etc.) to create a collection, perform topic modelling using top2vec, create summary visualizations, and allow users to perform queries to find documents under specific topics.

I will be examining a variety of topic modelling techniques including top2vec, LDA, PLSA, and lda2vec, and integrate the best one – or multiple algorithms and allow users to choose – into the program. If time permits, additionally, I would like to implement a document search functionality using the BM25 algorithm, as well as a recommender system based on user input.

I will be using libraries such as nltk and spacy for NLP processing, sklearn for document vectorization, and genism and top2vec for topic modelling. I will evaluate the program by testing on various types of text collections and examining the coherence of the topics found. Moreover, after each instance of topic modelling, I will also run different queries on the text collections to see if useful information is returned the users.

	Hrs	Tasks
1.	3	Develop a script to scrap specific relevant websites
	3	Explore, clean, and preprocess data from various sources
2.	5	Experiment different topic modelling approaches while testing different algorithm parameters
	2	Integrate best algorithm and parameters into code base
3.	5	Develop web app for visualization and topic/document query
4.	(5)	(If time permits) Allow user to select topics or documents/items of interests, and build a recommender system based on user input
5.	2	Write progress report
	3	Write final report (compile results, create infographics)
	1	Clean and package source code / jupyter notebooks
	2	Produce presentation / demo video