Just to want to share a few tips about the first assignment, and hope you have started working on that already. :)

1. One important decision in converting a document to a feature vector is *what should each attribute be*. In class we discussed using each word as an attribute, and the caveat that the total number of distinct words can be huge. This is where feature selection and dimensionality reduction can help. Maybe some words, such as “this” and “that”, are not worth of keeping (examples of so-called *stop words*, of which you can find lists online). Maybe in each document, some words are more important than others, and it may not be necessary to keep unimportant words (hint: TF-IDF and related measures can help quantify word importance). It also has impact on the type of attribute. Will your attribute value be binary, or integer/floating point representing some kind of weight?

2. Does an attribute have to be a word? How about a sequence of two words (*bigram*), or three words (*trigram*)? How will that affect the total number of attributes?

3. Did you do anything to deal with punctuation, upper/lower case, or non-ASCII characters? Some packages such as *NLTK* in Python can probably help.

4. It may take some time for your code to run, so it helps to start early. Don’t worry about having to create a perfect feature vector in one shot, as the definition of “perfectness” is task-dependent and ever-changing.

5. Make sure to uploaded output files to your STDLINUX directory, and set the permission correspondingly. Using 755 for both the directory and the files should work. All "parent" directories (between your home directory and your work directory) need to have execute permission for "other" users (e.g. "chmod o+x parentdirname"). You should include a Makefile, so that a grader can just type the "make" command to run your code.

To be perfectly clear – you do not have to do or have answers to all these questions but please do think about them. The intent here is to give you a real world data analysis task experience where things are often not clearly specified. That said we are readily available to offer guidance or suggestions if you are stuck so please contact us via email or we will be glad to answer questions during office hours. Please see the course webpage ([http://web.cse.ohio-state.edu/~srini/674/](https://email.osu.edu/owa/redir.aspx?C=lRzJv9y8Bkynw_rJamaERUJCgatdotEIt11GB-IbQl3hJERYR_SFhigTTiET75rZH_cAXLpcuE8.&URL=http%3a%2f%2fweb.cse.ohio-state.edu%2f%7esrini%2f674%2f)) for listed times, and email us if you need to arrange a different time.

Important tip: **please get started if you have not already.**