My project proposal is to predict the results of votes in Dáil Éireann based on the features of the preceding debate. The Houses of the Oireachtas provide extensive open data APIs that can be used to collect data on parliamentary proceedings[1]. This includes data on divisions (votes of yes/no/abstain by members) which are linked to their preceding debates by a unique identifier. The data available on debates includes information such as date, time, duration, number of distinct speakers, number of questions, the procedural structure of the debate, identities of the speakers, and full transcript. I propose to transform this debate data into a set of features which will be used to train a machine learning model which will predict the vote outcome. The scope of the study can be limited to the 32nd Dáil (2016-2020) to look at data within a specific period, although this may be refined as the project progresses.

To familiarize myself with the format of the data, I reviewed the interfaces available on the Oireachtas data portal and experimented with accessing the data and determining its structure on a small sample of data (filtering both data sources by using the same dates). I determined that to extract data from this portal I can submit curl commands to retrieve filtered data in JSON format. This JSON contains summary data as well as links to XML documents with a structured representation of the full transcript of Oireachtas proceedings for a given day. The data on each division indicates which section of a particular day's proceedings contains the debate that led to the division.

I believe the work involved in this project would be to write several Python scripts to extract data via the Oireachtas APIs and transform that data into a feature set suitable for input into machine learning models. To do this I would need to determine potential features to use for prediction. I expect the process would be iterative, as knowledge gained from later steps may lead me to return to and refine earlier steps. For example, I can see that there is irrelevant data associated with the division results such as specific votes of each TD, as I would likely only require the tallies and not the identities of those who voted each way.

Once I have generated the feature set, I can use this as training data for supervised learning. I will need to choose a suitable model to train and evaluate. I expect a significant amount of exploration and experimentation would be involved, for example I may choose to restrict the analysis to a certain type of debate (for example debate on Private Members Bills) as I think it's possible that initial analysis could indicate that the debate features are more or less useful as a predictor for particular debate types.

[1] Houses of the Oireachtas publish data on legislation, debates, constituencies, parties, divisions, questions, houses, and members (https://data.oireachtas.ie/)