# Problem Statement

Goal of this project is to build an abstractive summarization algorithm that will take any big text or speech and generate several new grammatical sentences using phrases in those text or speech. Since we are using several different speeches, this algorithm can be classified as multi-document and we will be using a sentence fusion algorithm to actually generate the sentences.

Sentence fusion has traditionally been used for news summarization, gathering many articles that often contain roughly overlapping parts of sentences as each other, but since we are using different inauguration speeches covering a variety of topics from different years, we will not expect there to be several sentences with a great deal of direct lexical overlap. For this reason, we will be using a method of embedding sentences into vectors in order to cluster them topically called Skipthoughts, a recurrent neural network which takes each sentence and tries to predict the next and previous sentences. With these vectors, we can cluster groups of sentences together using k-means on the vectors, and feed these smaller groups to the sentence fusion algorithm to work with. This method we hope will be useful in identifying topically-related sentences, and ultimately be used to produce new grammatical sentences that consist largely of phrases that presidents have actually used in close succession to each other.

Our sentence fusion algorithm will take up these groups of sentences, break them up into dependency structures akin to sentence diagrams, find similarly aligned grammatical structures in other sentences, and put together these parts of sentences together into new sentences that can serve as an abstractive summarization of the original source texts. With this tool in hand, political speechwriters can take inspiration from a group of source documents and find phrases that the imitated speakers tend to use and string together new speeches to sound like someone else. With a view to that end, we plan to use a measure of lexical similarity to evaluate whether the sentences we are producing ‘sound like’ the original speakers.

Data set :

<https://www.presidency.ucsb.edu/documents/app-categories/spoken-addresses-and-remarks/inaugural-addresses>

Project plan :

https://trello.com/b/5DkovNmZ/rajeev201902data-science-capstonedats650110