Report:

In this question we are asked to generate a summary of around 100 words based on a corpus of 3 or more news articles on the same event. To do so we implemented three extraction algorithms: the original SumBasic algorithm, a simplified version which doesn't include the non-redundancy update, and a simple baseline that just returns the first few sentences of one document.
We experimented the three algorithms on four different clusters. The first cluster discuss Google's investment in MILA, the second is about a fire on Parc avenue, the third is about Black Friday and the last cluster is about the recent US elections.
The SumBasic algorithm is a simple heuristic that returns the most probable sentence according to word probabilities. The first step is to compute the word probabilities for each word in all the pre-processed news articles from one cluster. For each pre-processed sentence in the corpus of texts we compute its score by taking the average of all its word probabilities. We add the original sentence corresponding to the best sentence score (best average word probability) to the summary and perform a non-redundancy update. This update squares each word's probabilities that were in the pre-processed sentence in order to not pick the same one again. The algorithm repeats those steps until the limit of 100 words is reached.
We preprocessed all sentences by putting everything in lower case, lemmatizing each words (using the WorNetLemmatizer), and removing stop words and punctuation. This allowed to reduce the vocabulary size when computing the word probabilities and not to differ between upper-case and lower-case words.

The first thing we noticed is that running the simplified version of the SumBasic algorithm produced the worst possible summary. Indeed, without the non-redundancy update the same most probable sentence is picked over and over again until we reach 100 words. This was expected: since we don't update word probabilities, the sentence with the best score will always be the same one.
The original version of the SumBasic produced much more coherent summaries since it has the non-redundancy update. However we notice that the SumBasic algorithm returns more key-phrases instead of a coherent, easy-to-read summary. That is because the sentences forming the summary are often not from the same source document or not in the same order as the source document. This kind of summary may be more informative, yet less grammaticaly correct. This algorithm may not be very robust to contradicting informations between two opposite documents, since we can have two sentences with different opinions using different words but both very probable in their own document. Yet this is not a problem in our dataset.
On the other hand, the baseline algorithm returns the first few sentences from one random document. This produced much less informative summaries since only one document is taken into account, yet they are grammaticaly correct and more “easy-to-read” since written exactly in that order by humans. Furthermore we can assume that information should not be contradicting in that case.
Both SumBasic and the baseline algorithms are producing generic summaries that don't take any point of view and preserve the authors' view.

In order to get the best of both worlds I would slightly change the SumBasic algorithm to also take into account the sentence index in a document as part of it's selection score. For instance at early iterations of the algorithm I would put more weights on sentences at the top of each documents as they often introduces well the topic. As we are adding sentences to the summary, I would slightly move the index weight towards sentences in the middle of each article. This would try to keep the chronological order of the sentences, while making sure that we output high probability sentences.
Another extension that one might think of would be to preprocess the documents by replacing every referent pronouns to the actual referee (noun). This would avoid having sentences with “he said” when we don't know who “he” is refereing to. This seems to be particularly relevent in news articles where we have bits of interviews like in our four clusters.