Task 1: Summarizing A Document Report

1. Present and analyze result of the task:

- After receiving the task, I read the article and 2 related papers this and this. I chose to dive deeper into extractive summarization because the technique is more intuitive than the abstractive one. I implemented an approach utilizing LSA (Latent Semantic Analysis). For this approach, I need to create a matrix representing the input document. I implemented 3 different methods for creating a matrix:
  1. Frequency of words: the cell is filled in with the frequency of the word in the sentence.
  2. Binary representation: the cell is filled in with 0/1 depending on the existence of a word in the sentence.
  3. Root type: : the cell is filled in with the frequency of the word if its root type is a noun, otherwise the cell value is set to 0.

  - For sentence selection, I used 2 methods:
  1. Gong and Liu
  2. Cross method.

  - Then, I combined matrix creation methods and sentence selection methods to summarize a document. The results I got are pretty good. Let's look at a given example in this paper.

  d0: 'The man walked the dog'.
  d1: 'The man took the dog to the park'.
  d2: 'The dog went to the park'.

V^T matrix ($k = 2$)			
	Sent0	Sent1	Sent2
Con0	0.457	0.728	0.510
Con1	-0.770	0.037	0.637

	0	1	2
0	0.5	0.707107	-0.5
1	-0.707107	3.2487e-15	0.707107

- The example used frequency of words for matrix creation and Gong and Liu for sentence selection. They chose concept 0 and sentence 1 which is d1 for the solution. On the left is the result from the paper, on the right is my result. My result is slightly off I think because I removed stop words and normalized words. Therefore, my d0, d1 and d2 look like
- d0: Man walked dog d1: Man took dog park d2: Dog went park
- This explains why numbers for sent0 and sent 2 are opposite.However, Vt[0,2] for my answer is not correct. I got a negative 0.5 instead of a positive 0.5. I guess because the order of rows of the matrix matters when calculating SVD. As the paper explained, how the matrix is created is very important, since it will affect resulting matrices calculated with SVD. Additionally, I did a little tweak to make Vt[0,0] look correct (explanation in video). However, my answer is sent1, and it is the correct answer because its number is the highest cell value for the most important concept which is con0.
- Then, they used the cross method for sentence selection. My numbers are still slightly different because of the reasons explained above. However, I still got the same answer.

V^T matrix ($k = 2$)				
	Sent0	Sent1	Sent2	Avg.
Con0	~~0.457~~ 0	0.728	~~0.510~~ 0	0.565
Con1	~~-0.770~~ 0	0.037	0.637	-0.021
Length	0	0.765	0.637	

Sent0	Sent1	Sent2	Avg.
[5.00000000e-01	7.07106781e-01	0.00000000e+00]	[2.35702260e-01]
[0.00000000e+00	3.24869715e-15	7.07106781e-01]	[1.14723045e-15]
[-5.55111512e-17	7.07106781e-01	2.07106781e-01]	

2. Answer Questions:

- What are current techniques and how well do they work?
+ There are 2 current techniques: extractive summarization and abstractive summarization.
+ Currently, according to the papers the extractive summaries give better results compared to automatic abstractive summaries. Because the abstractive one has to deal with “semantic representation, inference and natural language generation” which is more complex than the extractive summarization, so most automatic text summarization systems are extractive. Among the extractive approaches the algebra-based LSA is well-known. The approach uses SVD to extract the similarity between sentences - sentences, between words - words.

- Can you try to implement (write your code) one or two techniques? Implement at least one method:
+ As mentioned above, I wrote code for an LSA-based approach. The code can be found here.

- Comparison: Compare at least two methods, including at least an extractive and an abstractive one, and see how well they work.
+ Extractive summarization vs Extractive summarization: GongLiu vs Cross
+ I ran many configurations of matrix creation methods and sentence selection methods on the article.txt which contains the text of the given article. Yes, I used a summarizer to summarize an article about summarization. All configurations are: Frequency of words + GongLiu, Binary representation + GongLiu, Root type + GongLiu vs Frequency of words + Cross, Binary representation + Cross, Root type + Cross. I found that Binary representation + GongLiu and Root type + Cross gave decent results. The summaries preserve key informational elements and the meaning of content, and they are linguistically fluent. I put 2 summaries here:
+ One used “article.txt 10 binary gongliu”. It means extracting 10 sentences using binary + GongLiu. The summary can be found in the result folder. The file name is 10_binary_gl.txt
+ This one is ran using “article.txt 10 root cross”. It means extracting 10 sentences using root + cross. The summary can be found in the result folder. The file name is 10_root_cross.txt

+ The 2 summaries are far from perfection. The first one contains duplicate sentences. They contain useless information such as “The code to reproduce the experiments from the NAMAS paper can be found here.”, “ Example output of the attention-based summarization of Alexander et al.” However, the summaries are still pretty good. In the article, the author also gave a summary of the article. I compared 2 summaries with the author's summary. The generated summaries still cover important information the author wrote.

+ One of the cons of extractive summarization is that some 2 adjacent sentences covering the same topic are not meaningful. Suppose the topic is extractive summarization. These are chosen sentences “A typical flow of extractive summarization systems consists of:. Recent studies have applied deep learning in extractive summarization as well”. Additionally, a sentence following another sentence can cause misleading information. “For instance, Sukriti proposes an extractive text summarization approach for factual reports using a deep learning model, exploring various features to improve the set of sentences selected for the summary. The author carried experiments on both single and multi-document summarization tasks to evaluate the proposed model.” Both sentences are not relevant, but the second sentence misleads the information of the first sentence.

+ One of the interesting results I found was that the extractive technique also works well with non English documents. I used a Vietnamese document for summarizing. The summarization was still informative and fluent enough. Now we're going to compare extractive technique vs abstractive one.
+ Extractive summarization vs Extractive summarization: GongLiu and Cross vs T5 — Text-To-Text Transfer Transformer
+ Once again, I used article.txt to run both techniques. The result of extractive techniques explained above. This is the result from T5 using the given code in T5's article.
+ extractive summarization works by identifying important sections of text cropping out and stitching together portions of the content to produce a condensed version. results have shown the method achieved competitive or even better performance compared with baselines.
+ We can see that the summary contains information about extractive summarization, but not abstractive summarization. It doesn't cover all important information. The summary is also very short compared to the extractive summaries. I might have configured it incorrectly. Moreover, the abstractive technique didn't work well with my Vietnamese data. It resulted in nonsense sentences and incorrect grammar. I think this is because T5 is trained with English documents.
+ The running time of T5 is longer compared to the running time of extractive techniques.
+ This is the result of T5 when I used the example from paper "the man took the dog to the park.the dog went to park. he took a walk. the walked dog. The man went there."

- Are you able to reproduce the same or similar results as the original papers?
+ I am able to reproduce the same results as the original papers. My numbers are slightly different, but the final answers are correct (Explained in section 1.)

- Question: How are current methods different? Which technique did you observe work best? How do abstractive and extractive methods generate different results? Do you think the best results were satisfactory for practical usage? What can be improved?
+ The extractive summarization selects sentences directly from the document based on a scoring function to form a summary. Abstractive summarization is like how humans summarize a document; the method tries to understand the concept of a document ,and then generate sentences which are not from the document to summarize it. In each technique, people also use different scoring functions and different Machine Learning and Neural Networks models.
+ I think the best results were satisfactory for practical usage if we talk about summarization in terms of speed. For some documents, the results might not be meaningful. However, the overall performance is still decent.
+ For extractive technique, it should not select too specific sentences. Chosen sentences should be rephrased so that summaries are cleaner, clearer and more fluent. For abstractive technique, it should cover more key information.

3.Summarize 1 paper

- I will use Text summarization using Latent Semantic Analysis
- Its problem: How to extract important information from a document without reading the whole document using LSA.
- 3 key ideas:
+ Introducing text summarization using extractive summarization and abstractive summarization.
+ How to utilize LSA to summarize documents and why using LSA.
+ How to evaluate the quality of summaries and evaluate results produced by LSA.