This is the sample PDF document which I would use. This document should have a combination of simple, complex sentences and compound sentences. For the content, let’s discuss the task at hand and a little bit about myself.

The applicant is Anirban Saha. He grew up in Calcutta, India. He studies in Otto-von-Guericke Universität, Magdeburg. He is writing his thesis with Universität Potsdam, Hasso-Platner Institüt.

**So, what is the task?** The task is to retrieve text from a pdf and make it the knowledge base from which answers would be fetched to users queries. To break it into tasks, let’s broadly break this into two tasks. Task 1: To write a function that takes a PDF file and extracts the text. Ideally, the text should be stored in a database. But for this, we would store it in a text file. We will call this the knowledge base. Task 2: To write a function that takes a query from the user, find the answer to the query from the knowledge base and return it to the user.

Let’s get into the details of each task now.
For **task 1**: There are multiple open source python packages that can be used for this. We would try to use Apache Tika. For now, we will focus a little less on this and more on task 2.

For **task 2**: There are various ways that we could implement a cognitive search operation on the knowledge base. For this task, we would try the following three different approaches:

1. Allen NLP
2. Huggingface
3. CDQA

**Comparison**: For each of the three, we would record the answers retrieved to a fixed set of questions and the time taken for the process.

The final document would be submitted to Recobo on 24.02.2021 by email. The document should contain the analysis and the links to the Google Colab notebooks using the following format:

**Links:**

| | **Approach** | **Link to the program.** |
|---|---|---|
| 1 | AllenNLP | https://colab.research.google.com/drive/1d-CWmHjxmmD1q5EmAIiybbkOFrdEi3Vu?usp=sharing |
| | | |