A pdf summarizing project I've begun to develop. I'm positive there are better and well-established programs that do this but making my own has always seemed more exciting.

I like reading and learning about my field and recently I've been wanting to branch further into the world of published papers on topics of interest. The roadblock to my learning is a lot of important, peer-reviewed papers are dense with language (and rightfully so!). Rather than wrestle with this un-trained skill while trying to learn a dozen others I turned to Python for help!

Using PyPDF2, FPDF, and 🤗(Hugging Face) AI I've managed to get a satisfactory "study guide" companion for a paper.

Based on the structure of the pdf the information is broken down either page-by-page or paragraph-by-paragraph and loaded into a new pdf.

Now, pdfs are notoriously difficult to work with but not so notorious that I knew that going in. So, in lieu of a full-fledged summarized copy of the pdf things like images, charts, code-blocks are simply ignored. That's why I say the resultant pdf is to be used in tandem with the original document.

Another thing about pdf difficulty is getting the encoding right. Things like apostrophes and double-quotes are unable to be read by FPDF (or I'm still working on it), so they are removed and replaced by an asterisk. The printing is often choppy and odd spaces appear, essentially, the resultant pdf needs a lot of editing to be pleasantly legible.

I like to talk a lot and this post is long-ish, so I figured I could make it into a a pdf, run it through my summarizer, and provide the results to make it faster for you to read.

Here we go!