**Explainer**

I have used a flask server so that the user can upload files via a webpage. The architecture is designed in an extensible way. Currently there is just one file type supported(.txt). It can be extended to include other types such as pdf, docx and so on.

Major libraries used

Nltk

Flask

Scalability

1. Instead of flask server we can use an asynchronous server like Tornado which can help scaling up the number of request we can process at a time
2. We can scale out using kubernetes+knative stack along with Istio service mesh. Knative provides serverless capabilities along with traffic routing and load balancing on top of kubernetes. Can scale the number of pods based on the requests load and will scale down to zero if no requests are there.

Bonus tip: The above approach still has some aspects of devops in it as we will have to maintain the kubernetes cluster. In order to completely scale back devops, we can go with a managed Kubernetes service like Google Cloud Run. It’s built on top of knative and takes the best parts from AWS lambda. So you only pay for compute time, not idle time.