Semantic Caching Demo

1. Go over PowerPoint
2. Show **lab-semantic-caching** resource group
3. Show **foundry1-yjtp3yjn5hcpa**
    a. Briefly explain foundry
    b. Show the chat completions / embedding model
    c. Show deploying a new model
4. Show Azure APIM
    a. Brief overview of API section
        i. Many operations but only these ones being used:
            1. POST /deployments/{deployment-id}/chat/completions
            2. POST /deployments/{deployment-id}/embeddings
        ii. Go over the policy
    b. Show backends
        i. Embeddings-backend
        ii. Foundry1
5. Go through Jupiter notebook
    a. Run through it as is
    b. Run **check-redis-cache.py**
        i. Show vector representation / prompt storage
        ii. Show that only **one prompt** I stored
    c. Run **clear-redis-cache.py**
    d. Change APIM threshold score to **0.05**
    e. Run **check-redis-cache.py**
        i. Show vector representation / prompt storage
        ii. Show that only **four prompts** are stored
    f. Run **clear-redis-cache.py**

Rate Limiting Demo

1. Go over PowerPoint
2. Show policy in VSCode
3. Run through Jupiter Notebook