Python project part 2 :

- tf idf of question  
- get corpus of document with most words in common with question   
- Generate answer based on selected document

~~Step 1:~~

~~Turn question into list of words in the same way as documents : lowercase and no punctuation : function takes question as parameter and returns it as a list of words~~

~~Step 2:~~

~~Function that identifies terms in the question that are also in the document corpus. Ignore terms absent from corpus (bc no associated tf-idf values)~~

Step 3:

N (nb documents) rows and M (nb words in doc) columns   
- give each word in question their tf score. Words in corpus that are not in question get tf score = 0  
- give each word in question the same idf as word in corpus   
--> calculate tf-idf of words in question with that

Step 4: