



## Project 5: tf-idf

For this project you are going to implement the tf-idf algorithm for a couple documents. Each document is a short summary of a prominent technology company. These documents are given to you in the form of a text file with the punctuation removed. To get the words in each document all you have to do is read each file and split it up by spaces. For the purposes of this project you don't have to do anything more than that, like checking if a word exists in the english dictionary, etc. Just split by spaces, and that will be enough.

In a file called `proj5.py`, your task is to implement the tf-idf formulas to report the top 5 words in each document with the highest tf-idf scores. Make your output similar to the following. Here are the top five tf-idf scores for each document that I computed:

### Top words in document **apple.txt**

Word: Mac, TF-IDF: 0.02134  
Word: iOS, TF-IDF: 0.01909  
Word: Store, TF-IDF: 0.01909  
Word: TV, TF-IDF: 0.01685  
Word: Apple, TF-IDF: 0.01283

### Top words in document **facebook.txt**

Word: people, TF-IDF: 0.03308  
Word: Instagram, TF-IDF: 0.01743  
Word: Messenger, TF-IDF: 0.01743  
Word: share, TF-IDF: 0.01323  
Word: world, TF-IDF: 0.01162

### Top words in document **google.txt**

Word: Google, TF-IDF: 0.02419  
Word: Chrome, TF-IDF: 0.01345  
Word: engaged, TF-IDF: 0.01345  
Word: advertisers, TF-IDF: 0.01345  
Word: properties, TF-IDF: 0.01345

### Top words in document **microsoft.txt**

Word: Office, TF-IDF: 0.01825  
Word: Xbox, TF-IDF: 0.01642  
Word: server, TF-IDF: 0.0146  
Word: Server, TF-IDF: 0.01277  
Word: Windows, TF-IDF: 0.01143

### Top words in document **tesla.txt**

Word: energy, TF-IDF: 0.03737  
Word: solar, TF-IDF: 0.02242  
Word: electric, TF-IDF: 0.01993  
Word: Model, TF-IDF: 0.01744

Word: generation, TF-IDF: 0.01495

---

Assigned on 5/10/2017 12:01:00 AM | Due on 5/23/2017 11:59:00 PM

500 Submissions Remaining



**Looks like you don't have any submissions yet**

Click submit to submit your code!

[Submit](#)

[Download Starter Code](#)

[Open IDE](#)