CS 482 Machine Learning

Homework Questions

Text Based Learning

Given the following set of two documents, containing words of Albert Einsten

“Two things are infinite: the universe and human stupidity; and I'm not sure about the universe.”

“God does not play dice with the universe”

a) Construct a bag of words and vocabulary.
b) Remove all stop words from the English stop word list of scikit-learn (It can be found here https://gist.github.com/ethen8181/d57e762f81aa643744c2ffba5688d33a
c) Calculate tf-idf for the remaining words
d) Remove 5 words with the lowest tf-idf