StackOverflow Helper

Kevin Du

Introduction

- StackOverflow is an online forum where you can ask and answer questions about computer programming
- Project - build a web app that takes in a string of user-input keywords and uses NLP to find the answers that most closely matches the keywords
- Also identifies the “power-users” or gurus of a particular topic
- Complementary to StackOverflow search
  - Searches answers rather than question titles
  - Works offline after you download the dataset

Dataset

- Published on Kaggle in October 2016
- Around 500 MB in .csv format
- Includes almost 1 million StackOverflow answers with Python tag
- Data include votes, user IDs, answer text, and date posted
- Narrowed down to 100k answers by setting a filter of vote score > 4

Modeling

- CountVectorizer and TF-IDF
- English stop words
- Token pattern: at least two consecutive letters
- Minimum document frequency = 10
- Evaluated with cosine similarity
    - Slightly higher for TF-IDF than countvectorizer

Live demo time!

May the demo gods smile upon us.

Future improvement

- Try using word2vec or deep learning
- Weigh in the score of the answers
- Continuously update new answers into the model
- Improve functionality of web app