

Stack Overflow Recommendation System

Recommending based on distance and
similarity

Crystal Han



Introduction

Objectives

- Topic modeling for question
- Similar question recommendation

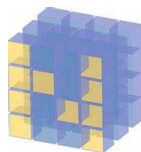
Algorithm

- NLP Unsupervised Learning
- Cosine Similarity



Methodology

- Tools:
 - BeautifulSoup
 - Google Cloud Big Query, SQL
 - Scikit-learn, NLTK



NumPy



pandas

matplotlib

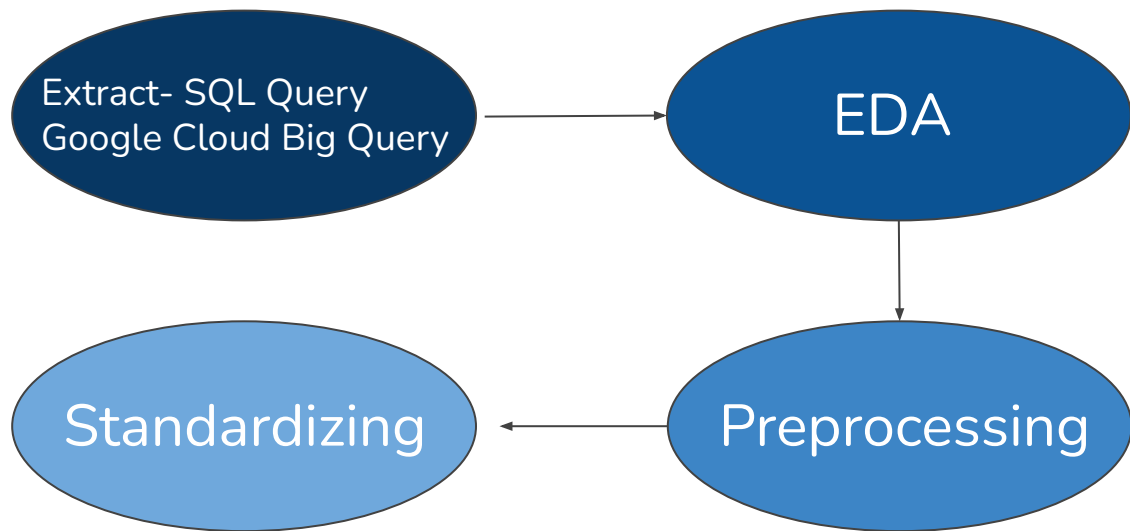


seaborn

Data Extraction



Special Characters, Code,
Links

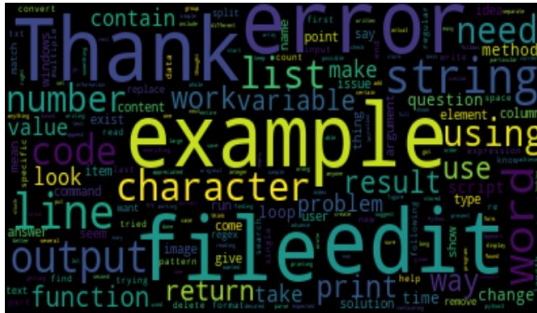
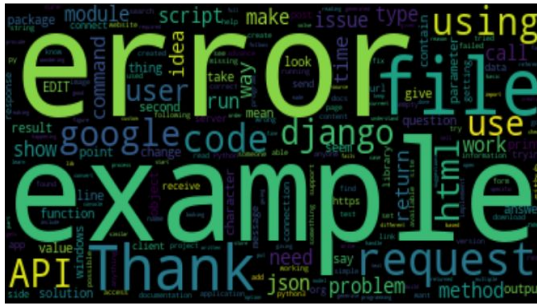
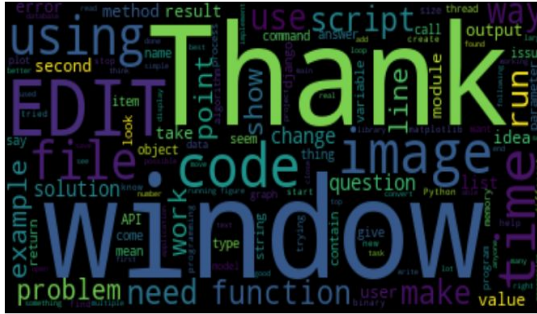


Document-Term Matrix
CountVectorizer & TFIDF

Pipeline: spaCy
-Part of Speech Tagger
-Lemmatizer

- SQL query from Google Cloud Big Query
- Combined 3 tables:user, questions, answers
- 10M+ data points collected -> filtered to 70k
- 'Python' in title/question

Topic Wordcloud



- LDA
- 3 top topic word clouds
- Tfidf Vectorizer

Topic 0:
image it code time plot using
data my have want process
there window file way use is
images do be thread graph
program script need memory
size make function get

Topic 0:

image it code time **plot** using data my have want
process there **window** file way use is **images** do be
thread **graph** program script need memory size make
function get

Plotting, graphing, images

Topic 1:

server request client send com using **API** https
requests code response error http my google **api** app
data get connect it connection message **django** trying
Google use post library side

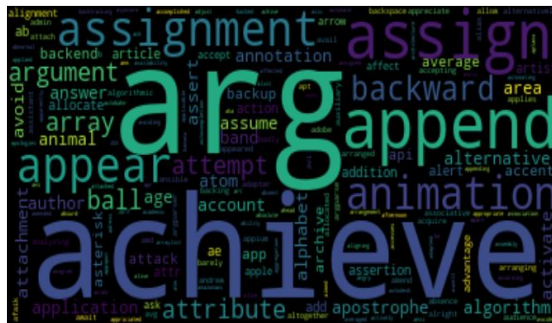
Flask, Django, Server, API, App

Topic 2:

string text file **line** word **regex** **characters** it words
want **lines** have output strings character list **split** code
match using **replace** **remove** number expression need
search get extract **txt** pattern

Filtering

- LDA
- Count Vectorizer



Topic 0:

and andrew adapter avi **arg** b adobe applies
alternatively backing arranged avail badly **algorithmic**
backspace ax **analyzing** amend affects achieve **avg**
alright arranging assistant appreciate append attach
apostrophes allocated allow

Analysis

Topic 1:

type asp adapt asterisk adb b aa amazon aspect
adapter aspx ahead amazing applies alphanumeric
accessing accidentally adobe allow addr andrew
address appended affecting await abc ax arguments
ansi atleast

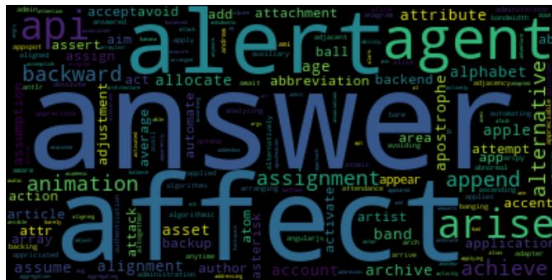
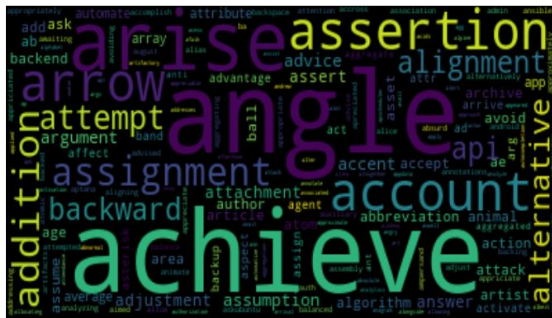
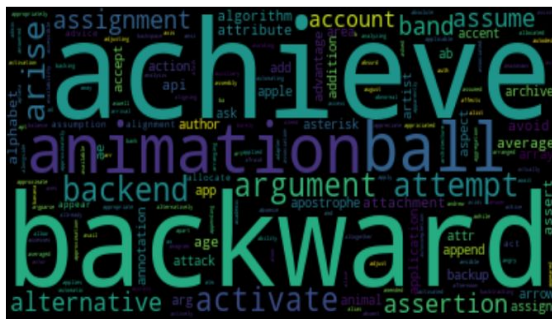
Filtering and Amazon Access

Topic 2:

august automating algorithmic answer balance asked
 actor andrew balanced backing answered
 alternatively aptana auth activity ant attention adapter
 appdata b asn askubuntu appropriate alchemy
 appreciate associated allow alert aware
 arduino

Algorithm, Authentication

- NMF
- 3 top topic word clouds
- Count Vectorizer



Topic 0:

andrew allocated adapter affects backspace applies
avail august alternatively ax analyzing backing
achieve actually backward barely animation b
apostrophes appreciate ball appeared aimed
arranging avi attempt appending backend appropriate
approximate

Analysis, App

Topic 1:

ant anti aggregated aggregate alternatively backing
advised august backed angle aptana ampersand apis
auth allow android ba appropriately appropriate
appriciate alice backspace achieve adjust arise
attempted advise askubuntu aimed affects

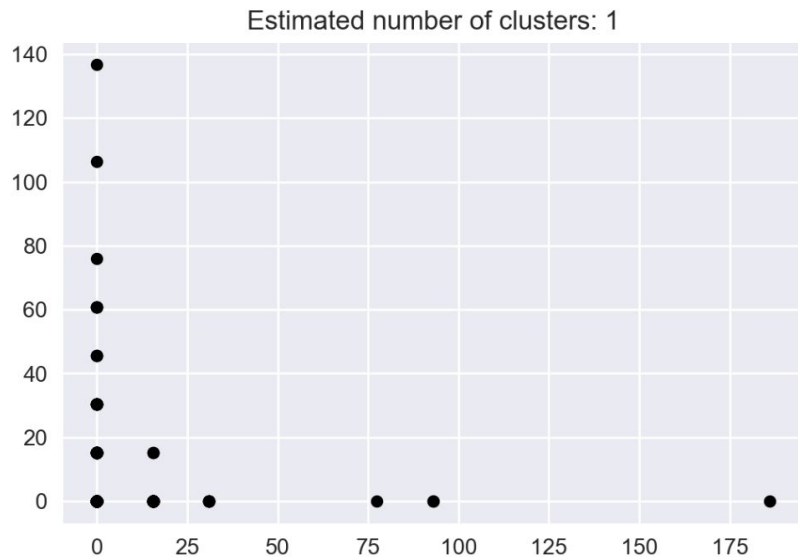
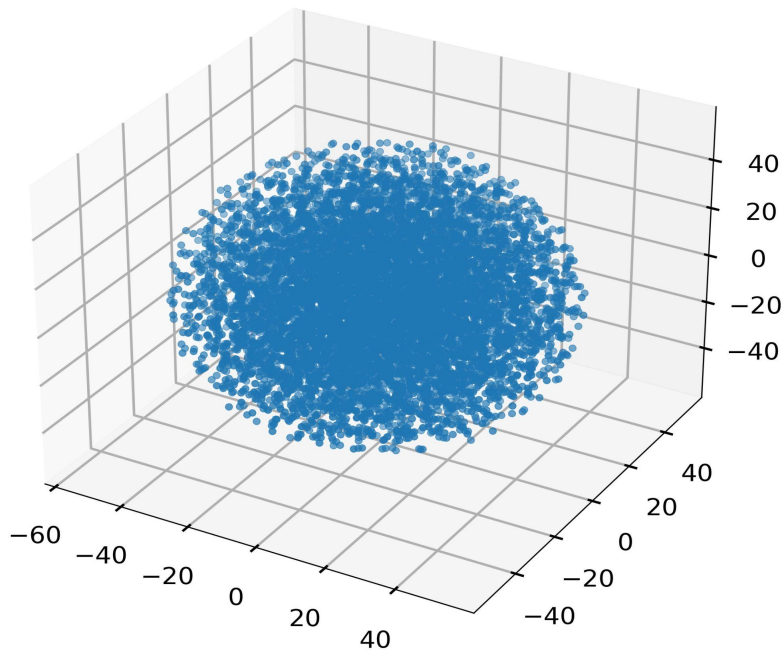
Authentication

Topic 2:

algorithmic algorithms automating artists answer
administrator bands answered aware appspot affect
andrew alternatively aptana aligned b arrive apply
assigns backing ascending applies appreciated arcpy
adapter bare appreciate aaa adjacency bandwidth

Algorithm/Automation

T-SNE of the Doc-Term Matrix & DBSCAN





Content Recommendation

Business can recommend other questions based on the content and question asked.

Topic: API

Top 5 Similar Topics:

Use, get, google, response,
request

Topic: Django

Top 5 Similar Topics:

Database, py, app, project,
server



Future Work

- Streamlit App
- Apply analysis to more topics
- Improve visualization



Conclusion

- Can find similar topics based on the question asked
- Find the closest question/topic based on the tags and words



Appendix

- Google Cloud Big Query
 - Stack Overflow
 - 10M+ data points
 - Combined 3 tables: User, Question, Answer
- Data Cleaning:
 - Combine question and its paragraph
 - Comb through words and filter for specific words

Thank you

