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

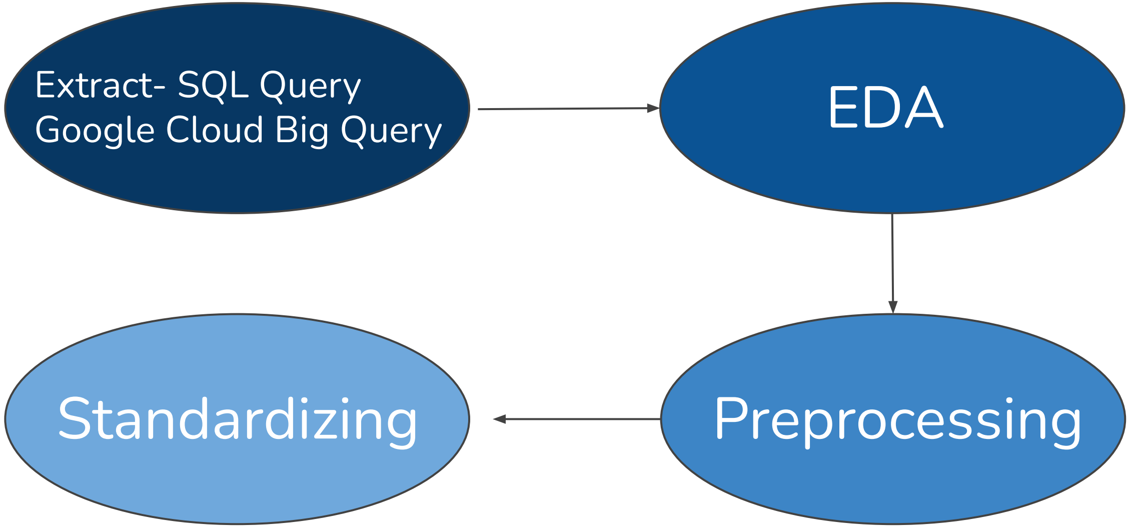
- Tools:
    - BeautifulSoup
    - Google Cloud Big Query, SQL
    - Scikit-learn, NLTK

NumPy pandas matplotlib seaborn

Data Extraction

stackoverflow

Special Characters, Code, Links



Extract- SQL Query
Google Cloud Big Query

EDA

Preprocessing

Standardizing

Document-Term Matrix
CountVectorizer & TFIDF

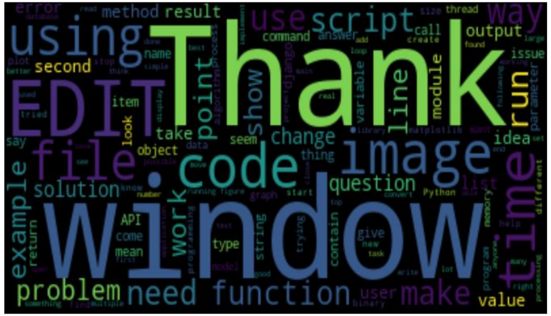
Pipeline: spaCy
-Part of Speech Tagger
-Lemmatizer

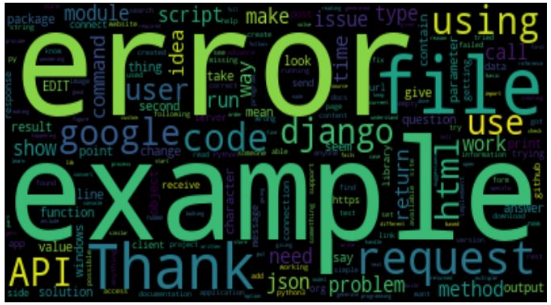
- SQL query from Google Cloud Big Query
- Combined 3 tables:user, questions, answers
- 10M+ data points collected -> filtered to 70k
- ‘Python’ in title/question

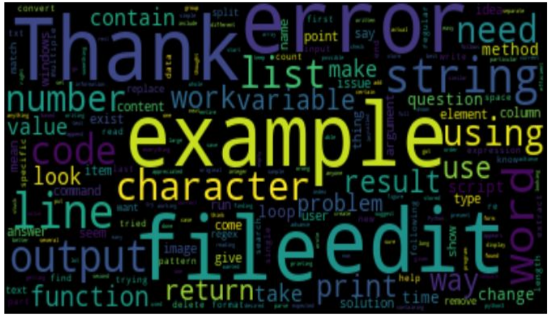
Topic Wordcloud

- 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

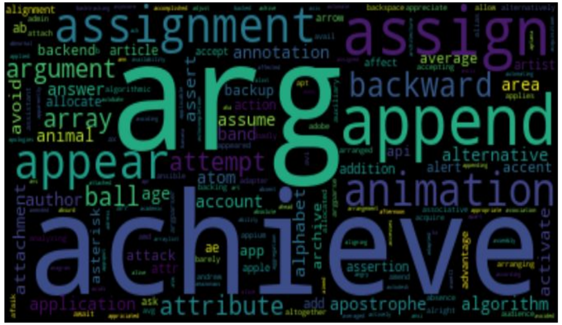
Flask,Django, Server,API, App

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

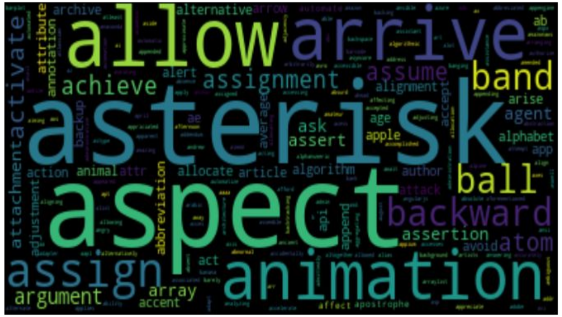
Topic Wordcloud

- LDA
- Count Vectorizer



Topic 0:
amd 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

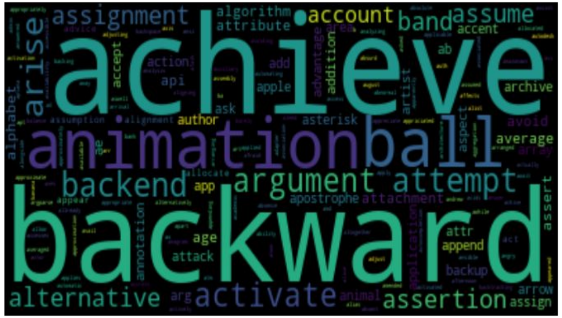
achieve
alert
backend
attempt
animation
atom
append
avoid
arise
assignment
account
algorithm
artist
animal
assign
affect
ball
attribute
asterisk
app
array
backward
assertion
area
accent
apostrophe
assume
band
advice
apple
assert
addition
attachment
alternative
argument
alignment
alphabet
archive
adjustment
abbreviation
annotation
allocate
arrive
backup
automate
author
attack
api
age
ab
adapter
balance
advantage
accept
add
ask

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

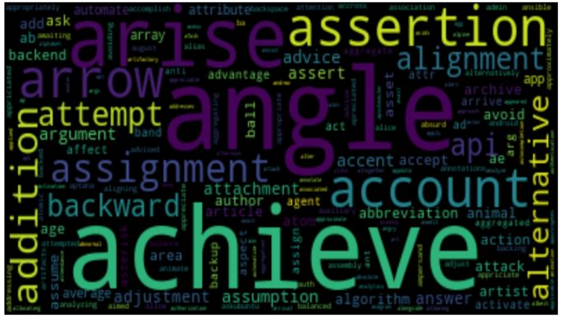
Topic Wordcloud

- NMF
- 3 top topic word clouds
- Count Vectorizer



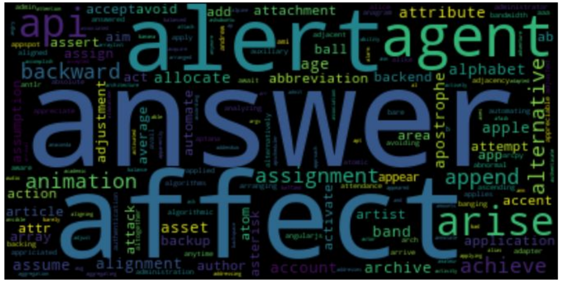
Topic 0:
andrew allocated adapter affects backspace applies avail august alternatively ax analyzing backing achieve actually backward barely animation b apostrophes appreciate ball appeared aimed arranging avi attempt appending backend appropriate approximate

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

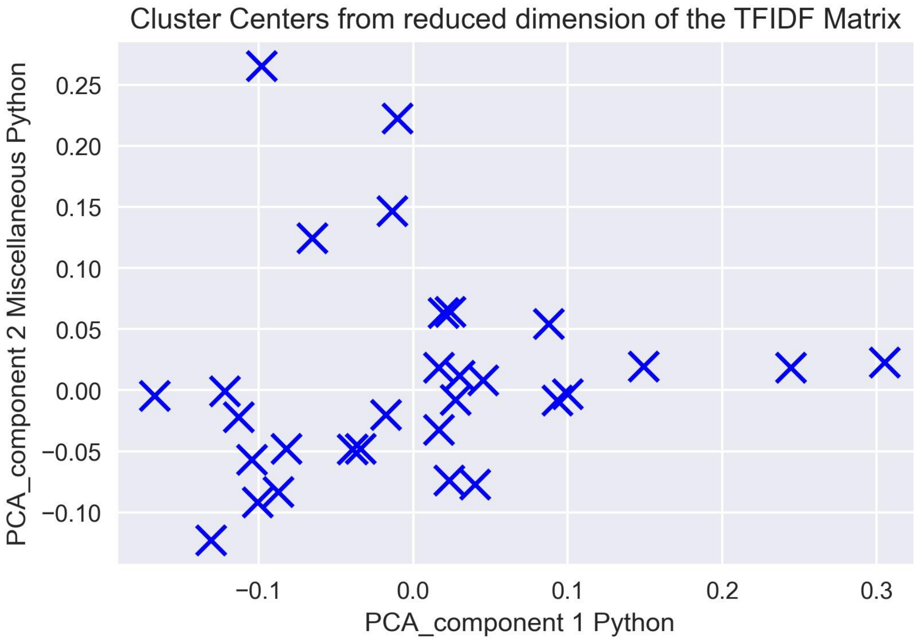
Authentication

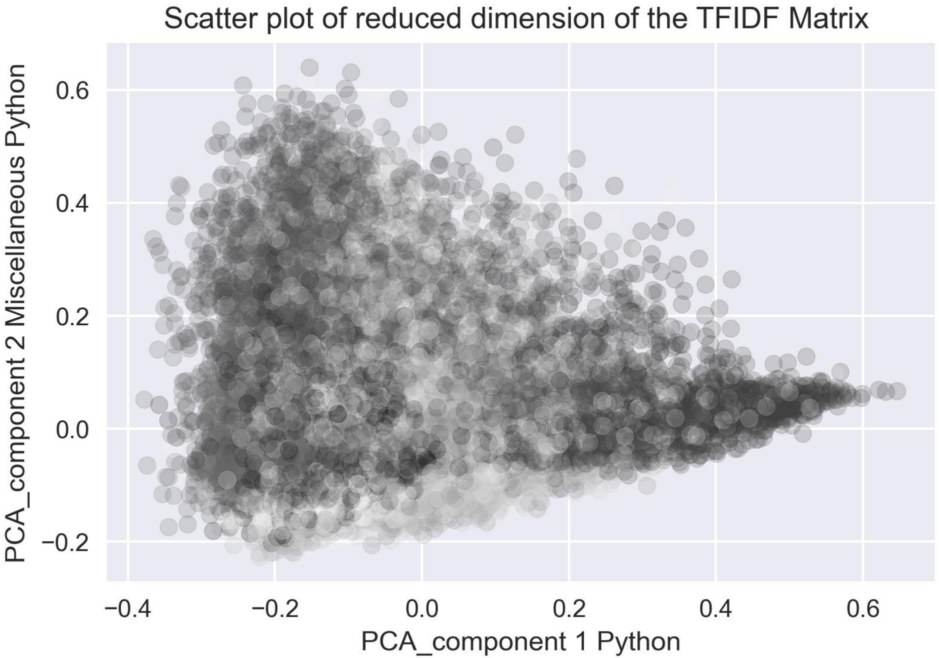


Topic 2:
algorithmic algorithms automating artists answer administrator bands answered aware appspot affect andrew alternatively aptana aligned b arrive apply assigns backing ascending applies appriciated arcpy adapter bare appreciate aaa adjacency bandwidth

Algorithm/Automation

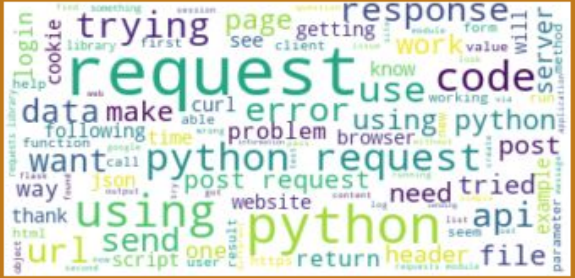
KMeans

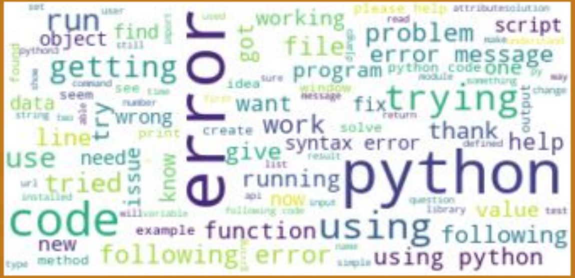

Cluster Centers from reduced dimension of the TFIDF Matrix
PCA_component 2 Miscellaneous Python
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.1
0.0
0.1
0.2
0.3
PCA_component 1 Python

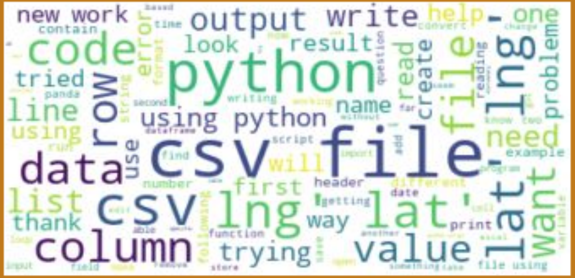

Scatter plot of reduced dimension of the TFIDF Matrix
PCA_component 2 Miscellaneous Python
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4
0.6
PCA_component 1 Python

KMeans Word Cloud

- Using TFIDF Document Term Matrix
- Cluster 0 consists of request, errors, response, api -> most likely web app related such as Flask, Django
- Cluster 1 consists of errors, functions, messages, script-> likely python function error related
- Cluster 2 consists of file, csv, row, list -> related to python taking in files, writing, cleaning, and filtering







Scores

Homogeneity Score	Silhouette Score
1.0 (didn’t make sense)	0.0271

- Homogeneity Score: 0 to 1, 1 stands for perfect homogeneous labeling
- Silhouette Score: -1 to 1, -1 being the worst score, values near 0 indicate overlapping clusters

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
- Based on these clusters, may be able to assign closest topic or topics from the cluster to the user

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
- Clsutering Algorithms:
    - K-means
    - DBSCAN

Thank you