Information Retrieval and Ranking Datasets based on Similarity Measures

Vaihhav Reddy K (1213083974)

Arizona State University

Team Members

- Sai Sudarsan Gollanapalli
- K Divya
- Vaibhav Kalakota
- Harish Bhutra
- Riya Saxena
- Gaurav Mittal

Abstract:

There are various measures that can be used to compare and rank the data. The data that is ranked can be obtained from various attributes. Some of them can be the textual data, some visual data etc. Based on these, the representations of color, time, semantic concepts, and other underlying information can drastically differ from one another in the case of Multimedia data. People can be a significant factor in the judgement. Example is the data that is obtained from them through their perception. The current project aims to experiment with text and image features, vector models, and similarity/distance measures. The data is obtained from the image sharing app - Flickr. Various of the users uploaded some of the images in their Flickr account. Through this, both the visual information and additionally the literary information is gotten. Utilizing the literary descriptors of the information, we discover the comparability between picture highlights.

Introduction:

The dataset which was initially gathered from Flickr utilizing the name of the area as inquiry, has Textual descriptors, visual descriptors alongside different descriptors The closeness between two pictures relies upon the how far the picture highlights are in the vector space. Picture highlights incorporate it's literary depictions which incorporate the user labels, remarks, portrayal of the pictures, Visual descriptors of 10 models and User comment believability descriptors.

A comparison between the data's from text as well as the visual data is below -

Normal Data:

taiwan" 1 15 0.06666666666666667 "third" 1 66 0.015151515151515152 "tour" 1 169 0.005917159763313609 "tower" 1 805 0.0012422360248447205 "vario" 2 16 0.125 "vision:mountain058" 1 1 1.0 "vision:outdoor0979" 1 1 1.0 "vision:sky0711" 1 1 1.0 "women" 1 26 0.038461538461538464 "國" 2 15 0.13333333333333333 "塔" 2 15 0.13333333333333333 "塞" 2 46 0.043478260869565216 "巴" 3 45 0.06666666666666667

Image data from various metrics:

angel of the north	4.85E+09	0.313432	0.262579	0.002287	0.226021	0.171794	0.000937	1.025359	1.11332	-0.24428
big ben	5.06E+09	0.251238	0.310628	0.002223	0.239145	0.17506	0.000933	2.470193	0.690616	-0.17429
hearst castle	3.9E+09	0.356659	0.315109	0.002075	0.211574	0.205501	0.001027	0.3826	0.839563	-0.21384
la madeleine	4.29E+09	0.258194	0.312429	0.00204	0.209511	0.176039	0.000855	2.880717	0.901565	0.015209
pont alexandre iii	4.43E+09	0.24735	0.228768	0.001805	0.198438	0.154496	0.000894	1.776605	1.194554	0.301015
neues museum	2.76E+09	0.373244	0.287769	0.002252	0.231821	0.198929	0.001036	0.350101	0.97054	-0.37684
montezuma castle	5.71E+09	0.316559	0.260911	0.001871	0.2333	0.187671	0.001048	1.093235	1.267811	0.152666

System Used:

Python 3.6 Release for Windows on Pycharm. Also used Juniper Notebooks

Data Format:

Task 1:
Number of unique users = U
Total unique words = W

Using this we formed a U*W data frame using these variables

Task 2:
Number of unique images = I
Total unique words = W

Using this we formed a I*W data frame using these variables

Task 3:
Number of unique users = U
Total unique words = W

Using this we formed a U*W data frame using these variables

Task 4 & 5:
No of Images = 300 (I)
No of Locations = L
Total no of comparitions for locations = L*I

Example dataframe: (L*I)

ce	greece	view	407	421	421bce	ancientgreece	...	ä®®	æµ®	é¢¨	ä¤□	æµª	æ¼«	vision:mountain058	vision:outdoor0979	vision:sky0711	vision:sky0662
0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0

Description:

Task 1:

The objective of the undertaking 1 is that when given a user ID, a model (TF, DF, TF-IDF), and esteem "k", the most comparable k users in view of printed descriptors must be returned. Also, the general coordinating score and in addition the 3 terms that have the most noteworthy comparability commitment ought to be shown Query which lists a 'k' similar users and the 3 terms with highest score contribution per similar user given a user ID and a model

- Parse the input file
- Get the csv out of it
- Create a data frame of Users and Unique words
- Do similarity calculations
- Get the k nearest Users and 3 most weighted words

Task 2:

The objective of the task 2 is that given a picture ID, a model (TF, DF, TF-IDF), and esteem "k", the most comparable k pictures in light of literary descriptors must be returned. Also, the general coordinating score and in addition the 3 terms that have the most elevated similitude commitment must be computed and appeared to the users.

- Parse the input file
- Get the csv out of it
- Create a data frame of Images and Unique words
- Do similarity calculations
- Get the k nearest images and 3 most weighted words

Task 3:

The goal of the task 3 is that given a location ID, a model (TF, DF, TF-IDF), and value "k", the most similar k images based on textual descriptors must be returned. Additionally, the overall matching score as well as the 3 terms that have the highest similarity contribution must be calculated and shown to the user.

- Parse the input file
- Get the csv out of it
- Create a data frame of Locations and Unique words
- Do similarity calculations

- Get the k nearest locations and 3 most weighted words

Task 4:

The goal of the errand 4 is that given an area ID, a model (one among CM, CM3x3, CN, CN3x3, CSD, GLRLM, GLRLM3x3, HOG, LBP, LBP3x3), and esteem "k", the most comparative k areas in light of the relating visual descriptors of the pictures must be returned. Moreover, the general coordinating score and in addition the 3 picture combines that have the most elevated closeness commitment must be figured and appeared to the user.

- Parse the input file based on the model given
- Get the csv out of it
- Create a data frame of Locations and Unique words
- Do similarity calculations
- Get the k nearest locations and 3 most weighted words

Task 5:

The objective of the task 5 is that given an area ID, a model (one among CM, CM3x3, CN, CN3x3, CSD, GLRLM, GLRLM3x3, HOG, LBP, LBP3x3), and esteem "k", the most comparative k areas in light of the relating visual descriptors of the pictures must be returned. Moreover, the general coordinating score and in addition the 3 picture combines that have the most elevated closeness commitment must be figured and appeared to the user.

- Parse the input file based on the model given
- Get the csv out of it
- Merge the CSV of all the models

- Create a data frame of Locations and Unique words
- Do similarity calculations
- Normalize the data
- Get the k nearest locations and 3 most weighted words

Conclusions:

In this report, we have presented the queries to find the similarity between image features using distance metric and textual and visual models. Textual descriptors and visual descriptors are used to calculate the similarity between the images, similarity between the users who uploaded the images and the similarity between locations at which the images are taken are calculated.

BIBLIOGRAPHY

1. Bogdan Ionescu, Adrian Popescu, Mihai Lupu and Henning Müller, Div150Cred
2. Data Management for Multimedia Retrieval, Canden