RedditTextAnalysis

Jianyu Li

May 9, 2014

```
Populating the interactive namespace from numpy and matplotlib
```

Reddits is a social entertainment website where people create posts to share pictures and stories with other redditors who share their interest. To ensure posts will reach the desired audience, redditors will self select a category to place the post. These categories are called subreddits. Each subreddit is a small community centered aroudn a specific topoic.

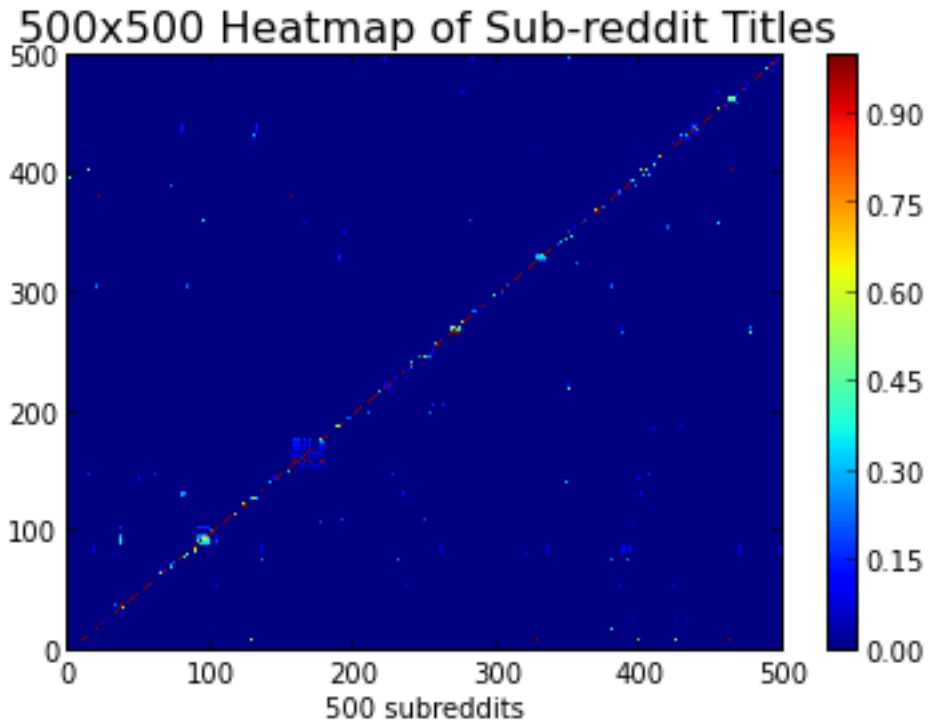
Names of subreddits are generally stand alone word that summarize the topic it contains. Examples of subreddits are: California, NASA, fashion, zombies, etc. If we want to find interesting facts about California, we should check the subreddit California.

There are over 300 thousand such subreddit communities in toal, both active and inactive. Some subreddits such as serendipity has a high subscriber base and is frequently updated. On the other hand, the most recent reddit post for Alteil, an indie game with a low player base, was created a month ago. There is a huge disparity in how much attention certain subreddits get.

This project aims to lower the attention disparities between subreddits by pairing up subreddits that talk about the same topic. Take a look at the names of the different subreddits. They are the very first thing people have to know in order to get to the subreddit page. Intuitively, names of subreddits should be able to tell us a lot about the contents of the subreddit.

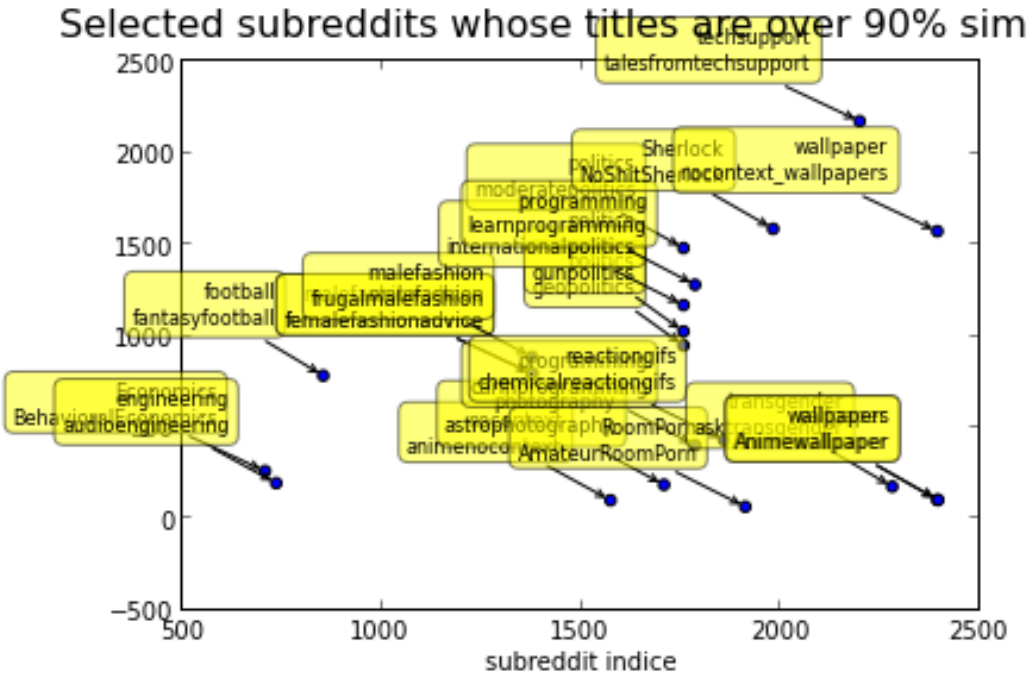
The heatmap below illustrated how similar a small subset of subreddit names are. This is done by examining all possible strings that can be form by a given name, and calculate how much overlaps are there between any two given names. We can see a lot of blue on the off diagonal and patches along the diagonal. This indicates two subreddits have names that are similar when they are close (mainly due to subreddits are sorted in alpabetical order).

(Please see appendix for a overall picture of the 2500x2500 subreddit similarity matrix)



There are also a few exceptions to this as there are small patches of light blue that are away from the diagonal. If we are willing to make the assumption that if two subreddits have names that are spelled the same way, the two must be related in some way, we can use subreddit names as a way of pairing.

Let's take a closer look at the subreddits thta are considered to have similar names. Set an arbitary threshold of 90% and only examine those subreddits that are off the diagaonal.

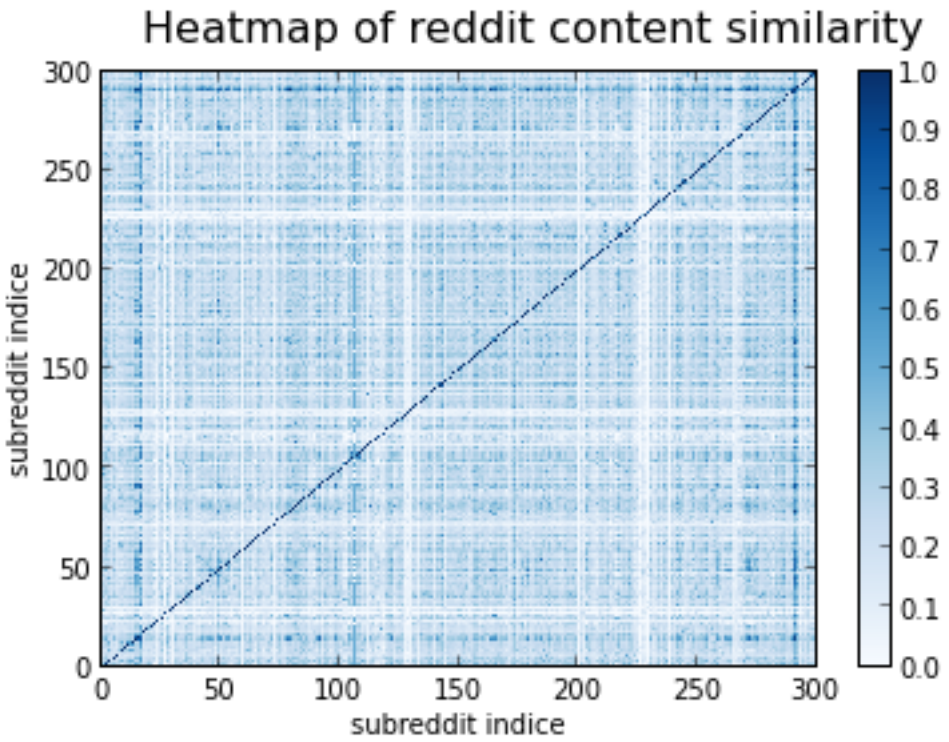


The names that we classified as similar share some features in common: one name tends to be very specific and the other name is broader or a generalization of the prevous one: politic and political checking; malefrugalfashion and femalefrugalfashion, etc. This suggests names of subreddits can be a good tool for pairing.However, how sure are we to say that the contents of subreddits malefashion and femalefashion are the same just by looking at their names. The names reactiongifs and chemicalreactiongifs are identified as a match. Judging from experience, reactiongif should be refering to human reaction toward certain events while chemical reaction gif is related to chemistry or chemistry photography, despite the two names do spelled very similarly. This is one downfall of the subreddit name approach because it fails to detect how the words are used in the context of a compound word.Let us look past the names of subreddits and examine the contents.

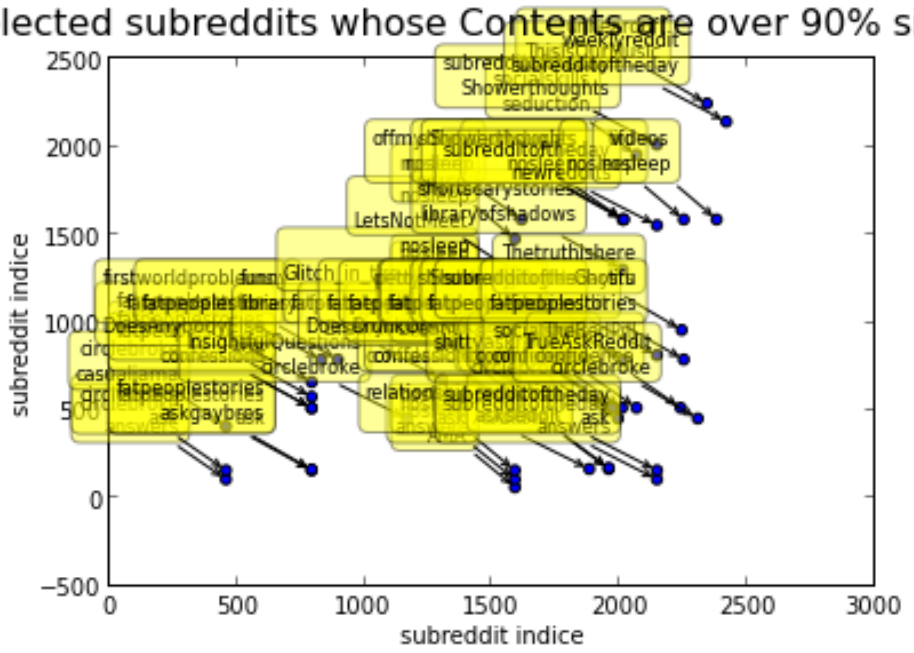
From early exploratory analysis, there is a non-negligible number of image based subreddits. These are subreddits with posts that only contain images. Therefore, looking at the contentin the body of each subreddit post will introduce bias that favors the text based subreddits. To resolve the problem, we will limit ourselves to just looking at the title of each post.

Two methods of content similarities are compared: Inverse IDF and the simple word count. The Inverse IDF approach weights down words that repeatedly appears and weight up the words that appear only occasionally. Thus, the Inverse IDF approach was eliminated since what makes a subreddit unique may be the repeated occurance of certain jargons (See appendix for comparisons).

Below is a heatmap of the similarity scores for 300 subreddit pairs.



Take a closer look at the subreddits that are considered 90% similar in their contents.


Selected subreddits whose Contents are over 90% similar
subreddit indice
subreddit indice
2500
2000
1500
1000
500
0
−500
0
500
1000
1500
2000
2500
3000
weeklyreddit
subredditoftheday
Showerthoughts
seduction
videos
nosleep
shortscarystories
libraryofshadows
LetsNotMeet
Thetruthishere
Glitch
fatpeoplestories
TrueAskReddit
circlebroke
askgaybros
relationships

Out [19]:

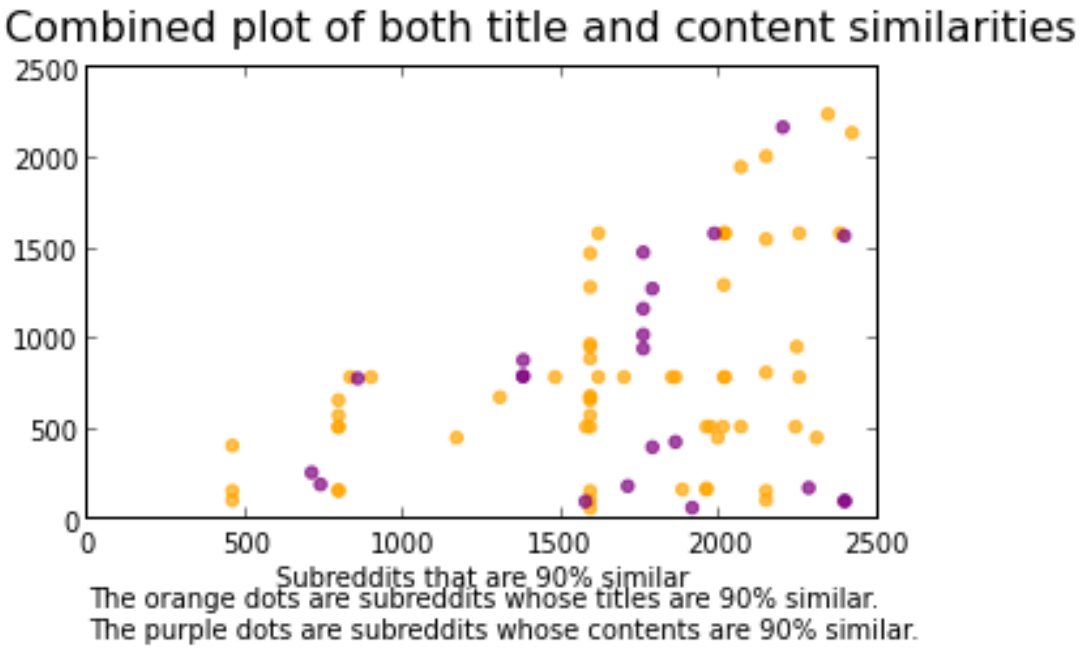
```
{'InsightfulQuestions': 'circlebroke',
 'MMFB': 'fatpeoplestories',
 'NoFap': 'confidence',
 'Showerthoughts': 'nosleep',
 'TheRedPill': 'confidence',
 'Thetruthishere': 'Ghosts',
 'TrueAskReddit': 'circlebroke',
 'circlebroke': 'casualiama',
 'fatpeoplestories': 'DoesAnybodyElse',
 'firstworldproblems': 'fatpeoplestories',
 'funny': 'fatpeoplestories',
 'libraryofshadows': 'Dreams',
 'nosleep': 'MMFB',
 'offmychest': 'nosleep',
 'pettyrevenge': 'fatpeoplestories',
 'rant': 'fatpeoplestories',
 'reactiongifs': 'fatpeoplestories',
 'relationship_advice': 'AskMen',
```

```
'seduction': 'confidence',
'sex': 'confidence',
'shittyaskreddit': 'circlebroke',
'short': 'confidence',
'shortscarystories': 'nosleep',
'shortstories': 'nosleep',
'socialskills': 'seduction',
'subredditoftheday': 'Showerthoughts',
'tifu': 'nosleep',
'unheardof': 'ThisIsOurMusic',
'videos': 'nosleep',
'weeklyreddit': 'subredditoftheday'}
```

The way the content similaries are computed is by grouping up all the words found in the title of posts within each subreddit, and count up the most frequent words. For any two given subreddits, we find the proportion of shared most-frequent-words.

If we are willing to assume the words found in the titles are meaningful, and related to the topic, content similarities may offer a more accurate pairing.

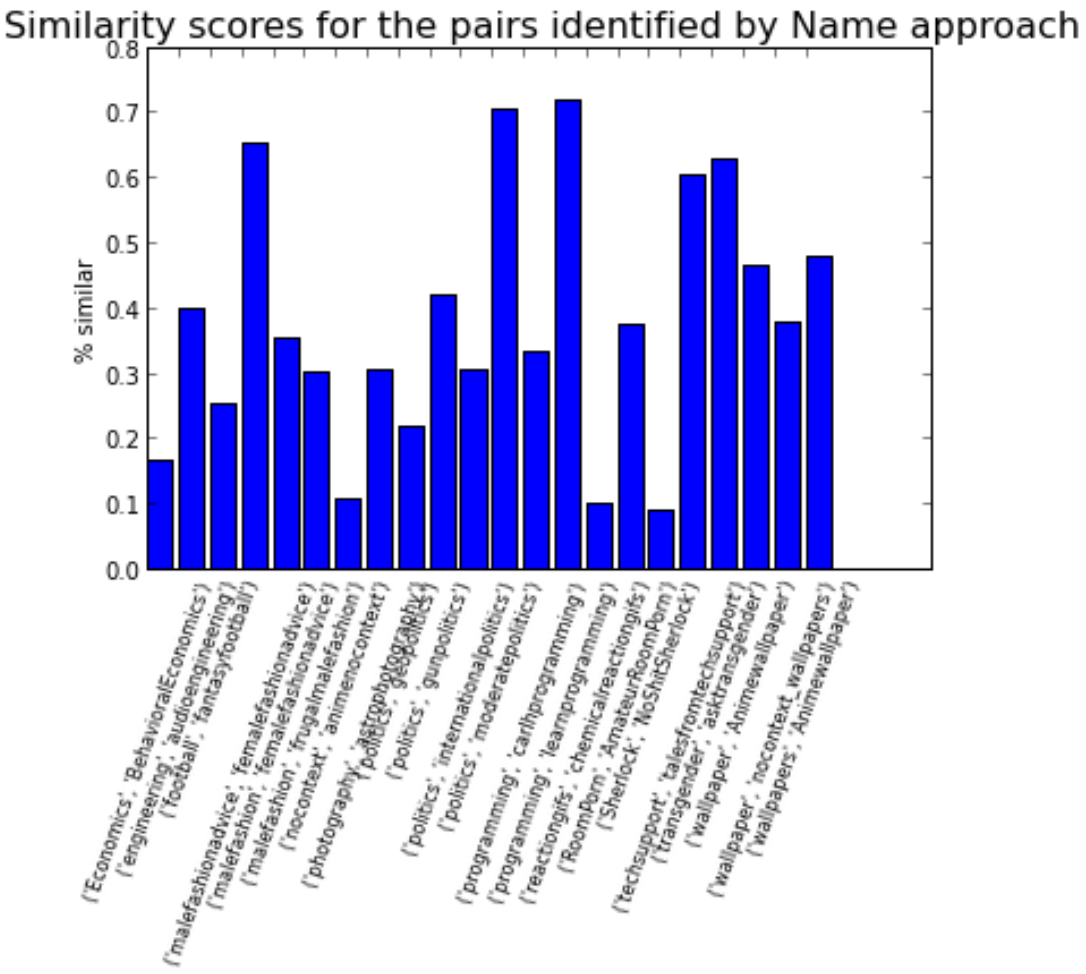
Moreover, the content similarity approach can reveal subreddits that are interesting but hard to find due to their obscure names as discussed in the article 19 Subreddits that have no reason to exist. By pairing those subreddits with others that are similar will increase their accesibility.The bad news is: subreddits that are identified to be more than 90% similar using the content approach have few overlaps with the pairs identified using the name approach.



This implies even though two subreddits may be 90% similar in their names, they do not necessarily have the same vocabulary in the titles of the posts.

Here is the content similarity scores for all the subreddit that are 90% similar in their names:

```
Out [22]: <matplotlib.text.Text at 0x3c897d0>
```



The highest pair has content similarity only up to 70% while some pairs are as low as 10%.It is hard to make a clear cut on when exactly are two subreddits similar. The name approach look for subreddit names, and round up those who have names that match. Not every pair of names matched may necessarily contain posts that talk about the same topic, furthermore, as you can see from the list above, even if the names are very similar, the vocabulary used in those subreddits are not that similar.

The content vocabuary approach identifies subreddit pairs that have similar words in titles, but there is no guarantee that the most frequent words that appear in the titles are related to the topic.

Despite the disagreements, we can still use the results from the name approach to reduce the number of subreddit that are devoted to the same topic, or to be used in a recommender system to offer viewers alternative subreddits to browse. The vocabulary approach can also be used to address the obscure subreddit name problem.

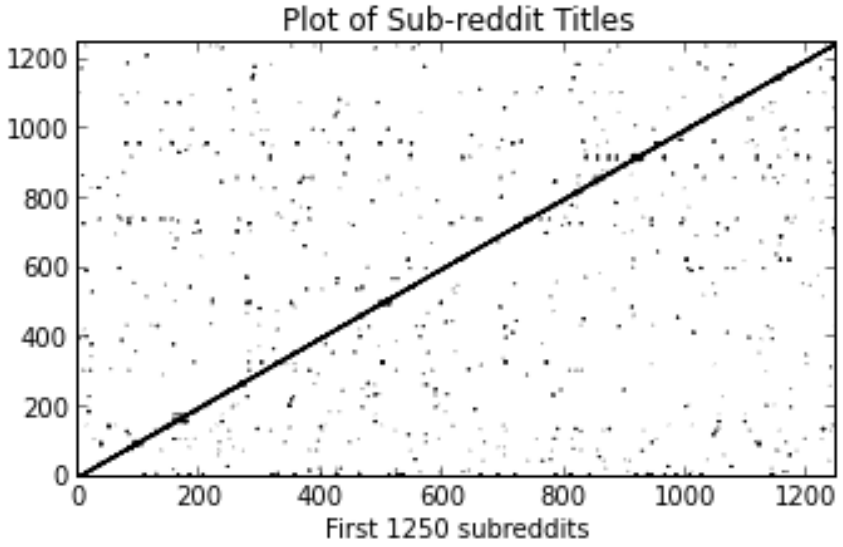
Full code and procedures can be found in Github .

Appendix

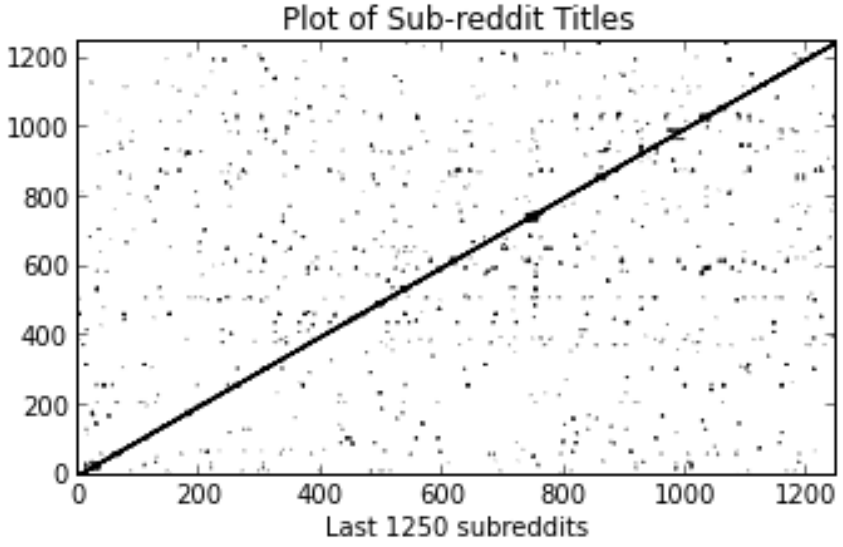
To view the over all pictures, we have to divide the space into four regions:

These are contour plots of the Name similarity matrix (mainly because they are faster to produce). Blank regions can be interpreted as thos subreddit pairs have low similarity score and they coincides with the blue regions in the heatmap above.

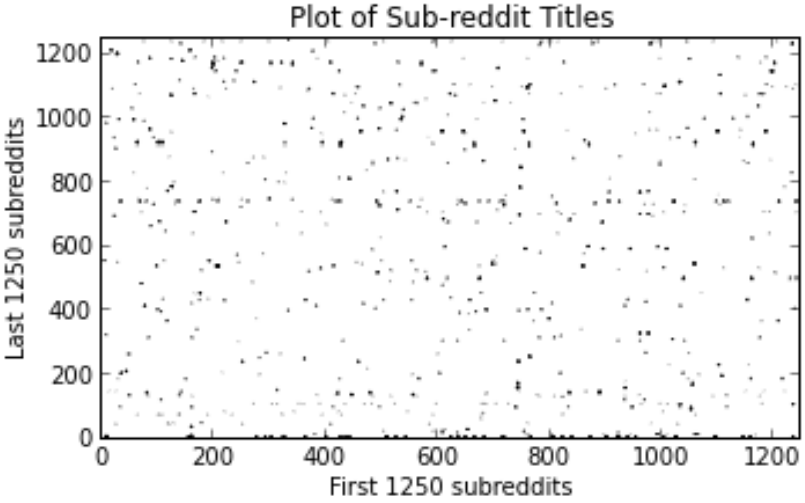
The following four graphs offers an overall view of the similarities of all the subreddit name pairs.



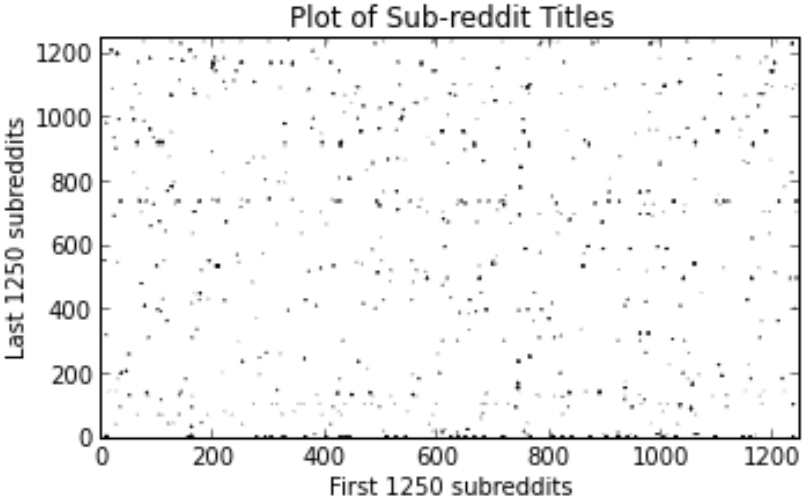
The first 1250 Names. Names that are similar appears in black.
Since the titles are sorted by alphabetical order, dots tend to cluster around the diagonal



The last 1250 Names. Names that are similar appears in black.
Since the titles are sorted by alphabetical order, dots tend to cluster around the diagonal



The first 1250 Names comopared to the last 1250 Names. Names that are similar appears in black.
Since we are away from the diagonal, we only see scatters with no apparent patches



The first 1250 Names comopared to the last 1250 Names. Names that are similar appears in black.
Since we are away from the diagonal, we only see scatters with no apparent patches