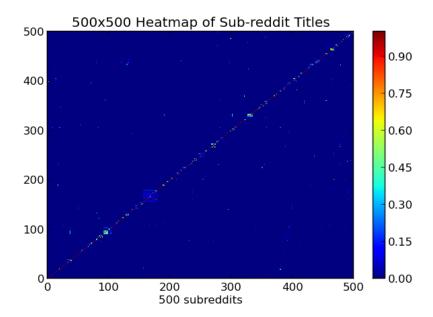
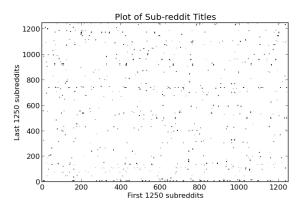
Figure 1



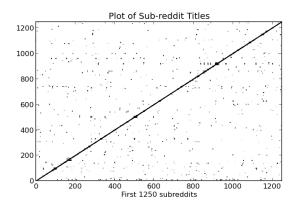
The heatmap of a portion of the  $2500 \times 2500$  Title similarity matrix (from 500-800). There are patches on the diagonal and sparse scatters everywhere else.

## Figure 2

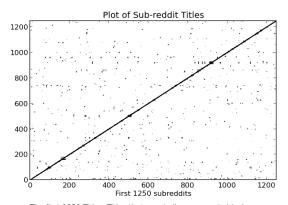


The first 1250 Titles comopared to the last 1250 Titles. Titles that are similar appears in black.

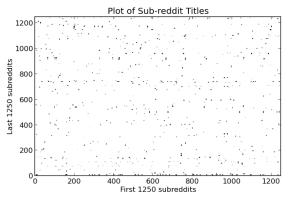
Since we are away from the diagonal, we only see scatters with no apparent patches



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



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

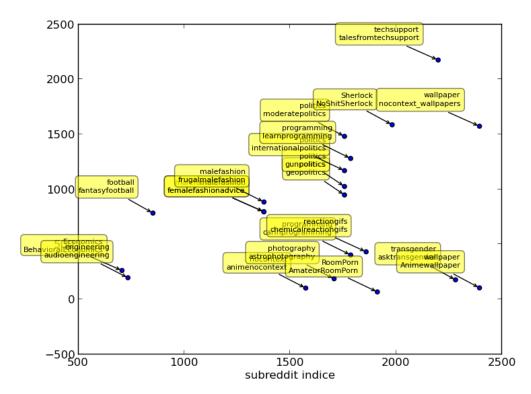


The first 1250 Titles comopared to the last 1250 Titles. Titles that are similar appears in black.

Since we are away from the diagonal, we only see scatters with no apparent patches.

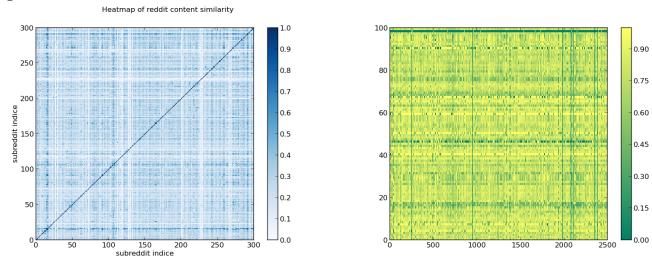
A full view of the title similarity matrix, divided into four regions.

Figure  ${\bf 3}$  Selected subreddits whose titles are over 90% similar



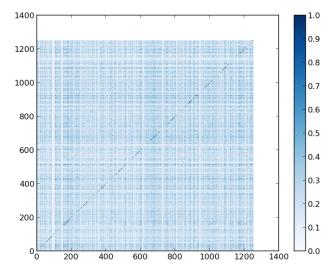
Subreddits that are more than 90 percent similar to each others, with corresponding names attached.

Figure 4



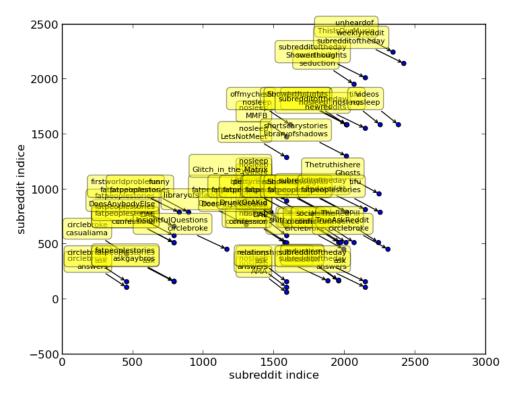
Left: a 300x300 region of the word-count similarity matrix. It has a large area of lightblue, only dark patches along the diagonal, implying most subreddits are distinct.

Right: a 100x1250 strip of the TFIDF-similarity matrix that aims to compare the similarity between any two given subreddits. It shows a high portion of yellow, indicating subreddits are quiet alike.



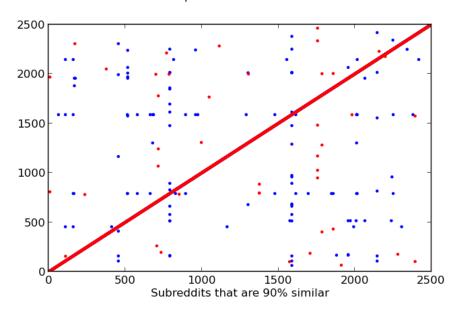
The same heatmap of content similarity computed base on word counts, at a large scale (1250x1250).

Figure  ${f 5}$  Selected subreddits whose Contents are over 90% similar



A plot of subreddits whose contents are more than 90 percent similar to each other. Note that several subreddits are similar to more than one other subreddits, therefore, appearing more than once.

Figure  ${f 6}$  Combined plot of both title and content similarities



The red dots are subreddits whose titles are 90% similar. The blue dots are subreddits whose contents are 90% similar.

Subreddits who are 90 percent alike in title and 90 percent alike in content plotted on top of each other.

Subreddits who are similar in title tend to be similar in content, but subreddits that are similar in content may not have the same title as there are more blue dots than red dots.