Categorization of Haiku from the Philadelphia Reflections website

George Fisher *

September 11, 2015

Abstract

The Philadelphia Reflections website contains several hundred haiku. This analysis shows how these haiku can be clustered (categorized) based on the apparent significance of the word usage in each haiku, compared to the overall word usage in the group of haiku as a whole.

See the **Technology** appendix on page 16 for the details of how this report was produced. The code is contained on GitHub: https://github.com/grfiv/haiku_analysis.

Contents

1 Introduction

This report shows the haiku organized into clusters of size 2 to 9 based on the significance of the word use in each haiku. At the top of each section is a graph showing the frequency distribution of the clusters (numbered starting at zero, by the way).

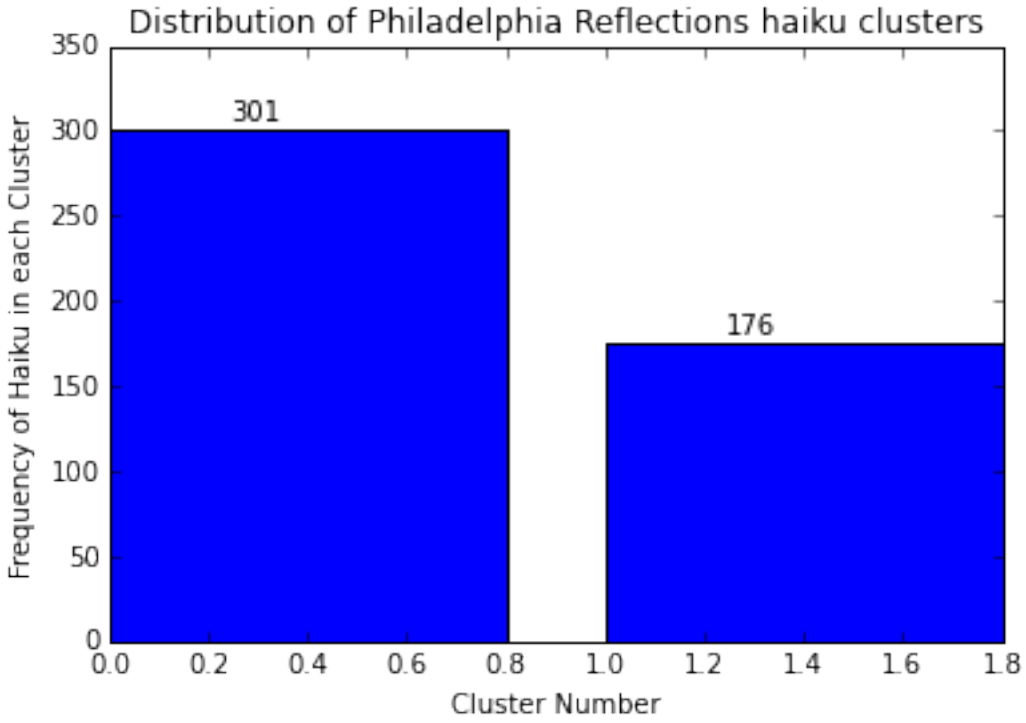
Underneath the graph is a list showing for each cluster the words that the algorithm considered most important for that group of haiku. Each word has a score and each cluster has a score. The numerical value of the score is not significant but its magnitude is: the larger the score, the more significant the word or cluster.

Below the list of clusters is a list showing the scores in descending order of each cluster.

The idea of the cluster's score is the strength of affinity within the cluster. In general, as the algorithm moves from n clusters to $n + 1$ clusters, it will attack the cluster in the previous set with the lowest score first since its members have the 'weakest attraction' to each other.

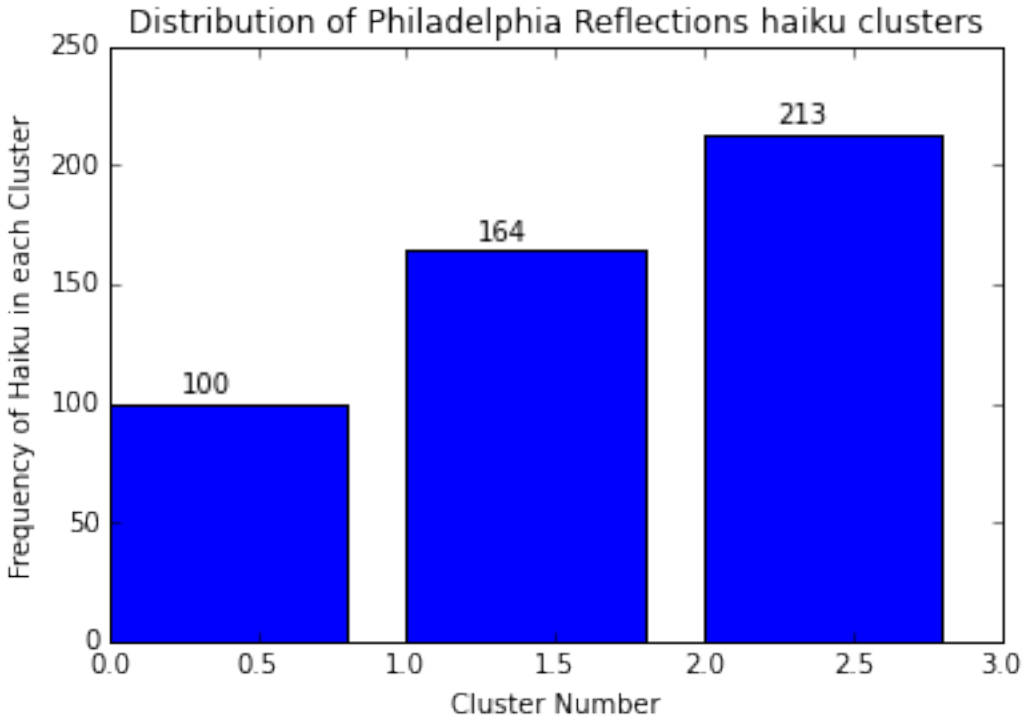
*George escaped from a 30-year international life of crime on Wall Street, got a Masters degree in quantitative finance from MIT, climbed Kilimanjaro and (some of) Everest and as of 2015 is pursuing an interest in machine learning. george at georgefisher dot com

2 Two Clusters



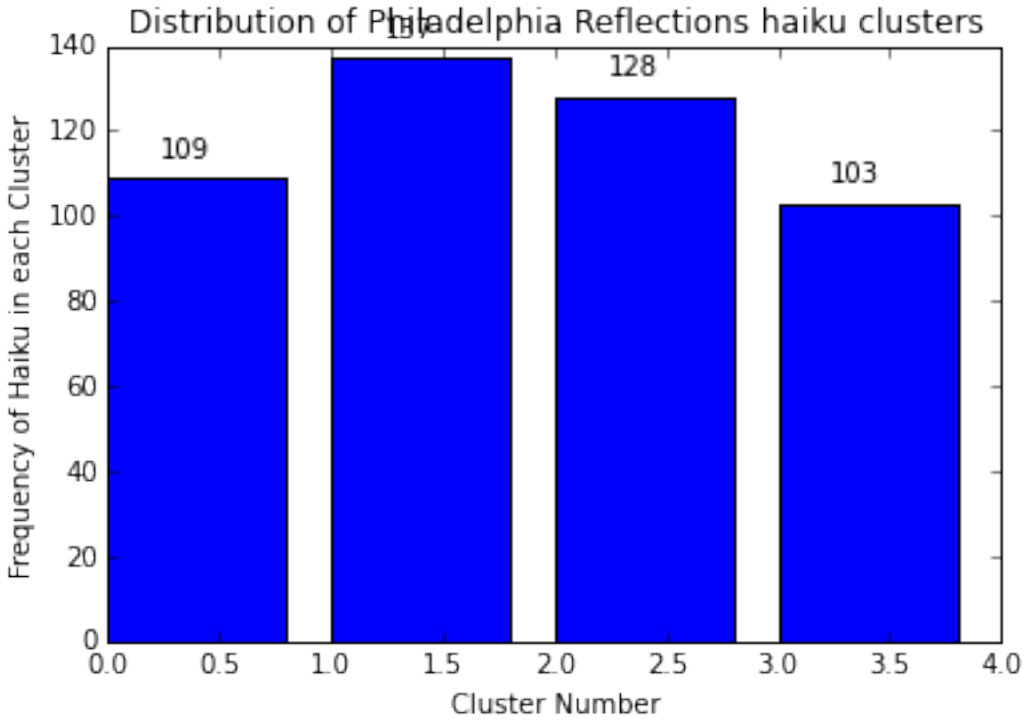

```
For cluster 0, 301 documents
1) old (score: 0.0209)
2) new (score: 0.0185)
3) eyes (score: 0.0154)
4) try (score: 0.0153)
5) quick (score: 0.0147)
Average cluster score 0.0169

For cluster 1, 176 documents
1) best (score: 0.0248)
2) man (score: 0.0225)
3) cash (score: 0.0203)
4) pay (score: 0.0202)
5) stop (score: 0.0202)
Average cluster score 0.0216

The clusters by average score, descending
1  0.0215889291333
0  0.0169423885829
```

3 Three Clusters




```
For cluster 0, 100 documents
1) buy (score: 0.0325)
2) loud (score: 0.0265)
3) crowd (score: 0.0251)
4) charge (score: 0.0220)
5) free (score: 0.0216)
Average cluster score 0.0256

For cluster 1, 164 documents
1) old (score: 0.0325)
2) new (score: 0.0256)
3) won (score: 0.0233)
4) age (score: 0.0213)
5) high (score: 0.0195)
Average cluster score 0.0245

For cluster 2, 213 documents
1) game (score: 0.0233)
2) ends (score: 0.0197)
3) play (score: 0.0193)
4) eyes (score: 0.0176)
5) words (score: 0.0175)
Average cluster score 0.0195

The clusters by average score, descending
0  0.025550027422
1  0.0244551417436
2  0.0194705330007
```

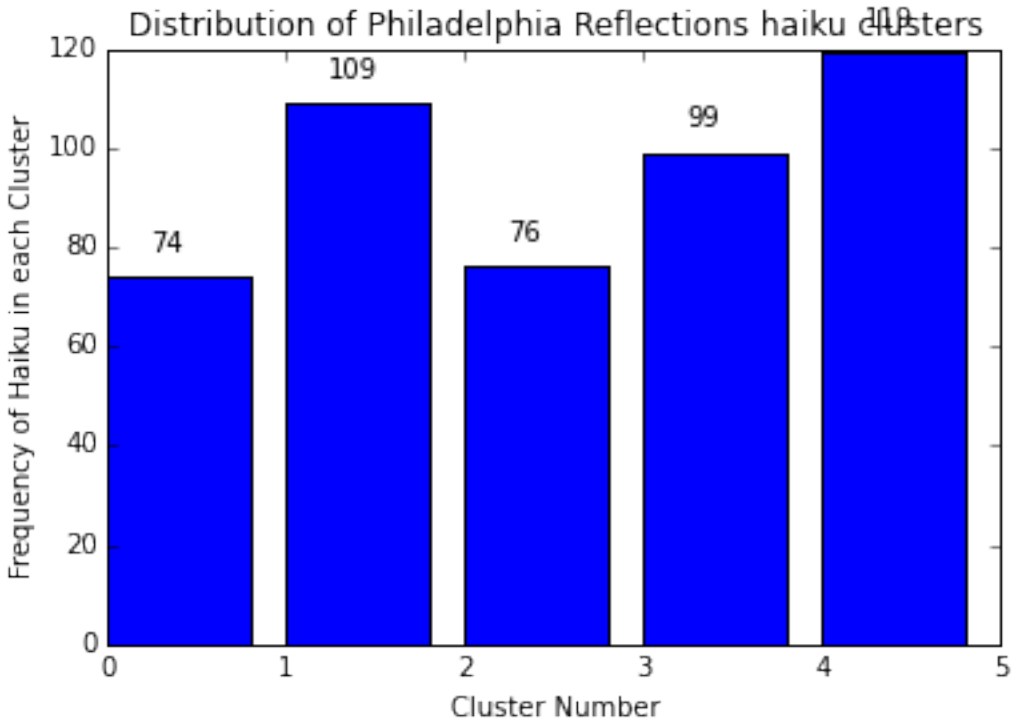
4 Four Clusters




```
For cluster 0, 109 documents
1) best (score: 0.0260)
2) love (score: 0.0216)
3) men (score: 0.0200)
4) new (score: 0.0191)
5) old (score: 0.0185)
Average cluster score 0.0210

For cluster 1, 137 documents
1) hope (score: 0.0216)
2) old (score: 0.0198)
3) feet (score: 0.0192)
4) feel (score: 0.0188)
5) watch (score: 0.0182)
Average cluster score 0.0195

For cluster 2, 128 documents
1) save (score: 0.0222)
2) gone (score: 0.0220)
3) cash (score: 0.0219)
4) stop (score: 0.0211)
5) words (score: 0.0208)
Average cluster score 0.0216

For cluster 3, 103 documents
1) end (score: 0.0267)
2) stress (score: 0.0257)
3) home (score: 0.0214)
4) just (score: 0.0200)
5) lights (score: 0.0195)
Average cluster score 0.0227
```

```
The clusters by average score, descending
3  0.0226744307088
2  0.0216042596299
0  0.0210393409326
1  0.0195132183468
```

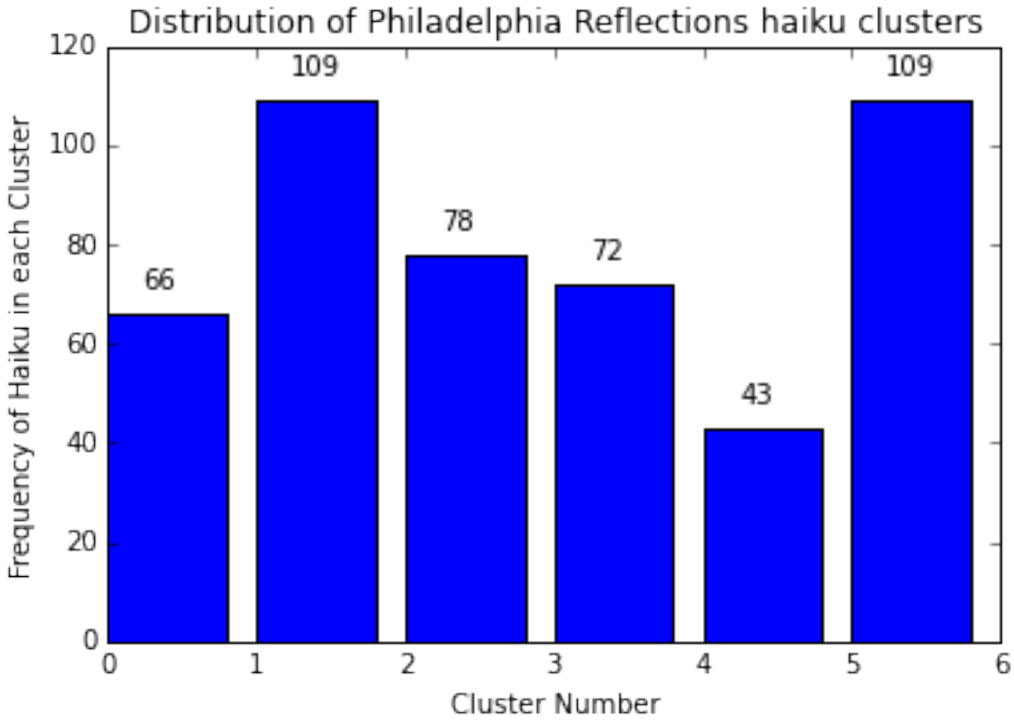
5 Five Clusters




```
For cluster 0, 74 documents
1) soon (score: 0.0270)
2) try (score: 0.0269)
3) won (score: 0.0259)
4) friends (score: 0.0238)
5) win (score: 0.0223)
Average cluster score 0.0252

For cluster 1, 109 documents
1) words (score: 0.0252)
2) wait (score: 0.0231)
3) cash (score: 0.0226)
4) free (score: 0.0223)
5) rich (score: 0.0221)
Average cluster score 0.0230

For cluster 2, 76 documents
1) head (score: 0.0492)
2) smile (score: 0.0305)
3) hide (score: 0.0274)
4) turn (score: 0.0240)
5) try (score: 0.0231)
Average cluster score 0.0308

For cluster 3, 99 documents
1) race (score: 0.0274)
2) time (score: 0.0261)
3) work (score: 0.0259)
4) don (score: 0.0237)
5) slow (score: 0.0229)
Average cluster score 0.0252
```

```
For cluster 4, 119 documents
1) old (score: 0.0337)
2) change (score: 0.0258)
3) years (score: 0.0257)
4) age (score: 0.0253)
5) new (score: 0.0235)
Average cluster score 0.0268

The clusters by average score, descending
2  0.0308363133164
4  0.0267930651694
3  0.0251993497109
0  0.0251725285107
1  0.0230334625804
```

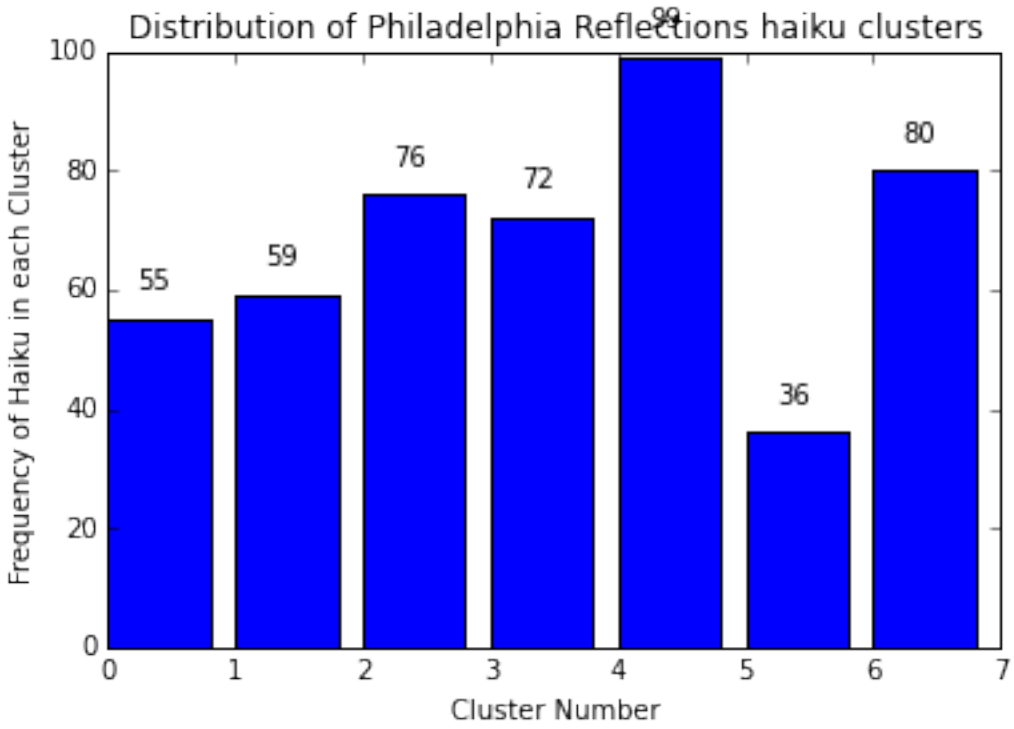
6 Six Clusters




```
For cluster 0, 66 documents
1) tough (score: 0.0262)
2) rage (score: 0.0252)
3) meat (score: 0.0244)
4) dead (score: 0.0222)
5) race (score: 0.0212)
Average cluster score 0.0238

For cluster 1, 109 documents
1) stop (score: 0.0224)
2) don (score: 0.0221)
3) wait (score: 0.0206)
4) buy (score: 0.0198)
5) cash (score: 0.0197)
Average cluster score 0.0209

For cluster 2, 78 documents
1) head (score: 0.0297)
2) smile (score: 0.0271)
3) cool (score: 0.0249)
4) rule (score: 0.0245)
5) old (score: 0.0236)
Average cluster score 0.0260

For cluster 3, 72 documents
1) feet (score: 0.0343)
2) win (score: 0.0302)
3) praise (score: 0.0291)
4) watch (score: 0.0260)
5) dance (score: 0.0241)
Average cluster score 0.0287
```

```
For cluster 4, 43 documents
1) waist (score: 0.0379)
2) work (score: 0.0347)
3) way (score: 0.0346)
4) pounds (score: 0.0327)
5) form (score: 0.0324)
Average cluster score 0.0345

For cluster 5, 109 documents
1) ends (score: 0.0294)
2) game (score: 0.0278)
3) check (score: 0.0264)
4) friends (score: 0.0239)
5) fun (score: 0.0226)
Average cluster score 0.0260

The clusters by average score, descending
4  0.0344558595641
3  0.028743253206
5  0.0260143525245
2  0.0259770275768
0  0.0238468214004
1  0.0209366814714
```

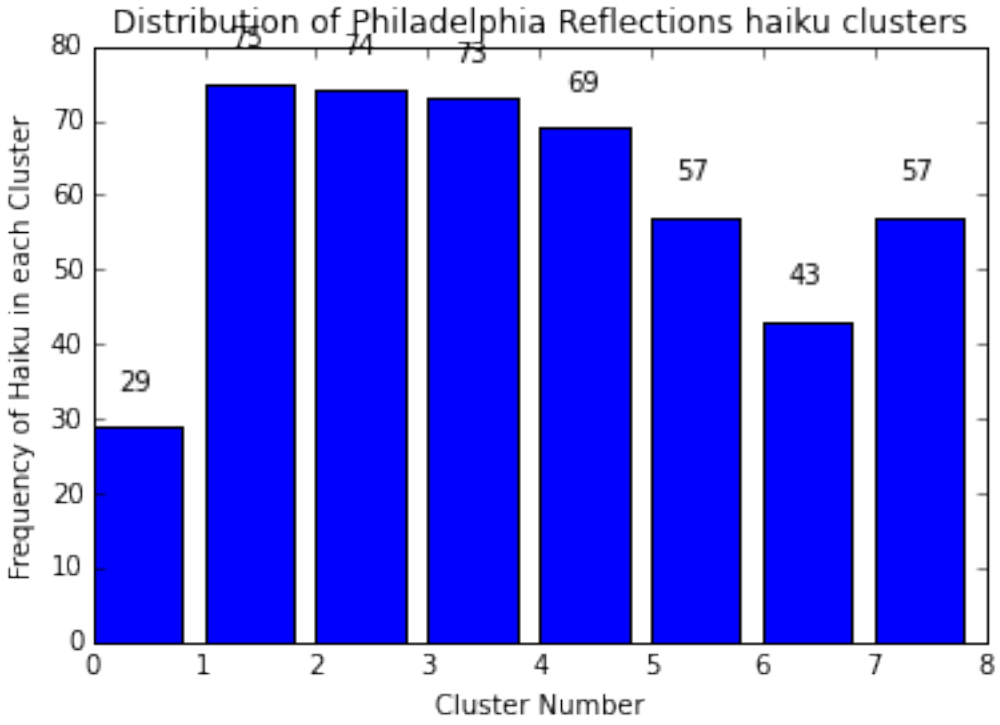
7 Seven Clusters




```
For cluster 0, 55 documents
1) talk (score: 0.0365)
2) don (score: 0.0316)
3) stop (score: 0.0285)
4) sweet (score: 0.0247)
5) hope (score: 0.0200)
Average cluster score 0.0283

For cluster 1, 59 documents
1) life (score: 0.0365)
2) youth (score: 0.0356)
3) years (score: 0.0351)
4) age (score: 0.0280)
5) face (score: 0.0248)
Average cluster score 0.0320

For cluster 2, 76 documents
1) try (score: 0.0277)
2) new (score: 0.0273)
3) use (score: 0.0247)
4) round (score: 0.0236)
5) fight (score: 0.0224)
Average cluster score 0.0251

For cluster 3, 72 documents
1) fans (score: 0.0301)
2) crowd (score: 0.0288)
3) loud (score: 0.0267)
4) big (score: 0.0262)
5) praise (score: 0.0261)
Average cluster score 0.0276
```

```
For cluster 4, 99 documents
1) game (score: 0.0282)
2) man (score: 0.0276)
3) best (score: 0.0270)
4) plan (score: 0.0256)
5) choice (score: 0.0246)
Average cluster score 0.0266

For cluster 5, 36 documents
1) dance (score: 0.0436)
2) feet (score: 0.0432)
3) shoes (score: 0.0326)
4) step (score: 0.0319)
5) fear (score: 0.0284)
Average cluster score 0.0359

For cluster 6, 80 documents
1) soon (score: 0.0340)
2) ends (score: 0.0296)
3) friends (score: 0.0281)
4) fun (score: 0.0217)
5) gals (score: 0.0209)
Average cluster score 0.0269

The clusters by average score, descending
5  0.0359255113684
1  0.0320015325584
0  0.0282661396908
3  0.0275813845331
6  0.0268625114736
4  0.0266101572609
2  0.0251354413953
```

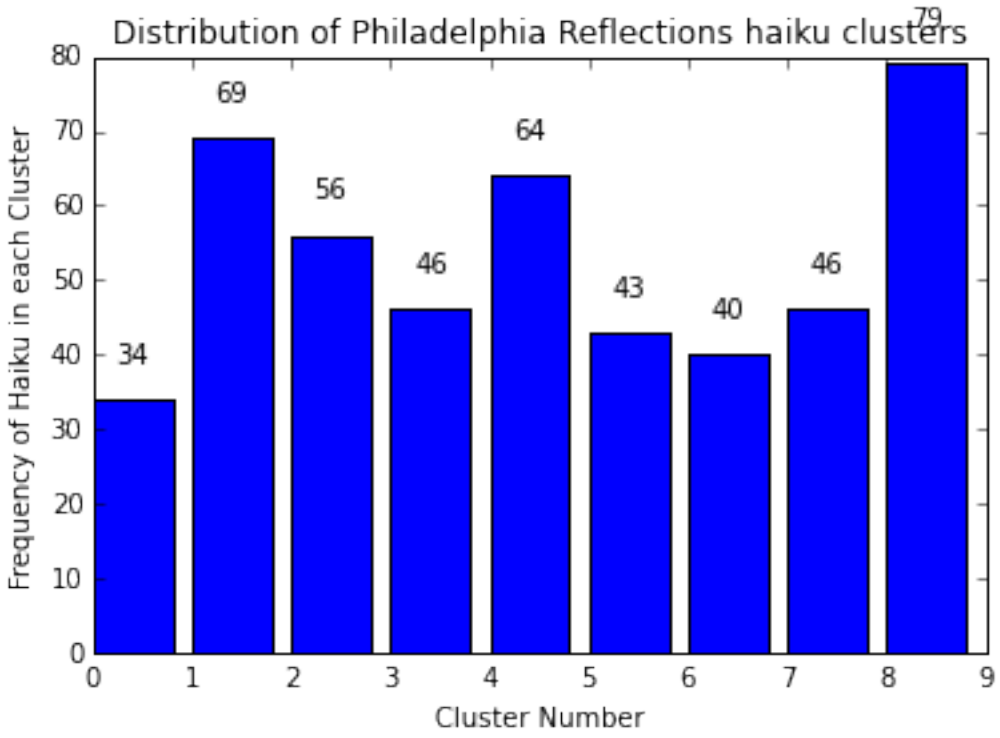
8 Eight Clusters




```
For cluster 0, 29 documents
1) tied (score: 0.0503)
2) fast (score: 0.0385)
3) tongue (score: 0.0362)
4) pride (score: 0.0324)
5) check (score: 0.0321)
Average cluster score 0.0379

For cluster 1, 75 documents
1) slow (score: 0.0298)
2) place (score: 0.0278)
3) right (score: 0.0277)
4) don (score: 0.0276)
5) race (score: 0.0274)
Average cluster score 0.0281

For cluster 2, 74 documents
1) head (score: 0.0336)
2) game (score: 0.0305)
3) rule (score: 0.0244)
4) seek (score: 0.0239)
5) fame (score: 0.0238)
Average cluster score 0.0272

For cluster 3, 73 documents
1) eyes (score: 0.0330)
2) feel (score: 0.0312)
3) beat (score: 0.0281)
4) feet (score: 0.0270)
5) age (score: 0.0248)
```

```
Average cluster score 0.0288

For cluster 4, 69 documents
1) stop (score: 0.0425)
2) rules (score: 0.0326)
3) ends (score: 0.0308)
4) pop (score: 0.0250)
5) gone (score: 0.0247)
Average cluster score 0.0311

For cluster 5, 57 documents
1) cash (score: 0.0360)
2) soon (score: 0.0298)
3) rich (score: 0.0292)
4) old (score: 0.0283)
5) ways (score: 0.0271)
Average cluster score 0.0301

For cluster 6, 43 documents
1) food (score: 0.0426)
2) good (score: 0.0304)
3) meals (score: 0.0299)
4) mood (score: 0.0272)
5) ho (score: 0.0270)
Average cluster score 0.0314

For cluster 7, 57 documents
1) wait (score: 0.0345)
2) cow (score: 0.0313)
3) high (score: 0.0279)
4) change (score: 0.0271)
5) bull (score: 0.0267)
Average cluster score 0.0295

The clusters by average score, descending
0  0.0378760600375
6  0.031409647666
4  0.0311147227688
5  0.0300991632311
7  0.0295004152491
3  0.0288251236892
1  0.0280709273252
2  0.0272439463521
```

9 Nine Clusters




```
For cluster 0, 34 documents
1) cow (score: 0.0310)
2) red (score: 0.0294)
3) gone (score: 0.0285)
4) bluff (score: 0.0284)
5) cards (score: 0.0277)
Average cluster score 0.0290

For cluster 1, 69 documents
1) feet (score: 0.0312)
2) dance (score: 0.0276)
3) words (score: 0.0272)
4) beat (score: 0.0260)
5) stop (score: 0.0259)
Average cluster score 0.0276

For cluster 2, 56 documents
1) food (score: 0.0355)
2) fast (score: 0.0340)
3) quick (score: 0.0302)
4) grown (score: 0.0277)
5) time (score: 0.0272)
Average cluster score 0.0309

For cluster 3, 46 documents
1) head (score: 0.0773)
2) style (score: 0.0360)
3) hide (score: 0.0348)
4) smile (score: 0.0323)
5) hair (score: 0.0316)
```

```
Average cluster score 0.0424

For cluster 4, 64 documents
1) time (score: 0.0346)
2) race (score: 0.0288)
3) work (score: 0.0284)
4) push (score: 0.0283)
5) place (score: 0.0251)
Average cluster score 0.0290

For cluster 5, 43 documents
1) wait (score: 0.0366)
2) poor (score: 0.0366)
3) rich (score: 0.0346)
4) fate (score: 0.0316)
5) gold (score: 0.0277)
Average cluster score 0.0334

For cluster 6, 40 documents
1) lights (score: 0.0331)
2) turn (score: 0.0312)
3) wheels (score: 0.0282)
4) home (score: 0.0277)
5) stay (score: 0.0277)
Average cluster score 0.0296

For cluster 7, 46 documents
1) free (score: 0.0376)
2) man (score: 0.0370)
3) know (score: 0.0357)
4) world (score: 0.0320)
5) run (score: 0.0273)
Average cluster score 0.0339

For cluster 8, 79 documents
1) good (score: 0.0347)
2) try (score: 0.0283)
3) hope (score: 0.0262)
4) best (score: 0.0250)
5) seek (score: 0.0249)
Average cluster score 0.0278

The clusters by average score, descending
3  0.0424074312387
7  0.0339217775439
5  0.0334125543697
2  0.0309348279601
6  0.0295708045007
4  0.0290280869937
0  0.0290063170299
8  0.0278211023041
1  0.0275629793656
```

A Technology

The haiku are kept as 'blogs' in the Philadelphia Reflections website, collected into groups known as 'topics'.

1. The website's MySQL database contents was downloaded in SQL format from the Apache server to a SQLite3 database on an Ubuntu 15.04 Zareason laptop.
2. Using Python 2.7, in an IPython 4.0 (Jupyter) Notebook, a list of the blog ids was extracted, given a list of the topic ids.
3. Each haiku and its title was extracted from the database, cleaned and saved as a text file in a folder containing only these files:
   (a) The `BeautifulSoup` package was used to remove the HTML tags
   (b) English 'stop words' were removed (these are words, like articles and pronouns, that are required for grammatical English but usually add no specific meaning)
   (c) Escape characters were removed.
   (d) A file name was constructed from the concatenation of the title and the blog id
   (e) The text was converted to lower case and strings of blanks were compressed to one
4. A list (known as the 'corpus') was created from these files, each entry of which a single string of all the words in a single haiku
5. In a loop, testing clusters in size from 2 to 20
   (a) The entire corpus was run through `scikit-learn`'s `TfidfVectorizer` to identify the significant words in each haiku
   (b) Clusters were created using the `KMeans` algorithm
   (c) The clusters were scored using the `silhouette_samples, silhouette_score` package using cosine similarity as the metric
6. Finally, the graphs and word clusters shown above were produced
7. This report was produced with the TeXstudio editor (available on Windows, Mac and Unix), using the LaTeX markup language