Clustering Analysis on Selected Texts Using NLP

- by Mark Ferguson

Unsupervised Learning Capstone Project- September 2018

The aim of this project is to group a disparate collection of short texts using unsupervised learning techniques. To facilitate this, the texts will first be parsed using Natural Language Processing, commonly referred to by its abbreviation NLP. NLP is a field within data science that interprets written or spoken data by characterizing constituent words or phrases. This powerful and increasingly widely used technique will be used to define each individual text using a technique known as vectorization, which will be explained in detail later. The analytical choices and decisions made will be discussed along the way. More formally, the goals of this project can be summarized as follows:

- To use several different clustering methods to group the texts.
- To evaluate and validate, using other techniques, the quality of the clustering obtained.

The Corpus

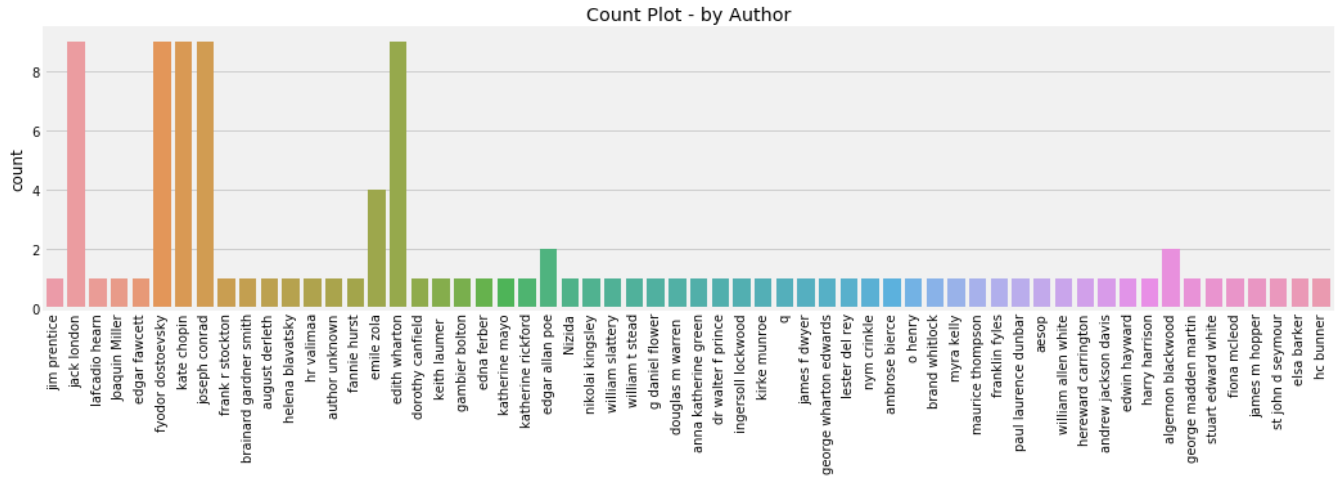
The corpus to be studied consists of 105 short stories downloaded from various sources on the internet. Sources include the Gutenberg Project and textfiles.com. Authors range from Jack London to Emile Zola, to Fyodor Dostoevsky, to a number of lesser-known authors. The texts are drawn from the romance, science fiction, detective and occult genres. Texts that cannot be easily identified as being from any of those genres are classified as 'miscellaneous'. In the analysis, each text will be identified by its author and title, and categorized according to its genre and the collection, if applicable, from which it was taken. Publication dates span the period from 1880 to 1992.

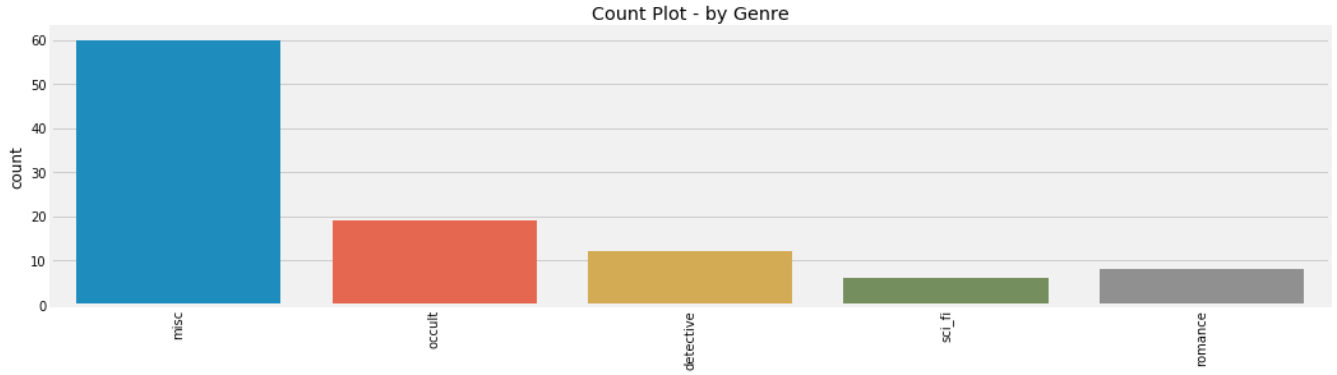
The full listing of texts, absent any grouping or order, is as follows:

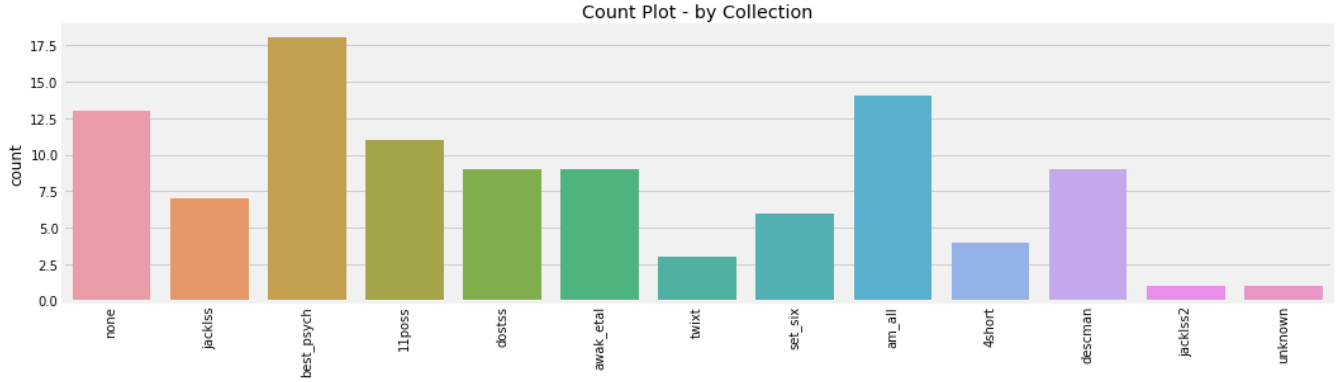
100 west by 53 north	bitter sweet	once upon a mattress	the end of all	the pit and the pendulum
a day's lodging	bobok	painkiller	the eyes of the panther	the portal of the unknown
a ghost	brown wolf	photographing invisible beings	the gift of the magi	the problem with time travel
a lion and a lioness	captain burle	prints of the city	the gold brick	the quicksand
a lost day	desiree's baby	rain dance mechwarrior fan fiction	the heavenly christmas	the reckoning
a novel in nine letters	expiation	shall he marry her	the informer	the repairman
a pair of silk stockings	flint and fire	some remarkable psychic experiences of famous ...	the kiss	the return
a reflection	freya of the seven isles	strange adventures of a million dollars	the lady's maid's bell	the right promethean fire
a respectable woman	gambler's world	the awakening	the land of heart's desire	the riverman
a smile of fortune	gaspar ruiz	the brute	the letter	the second generation
a thing that glistened	ghosts in solid form	the bushwhacker's gratitude	the locket	the secret sharer
a tragedy of high explosives	his mother's son	the cheated juliet	the miller's daughter	the sin eater
a traveler in time	il conde	the citizen	the mission of jane	the story of keesh
a witchs den	israel drake	the clavecin bruges	the mystic krewe	the struggles and triumph of isidro de los mae...
accounting for the cards	joseph a story	the crocodile	the only girl at overlook	the sun dog trail
after dark	ligeia	the death of olivier becaille	the ordeal at mt hope	the supernormal experiences
an anarchist	love of life	the descent of man	the other two	the sylph and the father
an honest thief	ma'ame pelagie	the dilettante	the parable of the vain crow	the tenor
an unpleasant predicament	nana	the dream of a ridiculous man	the passing of priscilla winthorp	the unexpected
another man's wife	nature spirits or elementals	the duel	the peasant marey	the white man's way
beyond the bayou	negore, the coward	the dwindling years	the phantom armies seen in france	when the world was young

Visual Summary of the Data

These plots give a quick visual overview of the dataset:

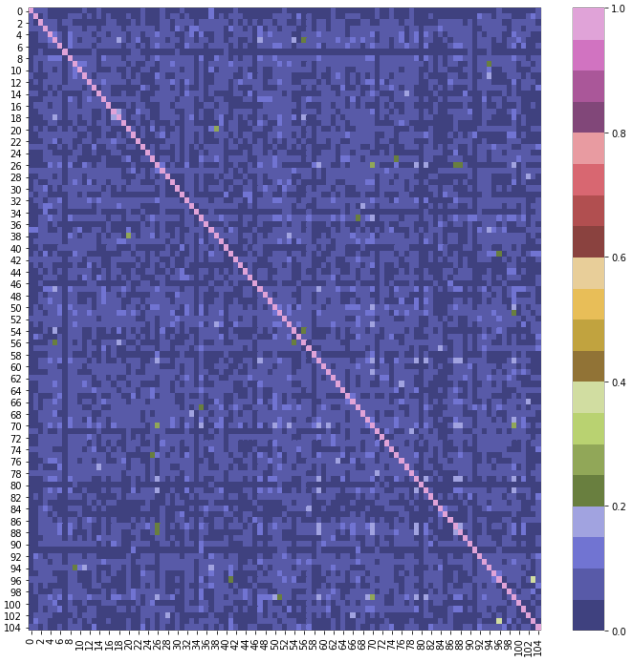






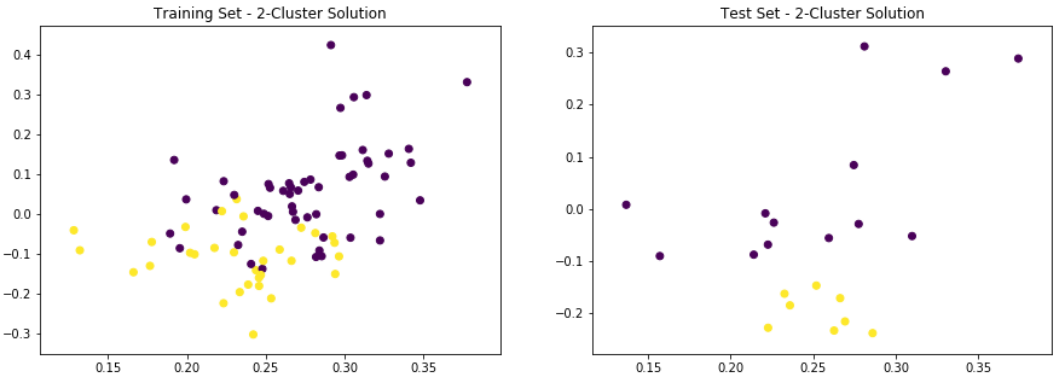
Methodology

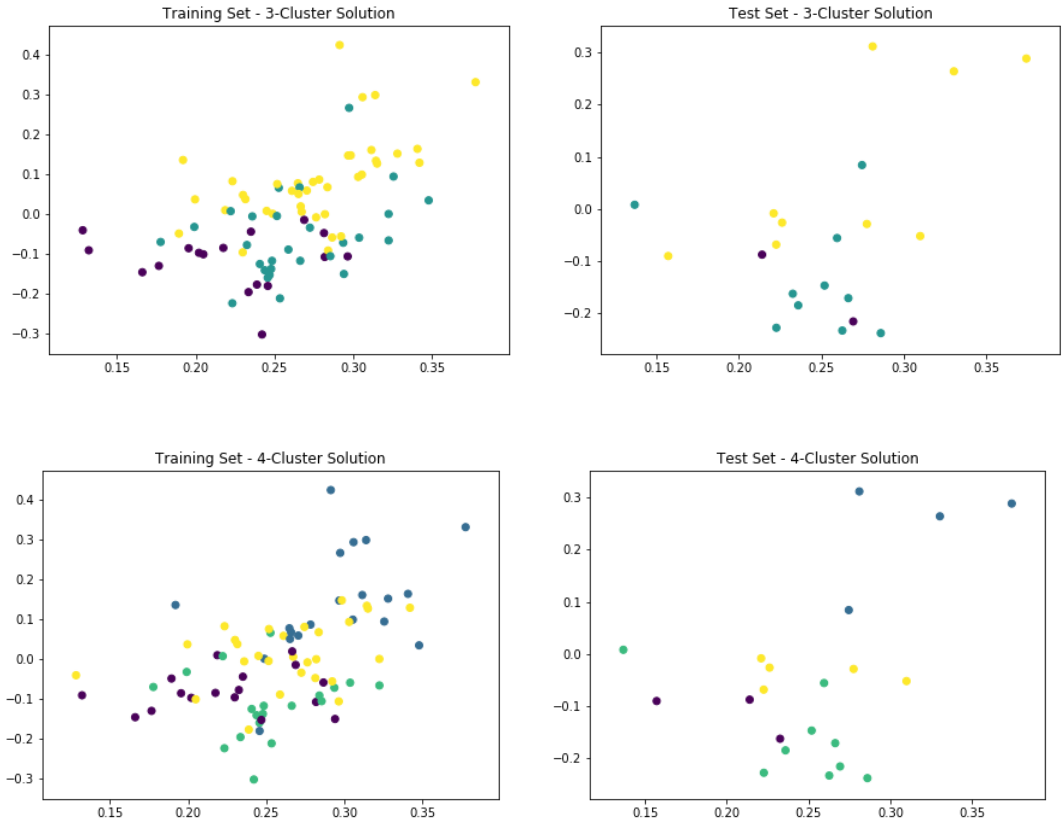
To conserve available computing power, the texts were first limited to their first 6000 characters, which is roughly equivalent to two pages. The texts were then vectorized using the TF-IDF model, such that each vector represented one text. The cosine similarity was then calculated as a measure of the similarity of each text to every other. These similarities are shown on the heatmap below. Clearly, most of the texts have a very low similarity to one another. A few pairs can be seen, though, whose similarity is higher than the rest. These are in greenish hues:



Clustering Techniques - 'Ground Truth' Methods

First, 'ground truth' methods were used to classify the texts by author, genre and then by collection. These attempts were fairly unsuccessful, as shown by their low Adjusted Rand Index scores – the K-Means model yielded scored of 0% for genre, 10% for collection, and 21% for author. The collection score may be partly due to the fact that some collections contained stories written exclusively by one author (e.g. A Set of Six, a collection by Joseph Conrad). Spectral Clustering produced a similarly unimpressive ARI of 0%.
Next, further runs were made with the K-Means model, with 2, 3 and 4 clusters, but without trying to tie these to any known groupings within the data. The data were first separated into training and test sets containing 84 and 21 texts respectively. Then, the vectorized training dataset was reduced from its original size of over 5000 dimensions, to 200. The K-Means model was run on both the training and test datasets, and the first two components of this reduced feature set plotted on x-y axes. The cluster was used as a third dimension in the form of dot color, as can be seen below. Clearly the quality of the solution decreases with increasing number of clusters. Consistency between training and test datasets can be inferred from the similarity between their respective scatter plots:





‘No Ground Truth’ Methods

The Mean Shift method was unable to establish any boundaries from the data, estimating a single cluster no matter what input arguments were passed.
We thus moved on to Affinity Propagation, and it was this part of the analysis that proved most successful. When applied to the training set, this technique discerned 14 clusters, ranging in size from 2 to 14 texts. The silhouette score, an evaluation technique used on cluster configurations having no ground truth, was 0.0072. This indicates moderate cluster quality.

Some text-text pairs were earlier designated as high-correlation or low-correlation, if their cosine similarity scores were more than 1.5 standard deviations above or below the mean. In 26.3% of the high-correlation pairs, both texts were within the same cluster, whereas this was true in only 2.5% of the low-correlation pairs. This provides a good substantiation of Affinity Propagation as a method for clustering these texts.

Supervised Learning

The next step was to use Logistic Regression to assign each of the texts in the test dataset to an existing cluster. The model was run without penalty, using a C coefficient of 1×10^9, and each test text classified as belonging to a cluster. The individual mean cluster coefficients were then calculated for each cluster, and compared with the values before the addition of the test data. While the overall silhouette score decreased slightly, from 0.0072 to 0.0068, 8 of the 14 clusters increased their silhouette scores, a further validation of the clustering solution. These results can be seen below:

	cluster_label	silhouette_score_orig	silhouette_score_final	delta	improvement?
0	0	0.003692	0.006037	0.002345	Y
1	1	0.008674	0.008283	-0.000392	N
2	2	-0.002922	-0.000556	0.002366	Y
3	3	0.019791	0.020268	0.000477	Y
4	4	0.006453	0.004814	-0.001639	N
5	5	0.002733	0.005331	0.002598	Y
6	6	0.004642	0.002754	-0.001888	N
7	7	-0.003492	-0.004038	-0.000546	N
8	8	0.012407	0.014682	0.002275	Y
9	9	0.104248	0.104248	0.000000	N
10	10	0.001898	-0.000735	-0.002632	N
11	11	0.013946	0.014089	0.000142	Y
12	12	0.019198	0.019918	0.000721	Y
13	13	0.006646	0.006729	0.000083	Y

Summary

We can summarize the findings from this study as follows:

- It is possible to model and group a corpus of short stories using TF-IDF and the Affinity Propagation clustering model.
- This combination of techniques on the training set produced a moderate solution with some overlap between clusters, having an aggregate silhouette score close to zero.
- Using logistic regression to cluster the holdout data into the existing groups, proved successful, with a negligible reduction in the overall silhouette score. In fact, 8 of the 14 clusters experienced an improvement in their mean silhouette coefficient after the holdout data had been added.
- Ground Truth clustering models were not effective in clustering the data according to genre, collection or author.
- However, some success was found using a K-Means solution specifying 2 and to some extent 3 clusters, as evidenced by plots of the first two SVD-reduced components of these data on x-y axes. These solutions showed reasonable consistency between the training and test sets.

Limitations

The main limitations of this study were:

- Only the first 6000 characters of each text were used. Better solutions, or further insights, may have been obtained if a larger portion of each text had been considered. Also, it would have been nice to have used a larger corpus of texts. Conversely, it could be that using more texts, or a larger portion of each text does not produce a discernible gain in the quality of the results. That in itself would have been a noteworthy finding; but the level of detail in the analysis had to be tailored to the available time and CPU power.

- There was a slight lack of familiarity with the texts used. This may have led to some inaccuracies in classifying texts as 'misc', for example, where some may perhaps have fit better into another category. A better classification might possibly have led to a better genre-specific clustering outcome.