THE USE OF UNSUPERVISED LEARNING TO IMPROVE THE STYLISTIC GENRE CLASSIFICATION FROM MIXED CORPUS

Author: Colin Power
Supervisor: Dr. Alexander O'Connor

A thesis submitted to the
University of Dublin, Trinity College
for the degree of
M.A.I. (St.)

Colin Power
School of Computer Science and Statistics
Trinity College, Dublin,
powerc6@tcd.ie

2014

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other University, and that, unless otherwise stated, it is entirely my own work.

Colin Power

07/04/14

Permission to lend or copy

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Colin Power

07/04/14

Acknowledgements

I would like to thank my supervisor Dr. Alexander O'Connor for his unwavering support and confidence in my work. I would like to also thank my parents for their support throughout the year.

Abstract

The purpose of this research is to determine the degree to which we can use unsupervised or deep learning as a method of improvement on the classification of stylistic genres from a a body of mixed corpus. Unsupervised approaches are considered over more traditional supervised methods due to the advantages of formation of more appropriate genres, the use of unlabelled data for the training set and less initial work is needed.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Introduction

In this chapter, the motivation and reasoning for the research being carried out are explored. The research question is stated and examined in detail, the objectives are introduced and potential research challenges that may present themselves are laid out. Finally an outline of the document structure is provided.

1.2 Motivation

The rise and fall of human civilization over time can be viewed as the rise and fall of sum of accessible human knowledge. In the later half of the 20th century this sum of human knowledge has expanded at an unprecedented rate and as we move into the second decade of the 21st century this trend shows no sign of slowing. As a consequence the vast amount data which holds this knowledge is starting to become truly unmanageable.

Another recent trend is the upsurge in the creation of *user generated content* online in all formats from text to video to audio. By far the largest part of interactive human media today is still text based and as such text classification will be the focus of this work. Text classification by natural language processing has a history stretching back several decades but only truly became prevalent with the advent of the world wide web in the early nineties.

The point of text classification is to sort similar texts in to a class or genre to aid people who are searching for a type of text. The old information retrieval question of, how do find something if you don't know what you are looking for, can be solved if all the information is sorted into relevant labelled categories or genres.

The need for the ability to classify text documents on a reasonably large scale is clearly present and with such a diverse global audience an unsupervised approach with no preconceived notions seems the most appropriate.

1.3 Research Question

The main research question this document hopes to answer is: Can the unsupervised learning approach offer improvements to the area of genre classification? The question can be broken down into three parts, the unsupervised learning approach, the improvements it can offer and finally how they can be utilised in the area of genre classification.

The unsupervised approach is the alternative to the traditional supervised approach to machine learning. In unsupervised the data is not pre-processed by humans and then fed to system. Instead the raw data is fed in and the system extracts patterns out of the data. These patterns differ from those used in supervised learning as they are not preconceived and are a result of the system and the data. In supervised learning the patterns are designed before the data is entered into the system. The system will instead be instructed whether an individual piece of data fits into given pattern and then adjust itself so that similar pieces in future will be placed correctly. The main differences in the two approaches can be summarized as this. Supervised approaches will alter the system for new data while unsupervised approaches will alter the categories when new data is received.

The improvements of the unsupervised approach revolve around the fact that the categories into which data is sorted are determined by the data rather than by expert human knowledge. These categories or genres as they will be referred to for the purpose of this document, are better fitted to the data than pre-made genres. Due to this better fit the texts in a given genre tend to be more closely related even though these relations may not be easily apparent. This is clearly a major advantage in the aim of improving the classification of documents into genres.

To be clear in how the area of genre classification can be improved the aims of the area must be laid out. The aim is to correctly group similar texts without leaving out any that should be in a grouping while also not including any which should be in a different grouping. Machine learning is highly useful for this style of grouping as it is based on comparing the similarity of new information with an established baseline model. By using unsupervised learning to create the model we can have a model that is designed to fit the data rather than having to fit the data to a pre established model.

Hopefully the research question of: Can the unsupervised learning approach offer improvements to the area of genre classification, will be answered in this document. The aim is to explore the question and the various challenges involved in meeting the objectives.

1.3.1 Objectives

To address the research question stated above, four specific research objectives have been defined to ensure the completeness of the answer:

1. To conduct a survey of the existing methods of genre classification, their effectiveness and an overview of the underlying machine learning and natural language processing theory.

2. To select a wide corpus which has a large variety of reasonably sized texts which cover various subject areas and has recognition in the research community for quality.

3. To design and build a system which is capable of using unsupervised learning to categorise texts using select features.

4. To conduct an evaluation of the approach taken, the methods used and the resultant classification of texts.

1.3.2 Challenges

In addition to the objectives stated above there are a number of domain specific problems which must be addressed for a successful conclusion to the research.

1. The concept of genre is ambiguous both as in what a genre should define and whether or not a text should belong to given genre. Many texts for example have the possibility to belong to several different genres, for example should a news article on a scientific breakthrough be in news or in science?

2. The corpus selection process is difficult as most public bodies of text are collected from a single source and tend to be narrow in scope which is unhelpful in identifying a number of genres. The few corpus that have a wide variety of topical texts tend to be smaller and naturally have less texts in a given topic than a more specialised corpus. This lack of depth needs to be balanced

3. The features that can be extracted from a text are many and varied however many features rely on already possessing some previous knowledge regarding the new data. The unsupervised learning approach restricts some the features which could be considered genre specific.

4. The evaluation of a system which categorizes texts based on similarity rather than on trying to fit them to a given set of genres is difficult as it must be done manually which due to the volume is problematic.

1.4 Thesis outline

A survey and analysis of the relevant theory and related work is given in Chapter Two along with an explanation of how this work has influenced the decisions taken while preforming the research. Chapter Three goes over the various design decisions made for the project, while Chapter Four covers the implementation of these decisions. Chapter Five contains a discussion of the evaluation of the research together with examples of the data. In Chapter Six the conclusions of the research are discussed along with the possible extensions and future work.

Chapter 2: State of the art

In this section the background and current state of the art in the areas of *machine learning* and *genre classification* will be discussed and in particular the areas where they overlap, which is the scope of this project. The background of *genre classification* which is *natural language processing* will be introduced. Next the field of machine learning will be introduced and broken down into two separate portions, supervised and unsupervised learning. Methods of which pertain to the topic under question here, that of *classification*, will be provided and the various benefits and problems of each will be examined in context of our research question.

2.1 Genre analysis

Genre classification for the purposes of this text will refer to the sorting of uncategorised text documents by genre and the grouping of similar texts. These genres will vary depending on corpus used but an example would be fictional, editorial, reference, technical, chat-room messaging or miscellaneous.

2.1.1 Background

The origins of classification of texts can be traced back to the paper published by Mosteller and Wallace in 1964[1], where they attempted to determine the authorship of the federalist papers. Of course this field has expanded since then to classifying unlabelled works by genre and time period as well as by author. We however will focus more on the genre classification problem which is more in line with the scope of this work.

2.2 Machine learning in Natural language processing

The discipline of machine learning revolves around the ambitious goal of creating computer programs that are able to learn. The word learn has many possible meanings, for the sake of clarity the definition used in this work is given here[2]:

A computer program is said to learn from experience E with respect to some class of task T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

It is an extremely broad field that has existed since Alan Turing first conceived of the idea of a machine thinking for itself. This resulted in the proposition of the Turing test in his paper[3] in the early fifties. The first actual example of a practical machine learning program was not complete until almost a decade later[4]. Arthur Samuel's program demonstrated an improvement in its play of the game of checkers as it encountered more opponents. While this was the first record of a machine being able to learn albeit in a limited capacity, the field has now expanded to be a research area which has blended with many others such as artificial intelligence, control theory and statistics. A brief overview of the basic concepts and techniques used in machine learning is offered below as well as a short demonstration of how they may be applied to our goal of examining the potential benefits of unsupervised learning of genres.

2.2.1 Core concepts

Working from the definition given above, a machine learning problem may be broken down in to three parts. The task to be addressed by the system must found and defined, the first step is to find the purpose of the system. In the case of this research the task is to evaluate the use of unsupervised learning in identifying genres. The next task is to obtain some form of examples from which the system can learn from. This will be provided by the Brown corpus collection of texts. Finally the problem must have some method of evaluation. Due to the unsupervised approach this will be carried out manually.

In order to give a more formal structure to our research, we will introduce some terminology to allow us to adequately address the problem. Supposing that a dataset has been collected where each record has n components, each record may be written as a vector

$$\mathbf{x = (X1,X2,X3,...,Xn)}$$

which is referred to as an instance. Each component is known as a feature. In this work the individual texts are records while the words are the features.

There are four main classes of problems which cover most of machine learning[5]. The first is referred to as classification. It takes a dataset of classified instances, instances which have already been assigned to predefined categories and the then uses these as a training set to create an algorithm which can deduce a way to classify new instances. These problems are mainly solved by supervised learning due to the need to pre-define the categories and label the data. The second is numeric prediction which similarly works on training data but replaces the search for a class value with a search for a numerical value for each instance.

Association learning forms the third type of problem and the first which is more suited to unsupervised learning, In such cases the goal is to discover significant associations between different features and uncover structure that may be hidden in the data. The final problem is known as clustering, the aim of which is to partition the instances into useful categories. Useful, in this sense can be hard to define properly and can often involve subjective evaluation. The research in this document aims to solve a problem which is a mixture of the third and fourth types. The aim is to find significant associations between features and use that to partition the instances into useful classes. Ideally these classes will have some connecting thread which can provide a use.

All four of the approaches outlined above remain subject to the familiar limits of statistics. Over-fitting the classifier to the dataset remains a problem and any algorithm selected be it supervised or not must be sufficiently robust to the noise inherent in real world data. We provide a brief overview of the supervised and unsupervised learning algorithms which have proven themselves in the past to be effective tools for research in this field.

2.2.2 Supervised approaches

To formally layout the purpose of the learning algorithms, one supposes that there exists some unknown function *f(x)* that will classify correctly each instance. The aim is to determine the this function, to which end we define the new function *h(x)* as a classifier and then attempt to approximate *h(x)* to *f(x).* The purpose of the supervised learning algorithms is to use the values taken by *f(x)* as a training set and fit *h(x)* to those values. This is done such that it can be used as accurately as possible to estimate the values of *f(x)* for new data. There are many alternatives which have been designed to solve this problem and while each performs best under different circumstances, two of the most widely used are included below.

2.2.2.1 Naive Bayes Classifier

Based on the Bayesian theorem this is a probabilistic classifier which works on the assumption that all features are independent. This is often unrealistic in the real world hence the term naïve. Despite this drawback it often performs very well at actual classification problems. Indeed research has been conducted as to which it performs better than expected on complex problems given its simplistic assumptions[6]. These assumptions make it naturally resistant to over-fitting and it works well as a classifier on datasets with a high feature count. Due to the training of the algorithm being linear in regards to both features and instances, it is a computationally efficient algorithm[7]. However despite these positives it is often outperformed by more specialised algorithms especially where feature dependence is common. In addition it works well on small datasets but doesn't scale optimally with larger training sets.

2.2.2.2 Logistic Regression

Logistic regression is used to classify an instance in to one of two categories, this classifier does not assume a linear relationship between the target class and the features. It operates through a combination of linear regression and a transformation through the logistic function. It has proven it's worth in the medical and social science fields mainly for it's ability to deal with data that doesn't fit with even statistical assumptions. The three main weakness of logistic regression are sensitivity to over fitting, a large number of categories and the requirement for a large sample size.

2.2.3 Unsupervised approaches

The unsupervised approach to a problem must be able to handle the raw data without any labels. To this end there are many algorithms which are capable of providing a solution. These range in complexity and style from artificial neural networks to data clustering, with many variations and fringe algorithms in between. As with the supervised algorithms there is no one algorithm which consistently exhibits superior performance across every environment[8]. However some algorithms are more commonly used than others and will be given a brief description below.

2.2.3.1 K-means

The K-means algorithm is one of the most popular data clustering algorithms. It was first referred to a as K-means in the late sixties even though the idea was around earlier[9][10].It creates K clusters or categories each of which have a mean or average of instances which are in that cluster. When the new instance is added to the system it is added to the cluster whose mean is closest to the new instance. Usually this is done via Euclidean distance but others methods for obtaining the distance between two vectors can be used. Various modifications of the algorithm use other distance measures.

The most common algorithm uses and iterative refinement technique. Given an initial set of *k* means the algorithm alternates between two steps.

Assignment step

Each instance is assigned to a cluster whose means yields the least within-cluster sum of squares(WCSS). The sum of squares is simply the squared euclidean distance, this provides the nearest mean.

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \leq \|x_p - m_j^{(t)}\|^2 \ \forall j, 1 \leq j \leq k\},$$

Update step

The new means are calculated to be the centroid of the instances in the new clusters.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

The algorithm has converged when the assignments of instances to clusters no longer change. Both steps optimize the WCSS objective, hence the algorithm will converge to a local optimum. One of the faults of the K-means algorithm is that there is no guarantee that the global optimum will be reached.

Chapter 3: Design

3.1 Introduction

The purpose of this chapter is to provide a detailed description of the design decisions for the research, which builds on the foundation of the related work discussed in **Chapter Two.** Firstly a set of requirements are gathered that will ensure that the research question outlined in **Chapter One** is answered, the objectives met and the various challenges overcome.

3.2 Requirements

Due to the wide range of research objectives and challenges, the requirements for this work will be split on the lines of specific aims. The first set of requirements are related to the process of corpus selection summarized by the objective to: *To select a wide corpus which has a large variety of reasonably sized texts which cover various subject areas and has recognition in the research community for quality.*

Wide corpus

- *To select a wide corpus which has a large variety of texts*

It is important in all research and particularly in the construction of machine learning systems, that the product is not overly fine tuned to a given sample. To create a system which has a balance between a high success rate and the ability to be used in multiple situations one must a have a training and testing set which balance the qualities of a high number of texts per genre versus a high number of genres. The ideal corpus would have an extremely large number of clearly defined independent genres each containing an extremely large number of unique texts. However this is unrealistic and as such compromises must be made in order to have the best available corpus to work with.

In this particular instance the width of the corpus or the number of genres was considered more vital to the good progress of the research than an extreme depth of texts in all available genres. This was due to the aim of the research being to discover the differences between the various genres rather than attempting to define the characteristics of a given genre. As a result the broadest corpus possible was selected which had integrity as resource for natural language processing.

Quality of the corpus

- *The corpus must have be recognized in the research community for its quality*

The quality of the corpus is related to the extraction of features and the actual content of the texts contained within. Due to the nature of the classifier and the features selected there were several items which had to be considered. The selection of individual words as the main feature for the classifier meant that while the grammar of the texts in question was not of supreme importance the spelling of the individual words was. Each spelling error would create a new variable for each text which would not only return an incorrect result but also significantly increase the amount of processing time needed as each text would have an extra variable.

In addition to the basic quality required to carry out functional research on a variety of texts there was the need for the resources used to be recognized as of high quality. As little research has been done currently in the area of unsupervised genre classification as can be seen in **Chapter Two.** There is a need for new research to be based on a solid foundation. Hence developing a corpus or selecting a more obscure corpus may have indicated a bias in the research. The other problem with developing a corpus is that it is time consuming and difficult to select and gather the texts and ensure they are of representative quality with no errors. As outlined above certain classes of errors in the source data can have crippling effects on the system.

The research also required some form of organization to be present in the corpus. To compare the classification between the standard methods and the unsupervised learning results, a baseline of classified documents was required. This would take the form of documents that were previously sorted into some form of genres and the information on the genres would be easily obtained. This naturally was a significant limiting factor on the corpora which could be potentially selected for use. While other corpora could have been adapted, using a pre-established corpus saved effort due to the need for manual pre-labelling and removed any chance of bias in the results.

Size of Corpus

The size of the corpus is an important factor in the selection process. One of the issues highlighted by the survey of related work is the need for a large corpus to provide a decent training and test set. Both of these are required to create a strong classifier and evaluate it fairly. Due to the nature of the results and the need to manually evaluate the classification a smaller size would have made the evaluation a simpler task. This requirement was not massively influential in the corpus selection decision but was considered as part of the selection process.

The next set of requirements was based on the need to extract features which would be common across all documents and could be used in an unsupervised learning system. This is derived from the objective: *To design and build a system which is capable of using unsupervised learning to categorise texts using select features.*

Select features

- *To categorize texts using select features*

The features of a text are to be considered any identifying characteristics that are either individually or in combination, can be used to uniquely identify a text. As it is impossible to comprehensively check ones features against features extracted using the same methodology from every other possible text reasonable assumptions must be made. For example the length of a text is one of the easiest characteristics to obtain but it is most definitely not a unique characteristic which can be used to identify a text. To this end it was decided to look at a lower level than the overall length of the text. While individual sentences structure can be extremely useful in NLP (generally referred to as POS or part of speech tagging) for this task of categorising texts it was seen a overly unique and it would be difficult to relate to other texts.

The next step down was that of considering the individual words in any given text. The most obvious and simplest one was to obtain the most common word used in a given text however that was quickly proven to be a trivial solution even with the removal of the stop words. The possibility of more words being used to form a unique identity for a text was considered. This is more commonly known as the bag-of-words approach, every word in a text is counted and then for each word that occurs in the text the number of appearances are counted and recorded. Each text is rendered down into a single vector with a length equal to that of the dictionary used.

Of course each word may not and in most cases will probably not occur in every text, this can be seen in the sparse vectors which are produced. As such some form of dictionary must be created which will create a place-holder for each word. The dictionary for this work will be constructed from the training set. This allows us to compare like with like and simplifies the calculations. This was considered to be a practical approach which had been used previously in other research both in combination and in isolation. It enabled easy comparisons between texts while also allowing texts to be uniquely identified.

Unsupervised learning system

- *To design and build a system which is capable of using unsupervised learning*

While there are a vast array of unsupervised learning algorithms available, it was felt best to start off with a simple approach. As with the supervised learning approaches where the simple naïve Bayes classifier can often produce excellent results, the option of using data clustering was selected as a primary approach. This approach managed to limit the amount of human input or supervision in the system. The K-means approach still required the input of the number of classes to be developed by the system. This was as much as could be limited without doing excessive calculations such as comparing every relation between each text.

The relation selected to compare between each text was that of the cosine distance or similarity between the sparse vectors which were the result of the feature selection process. The cosine distance is actually the cosine of the angle that exists between the two vectors and the formula is given below:

$$\vec{a} \cdot \vec{b} = \|\vec{a}\|\|\vec{b}\| \cos\theta$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

This is measurement of orientation rather than magnitude, it can be seen as a comparison between documents on a normalized space. The consideration of the magnitude of each word count of each document is not the only factor taken into account. The angle is also taken into account. The advantage of this method over alternatives like Euclidean distance is that it tends to remove the length (which manifests as higher term counts) as a variable. The angle between two vectors will not change due to the length of a vector extending.

K-means clustering is a method of partitioning *n* observations into *k* clusters in which each observation belongs to the cluster with nearest mean. In this case each observation is a single document and each cluster is a genre. It was selected as is designed for vector quantization which makes it ideal for the particular use found in this work. As the features could be easily represented numerically and arbitrary distance functions were not being utilised(with the cosine similarity being used instead) the usual drawbacks of the function were not applicable.

Chapter 4: Implementation

4.1 Introduction

This section will deal with the implementation of the system and the various choices made along with the reasoning for those choices. The major choices involved include the selection of programming language, the libraries that could be made use of and the algorithm for determining the classification. These choices will build out from the design decisions made in **Chapter Three** and the work covered in C**hapter Two.**

4.2 Language selection and library choices

The choice of language was heavily influenced by the resources made available in the area of the task of creating an unsupervised learning system for a large collection of texts (also known as a corpus). These resources came in the form of programming libraries some of which had built in access to the various corpus which are popular in the natural language processing community.

4.2.1 Python

Python was selected as the language of choice primarily due to the availability of the excellent natural language tool kit library. The advantages of selecting python went beyond the simple wide availability of libraries. Variables in python are not dynamically typed and as such are more flexible than the fixed variable types present in languages like C and Java. This is helpful in switching between the raw textual data and the processed numerical output both of which are used in this work.

The two other major considerations were that python is by design a highly readable language and also like other dynamic languages python can be utilised as a scripting language. The readable nature of python not only made the creation of a system easier it also allows more people to understand the finished product which is always important in research into systems which make choices for us. The scripting nature of python was helpful in testing the various methods of classification and manipulating the results. This was due to the initial processing of the data being approximately 90% of the total run time of the project as it is in most learning systems. It also allows future researchers to easily build on this work without a large investment of time.

4.2.2 NLTK

The main library selected for use in this project was the Natural Language Tool Kit[11], a library which specialises in tools for manipulating and extracting information from text documents. It also provides access to a large number of corpora which can be downloaded and easily called by functions which are included. This meant that the Brown corpus could be easily loaded and separated out into the various texts that comprise it. The ability to load the individual texts into an array with a single function call was a significant tool in choosing to implement the system with the aid of the NLTK library. This made dividing the texts for the purpose of processing them to be simpler. The NLTK library also provides a reasonable implementation of the K-means clustering algorithm which was selected for use in the system.

4.3 Variations

The Brown corpus was put through the system a number of time with several variations made to the system to view the changes altering the feature selection and means initialization would have on the results. This was done to see the effects these changes would have on the results and so the variations the changes made in the results could be used to evaluate the optimized method for genre classification. The changes can be split in to two distinct sections, the inclusion or removal of stop words and the inclusion or removal of predefined starting means. The effects of these variations will be discussed more in the next chapter.

Chapter 5: Evaluation

5.1 Introduction

The purpose of this chapter is to evaluate the results of the research carried out against the research objectives set out in Chapter One. To this end a plan for the evaluation procedure is first outlined. The major components of the system are then evaluated according to this plan, then the outcome of the evaluation is discussed. The various changes made to alter the results, provide an easy method to break down the results.

The results provide the categorization of the five hundred texts(title, author and order in Appendix) and the various categories are labelled by the most common words in each category. Due to the need for manual evaluation only a sample of results will be discussed and analysed due to resource constraints. The criteria used will be:

- *Does the placement of the chosen samples seem appropriate based on what we know of the text?*
- *Does the placement of the text seem consistent with other similar texts?*
- *Does the categories chosen seem internally consistent?*

The above criteria will evaluate the individual test runs while the modifications as a whole will have the following criteria:

- *Do the variations alter the clustering of the chosen sample texts?*
- *Do the changes made by the variations seem appropriate?*
- *Which setup provides the best classification for the analysed texts?*

5.2 Results

The samples chosen for analysis were the first ten texts which are collected under the Brown corpus category of political news articles. These texts were chosen due to the fact that they comprise a field which has a consistent theme and style of writing. The details of the samples is given below:

A01	Atlanta Constitution Political Reportage	
A02	Dallas Morning News Political Reportage	Chicago Tribune Political Reportage
A03	Chicago Tribune Political Reportage	
A04	Christian Science Monitor Political Reportage	
A05	Providence Journal Political Reportage	
A06	Newark Evening News Political Reportage	
A07	New York Times Political Reportage	
A08	Times-Picayune, New Orleans Political Reportage	
A09	Philadelphia Inquirer Political Reportage	Chicago Tribune Political Reportage
A10	Oregonian, Portland Political Reportage	

5.2.1 Stop words

These texts should be relatively similar however the differences which appear in the results show the vast differences which small changes can make. The most significant can be seen in the removal of stop words. In most research stop words such as *and, the, if, a* are removed from the sample texts to reduce the computation time and remove commonalties between texts however this can also affect the outcome of the experiment negatively. This can be seen when the results for the stop word included and excluded experiments are compared.

Stop words included

[1, 1, 1, 6, 1, 1, 1, 1, 6, 1]

Stop words excluded

[14, 5, 14, 5, 14, 14, 10, 10, 10, 5]

As you can see we have a much more fractured cluster when the stop words are excluded. Not only this but the breaking of the clustering is also far more inconsistent than it is where all the words are included. This implies that far from being inconsequential, stop words are some of the more important features that can be found when it comes to genre classification.

5.2.1.1 Included

[1, 1, 1, 6, 1, 1, 1, 1, 6, 1]

Where the stop words are included the breaks are where the Christian science and Philadelphia Inquirer are classified as religious articles rather than political. This is due to the nature of the classification algorithm. Each instance can only be assigned to one genre where it may be appropriate for several related genres. This limitation can again be seen in the effects on social science texts. These texts include ones on African demographics and bio-warfare which under this system were classed as political texts rather than scientific texts like others in the Brown corpus such as ones on physics and chemistry.

5.2.1.2 Excluded

[14, 5, 14, 5, 14, 14, 10, 10, 10, 5]

The excluded stop words clearly degenerated the ability of the system to give coherent clusters based on the types of writing or topics covered. However there is still some form of identification taking place even if the the classes are not easily identified based on the summary information provided by the corpus. This can be seen by the fact that there are only three genres present.

Chapter 6: Conclusion and Future Work

This chapter will discuss the contribution made by this work to the areas of unsupervised learning and genre classification, what work could be done to build on this work and what conclusions can be reached as a result of this work.

6.1 Contribution

This work has made several observations which can be considered of use to the relevant fields. The K-means algorithm was proved to be a strong viable choice for solving an unsupervised learning task. While simple in approach, like the naïve Bayes classifier it can produce results which are more than acceptable. This clearly is proof that feature selection is at least as important as the choice of algorithm. In fact as we can see from the differences when features were restricted it may be even more important.

In the area of genre classification we can see that the categories or genres defined are more judgement calls than hard and fast labels. They provide a means for us to sort texts but the way one person or system may label the texts is quite different to another. As we can see here the unsupervised approach provides us with classes which are defined by the most common parts rather than the one word labels provided by supervised systems. While this is a disadvantage in creating human readable labels for genres, the question remains which is more important,naming a genre or gathering similar texts together?

6.2 Future work

While much has been accomplished in this work there is still much left in this particular field to explore. The main points which could be improved upon are:

- More features being extracted
- Multiple genres per text, allowing overlap
- Combination of methods, possibly mixing supervised and unsupervised techniques
- More in-depth analysis of the results

6.3 Conclusion

The main conclusions of this work can be summarized in three key points. The advantage of an unsupervised approach is that it can remove most of the problems of bias present in setting up genres. The second point is that the preconceived genres are not necessarily the most appropriate for a given text and that most times there are several overlapping genres that a given text could be sorted into. The results only show the class *most* suited to a text not all the genres an instance could be placed in. The third and final point is that the feature selection is possibly the most important part of the process.

All features possible should be included not just those which emphasise difference for the most comprehensive results.

References

Bibliography

1: Mosteller and Wallace, The Disputed Federalist Papers, 1964

2: T.M. Mitchell, Machine learning, 1997

3: A.M Turing, Computing machinery and intelligence,, 1950

4: Arthur L. Samuel, Some studies in machine learning using the game of checkers, 1959

5: I.H. Witten and E. Frank, Data Mining: Practical machine learning tools andtechniques, 2005

6: H.Zhang, The optimality of naive Bayes, 2004

7: E. Frank, M. Hall and B. Pfahringer,, Locally weighted naive bayes, 2002

8: D.HWolpert, The lack of a priori distinctions between learning algorithms, 1996

9: MacQueen, J. B, Some Methods for classification and Analysis of Multivariate Observations, 1967

10: Steinhaus, H., Sur la division des corps matériels en parties, 1957

11: S. Bird, E. Loper and E. Klein, Natural Language Processing with Python, 2009

12: Henry Kucera, W. Nelson Francis, Computational Analysis of Present-Day American English, 1967, http://icame.uib.no/brown/bcm-los.html#a

Appendix

Full results

Stop words in, staring means at centre of brown corpus categories

[1, 1, 1, 6, 1, 1, 1, 1, 6, 1, 9, 4, 9, 9, 4, 4, 4, 4, 4, 4, 4, 4, 4, 6, 4, 0, 1, 1, 4, 4, 4, 4, 9, 6, 8, 6, 9, 9, 4, 5, 9, 9, 1, 2, 1, 1, 9, 1, 1, 1, 1, 10, 9, 10, 9, 6, 1, 1, 1, 1, 1, 1, 1, 2, 3, 6, 2, 4, 2, 1, 1, 4, 1, 4, 4, 2, 2, 4, 2, 2, 2, 2, 6, 6, 2, 4, 2, 4, 8, 3, 2, 2, 2, 2, 2, 6, 1, 1, 3, 8, 8, 6, 6, 0, 10, 1, 4, 1, 1, 9, 1, 7, 7, 0, 4, 4, 0, 9, 4, 0, 1, 1, 2, 1, 4, 1, 2, 2, 1, 7, 5, 0, 1, 8, 1, 0, 1, 1, 0, 1, 4, 2, 9, 2, 1, 9, 4, 1, 4, 10, 4, 1, 2, 1, 6, 8, 10, 6, 14, 2, 5, 7, 9, 1, 4, 1, 5, 2, 9, 2, 4, 10, 5, 1, 0, 14, 5, 8, 9, 0, 6, 7, 2, 6, 8, 9, 9, 1, 3, 8, 6, 1, 10, 2, 5, 2, 2, 8, 6, 8, 1, 8, 9, 2, 8, 2, 2, 1, 8, 6, 2, 2, 6, 8, 8, 2, 2, 8, 1, 5, 5, 2, 10, 1, 10, 9, 2, 4, 10, 9, 8, 3, 8, 10, 8, 13, 9, 2, 5, 4, 5, 5, 5, 1, 2, 0, 9, 6, 5, 6, 6, 7, 2, 6, 5, 5, 2, 2, 2, 8, 8, 2, 2, 5, 0, 1, 8, 1, 8, 0, 0, 3, 3, 0, 7, 3, 1, 6, 1, 3, 9, 1, 1, 0, 8, 3, 3, 1, 3, 9, 0, 0, 8, 1, 7, 7, 7, 7, 1, 7, 7, 1, 0, 2, 8, 0, 6, 0, 4, 6, 6, 7, 8, 7, 7, 8, 0, 8, 8, 0, 1, 1, 6, 5, 2, 7, 1, 1, 0, 6, 6, 1, 2, 6, 7, 6, 1, 8, 6, 8, 0, 2, 8, 3, 8, 2, 2, 6, 9, 6, 8, 6, 3, 9, 9, 8, 2, 9, 2, 2, 8, 2, 3, 7, 0, 0, 0, 0, 7, 0, 7, 7, 7, 7, 10, 10, 10, 5, 10, 10, 11, 10, 5, 10, 10, 10, 10, 10, 10, 5, 11, 12, 10, 10, 10, 5, 9, 11, 5, 10, 9, 11, 12, 13, 12, 11, 10, 11, 10, 10, 13, 10, 10, 10, 11, 10, 10, 10, 11, 10, 10, 10, 13, 11, 4, 10, 13, 11, 10, 10, 11, 10, 10, 11, 11, 11, 10, 11, 12, 11, 5, 10, 11, 10, 11, 10, 10, 10, 13, 13, 12, 10, 10, 10, 10, 10, 11, 10, 10, 10, 5, 10, 10, 5, 10, 5, 5, 11, 10, 11, 5, 5, 13, 11, 5, 11, 13, 11, 5, 13, 13, 13, 5, 13, 10, 11, 5, 5, 11, 10, 10, 5, 5, 13, 10, 14, 2, 11, 2, 4]

Stop words out no starting means

[14, 5, 14, 5, 14, 14, 10, 10, 10, 5, 5, 14, 10, 5, 5, 0, 0, 0, 0, 2, 0, 0, 0, 10, 2, 10, 14, 10, 5, 5, 5, 14, 5, 10, 10, 10, 5, 5, 2, 5, 5, 2, 10, 2, 10, 10, 5, 10, 10, 10, 5, 14, 5, 10, 5, 5, 10, 10, 10, 5, 1, 10, 1, 5, 2, 10, 5, 2, 2, 5, 10, 5, 5, 5, 2, 14, 2, 5, 2, 2, 2, 2, 5, 5, 14, 4, 2, 2, 10, 2, 2, 2, 2, 2, 1, 2, 10, 10, 3, 2, 4, 5, 5, 10, 10, 11, 10, 5, 4, 0, 5, 11, 10, 4, 4, 2, 4, 2, 10, 10, 10, 10, 2, 10, 10, 4, 2, 2, 10, 10, 5, 10, 1, 10, 1, 5, 10, 10, 4, 10, 5, 2, 10, 2, 10, 5, 1, 10, 5, 10, 5, 10, 2, 10, 2, 5, 5, 5, 13, 5, 5, 10, 5, 5, 5, 5, 10, 2, 5, 5, 5, 10, 2, 10, 4, 14, 5, 10, 2, 2, 5, 5, 2, 5, 8, 5, 10, 13, 2, 5, 5, 14, 10, 2, 5, 2, 2, 5, 2, 5, 11, 5, 2, 2, 10, 14, 5, 10, 4, 5, 5, 1, 10, 10, 10, 2, 2, 5, 4, 2, 2, 2, 14, 4, 14, 5, 2, 2, 14, 5, 10, 2, 2, 14, 5, 13, 10, 2, 5, 5, 5, 5, 5, 13, 2, 2, 2, 2, 10, 5, 6, 10, 14, 10, 2, 2, 2, 2, 13, 2, 5, 5, 2, 5, 4, 4, 5, 10, 10, 4, 4, 2, 4, 4, 10, 4, 5, 4, 10, 2, 5, 1, 10, 4, 4, 4, 4, 10, 2, 5, 4, 4, 4, 10, 4, 7, 7, 7, 4, 7, 4, 10, 4, 5, 4, 4, 2, 4, 4, 7, 4, 7, 10, 7, 7, 4, 2, 2, 10, 4, 10, 10, 10, 10, 2, 7, 10, 10, 4, 5, 10, 10, 2, 4, 4, 4, 10, 4, 10, 4, 4, 1, 10, 2, 4, 5, 2, 4, 5, 10, 8, 5, 2, 5, 2, 5, 2, 5, 2, 5, 4, 2, 4, 10, 10, 4, 4, 4, 7, 4, 10, 7, 7, 10, 11, 10, 10, 5, 10, 10, 11, 14, 2, 14, 5, 5, 14, 10, 14, 5, 11, 12, 11, 14, 11, 14, 5, 11, 2, 14, 9, 14, 13, 13, 12, 11, 11, 11, 10, 11, 2, 10, 10, 11, 11, 11, 1, 10, 11, 14, 11, 10, 13, 11, 5, 11, 13, 14, 5, 14, 11, 14, 10, 11, 11, 11, 11, 11, 13, 11, 5, 14, 11, 10, 11, 14, 11, 11, 13, 13, 12, 14, 11, 13, 11, 5, 14, 10, 11, 14, 10, 14, 10, 14, 14, 14, 5, 14, 14, 11, 5, 13, 13, 14, 2, 11, 14, 11, 5, 13, 13, 13, 2, 13, 14, 11, 14, 13, 11, 10, 14, 5, 2, 14, 11, 14, 14, 11, 2, 2]

Python Code

```
from __future__ import division
import nltk
import random
import re, pprint, os
import numpy
import nltk.corpus
from nltk.corpus import brown
from nltk import cluster
from nltk.cluster import util
from nltk.cluster import api
from nltk.cluster import euclidean_distance
from nltk.cluster import cosine_distance
texts = brown
print "Read in", len(texts.fileids()), "documents..."
numdoc=len(texts.fileids())
print "The first five are:", texts.fileids()[:5]
doc_id=0
#unique terms- stop words
unique_terms = list(set(texts.words())-set(nltk.corpus.stopwords.words('english')))
idf_count=[]
for word in unique_terms:
    idf_count.append(0)
print "Found a total of", len(unique_terms), "unique terms"
# Function to create a BOW for one document.  This is called with the fileid
# of one of the files in the corpus.  We convert its list of words into an
# nltk.Text object so we can use the count method.  Then for each of
# our unique words, we have a feature which is the count for that word
def BOW(document,_id):
    #print type(document)
    document = nltk.Text(texts.words((document)))
```

```
    #remove stop words
    filtered_words = [w for w in document if not w in nltk.corpus.stopwords.words('english')]
    document=filtered_words
    word_counts = []
    for word in unique_terms:#tf calc
        word_counts.append(document.count(word)/len(document))
        if(document.count(word)>0):
            idf_count[unique_terms.index(word)]+=1
    #_id+=1
    #print "done"
    return word_counts

#function to add in the idf metric to the vectors, not 100% if it works right
def IDF(vectors):
    idfvectors=[]
    #cant iterate over int need to use something else
    for v in vectors:
        for n in range (1,len(unique_terms)-1):#index out of range
            idfvectors[vectors.index(v)][idf_count.index(n)]=vectors[vectors.index(v)][idf_count.index(n)]*numpy.log(numdoc/n)
    return idfvectors

# And here we actually call the function and create our list of tf vectors.
vectors = [numpy.array(BOW(f,doc_id)) for f in texts.fileids()]
print "TFVectors created."
final_vectors=vectors
print "IDFVectors created."
print "First 10 words are", unique_terms[:10]
print "First 10 counts for first document are", vectors[0][0:10]
# We now have a vector ready to feed to our clusterer of choice.
#more rigourus mean selection possibly needed
means= [vectors[22],vectors[60],vectors[80],vectors[98],vectors[115],vectors[160],vectors[250],vectors[300],vectors[350],vectors[400],vectors[415],vectors[430],vectors[450],vectors[480],vectors[495]]
#setup the initial means
#currently testing with no inital means, uncomment below to change this
clusterer = cluster.KMeansClusterer(15, cosine_distance,initial_means=means)
```

```
clusters=[]
clusters = clusterer.cluster(vectors, True, trace=True)
#for i in xrange(0,10):
 #  clusters.append( clusterer.cluster(vectors, True, trace=True))
   #go cluster with kmeans
  # print'clustering done'
   #print 'Clustered:', vectors
print 'As:', clusters
print 'Means:', clusterer.means()
ntemp=clusterer.means()
#create space for means ids
x = [[0 for i in range(15)] for j in range(15)]
#identify the top 15 words in each category
for j in xrange(15):
        for i in xrange(15):
                idx = numpy.argmax(ntemp[j])
                x[j][i]=idx; ntemp[j][idx]=0
#print the key words
for y in x:
    for t in y:
        y=unique_terms[t]
```

Brown corpus Data[12]

A01	Atlanta Constitution	Political Reportage
A02	Dallas Morning News	Political Reportage
	Chicago Tribune	Political Reportage
A03	Chicago Tribune	Political Reportage
A04	Christian Science Monitor	Political Reportage
A05	Providence Journal	Political Reportage
A06	Newark Evening News	Political Reportage
A07	New York Times	Political Reportage
A08	Times-Picayune, New Orleans	Political Reportage
A09	Philadelphia Inquirer	Political Reportage
	Chicago Tribune	Political Reportage
A10	Oregonian, Portland	Political Reportage
A11	Sun, Baltimore	Sports Reportage
A12	Dallas Morning News	Sports Reportage
A13	Rocky Mountain News	Sports Reportage
	Dallas Morning News	Sports Reportage.
A14	New York Times	Sports Reportage.
A15	St. Louis Post-Dispatch	Sports Reportage
A16	Chicago Tribune	Society Reportage
A17	Rocky Mountain News	Society Reportage
	Dallas Morning News	Society Reportage
A18	Philadelphia Inquirer	Society Reportage
	Times-Picayune, New Orleans	Society Reportage
A19	Sun, Baltimore	Spot News
A20	Chicago Tribune	Spot News
A21	Detroit News	Spot News
A22	Atlanta Constitution	Spot News
A23	Oregonian, Portland	Spot News
A24	Providence Journal	Spot News
A25	San Francisco Chronicle	Spot News
	Chicago Tribune	Spot News
A26	Dallas Morning News	Financial Reportage
A27	Los Angeles Times	Financial Reportage
	Philadelphia Inquirer	Financial Reportage
A28	Wall Street Journal	Financial Reportage
A29	Dallas Morning News	Cultural Reportage.
A30	Los Angeles Times	Cultural
	Sun, Baltimore	Cultural Reportage
A31	Miami Herald	Cultural Reportage
A32	San Francisco Chronicle	Cultural Reportage
A33	Washington Post	Cultural Reportage

A34	New York Times	News of the Week in Review
A35	James J. Maguire	A Family Affair
A36	William Gomberg	Unions and the Anti-Trust Laws
A37	Time	National Affairs
A38	Sports Illustrated	A Duel Golfers Will Never Forget
A39	Newsweek	Sports
A40	Time	People. Art & Education
A41	Robert Wallace	This Is The Way It Came About
A42	Newsweek	National Affairs
A43A	U. S. News & World Report	Better Times for Turnpikes
A43B	U. S. News & World Report	A Plan to Free U. S. Gold Supply
A44A	John Tebbel	Books Go Co-operative
A44B	Gilbert Chapman	Reading and the Free Society
B01	Atlanta Constitution	Editorials
	Washington Post	Editorials
B02	Christian Science Monitor	Editorials
B03	Detroit News	Editorials
	Chicago Dally Tribune	Editorials
B04	Miami Herald	Editorials
	Los Angeles Times	Editorials
B05	Newark Evening News	Editorials
B06	St. Louis Post-Dispatch	Editorials
B07	New York Times	Editorials
B08	Atlanta Constitution	Columns
B09	Christian Science Monitor	Columns
B10	Sun. Baltimore	Columns
B11	Los Angeles Times	Columns
B12	Newark Evening News	Columns
B13	Times-Picayune, New Orleans	Columns
B14	Atlanta Constitution	Columns
B15	Providence Journal	Letters to the Editor
B16	Chicago Tribune	Voice of the People
B17	Newark Evening News	What Readers Have to Say
	Washington Post	Letters to the Editor
B18	New York Times	Letters to the Times
	Detroit News	The Public Letter Box
B19	Philadelphia Inquirer	The Voice of the People
	Detroit News	The Public Letter Box.
B20	Nation	Editorials
B21A	Gerald W. Johnson	The Cult of the Motor Car
B21B	James Deakin	How Much Fallout Can We Take
B22	Commonweal	Week by Week

B23A	William F. Buckley, Jr.	We Shall Return
B23B	James Burnham	Tangle in Katanga
B24	Time	Reviews
B25A	Alexander Werth	Walkout in Moscow
B25B	Peter Solsich, Jr.	The Armed Superpatriots
B26	National Review	To the Editor
B27	Saturday Review	Letters to the Editor
C01	Chicago Daily Tribune	Reviews
	New York Times	Reviews
C02	Christian Science Monitor	Reviews
C03	New York Times	Reviews
C04	Providence Journal	Reviews
C05	Christian Science Monitor	Reviews
C06	Wall Street Journals	Reviews
	New York Times	Reviews
C07	New York Times	Reviews
C08	Providence Journal	Reviews
C09	New York Times	Reviews
C10	Providence Journal	Reviews
C11	New York Times	Reviews
C12	Christian Science Monitor	Reviews
C13	Wall Street Journal	Reviews
	New York Times	Reviews
C14	New York Times	Reviews
C15	Life	Reviews
C16	Saturday Review	Reviews
C17	Time	Reviews
D01	William Pollard	Physicist and Christian
D02	Schubert Ogden	Christ Without Myth
D03	Edward E. Kelly	Christian Unity in England
D04	Jaroslav Pelikan	The Shape of Death
D05	Perry Miller	Theodore Parker: Apostasy With in Liberalism
D06	A Howard Kelly	Out of Doubt into Faith
D06B	Shirley Schuyler	Not as the World Giveth
D06C	Nathanael Olson	Are You in Orbit?
D07	Peter Eldersveld	Faith Amid Fear
D08	Schuyler Cammann	The Magic Square of Three
D09	Eugene E. Golay	Organizing the Local Church
D10	Huston Smith	Interfaith Communication: The Contemporary Scene
D11	Paul Ramsey	War & the Christian Conscience

D12	Kenneth Underwood and Widen Jacobson	Probing the Ethics of Realtors
D13A	Donald H. Andrews	The New Science & the New Faith
D13B	George Bo Longstreet	The Seeming Impossible
D14	Kenneth S. Latourette	Christianity in a Revolutionary Age
D15	Ernest Becker	Zen: A Rational critique
D16A	Anonymous	What the Holy Catholic Bible Teach
D16B	Harold Brenneman	Notice You May Do As You Please
D17A	Anonymous	Guideposts: 15th Anniversary Issue
D17B	J. I. Rivero	The Night Our Paper Died
E01A	Ben Welder	Henri de Courcy: Jr. Mr. Canada
E0lB	Joe Welder	The Mark of the Champion
E02A	Dorothy Schroeder	Plant a Carpet of Bloom
E02B	Anonymous	Avocado is Something Special
E03	D. F. Martin	Will Aircraft or Missiles Win Wars?
E04A	Harris Goldsmith	The Schnabel Pro Arte Trout
E04B	Robert C. Marsh	The True Sound of a Solid Second
E04C	R.D.D.	Review of Adam, Giselle
E05A	Paul Nigro	The Younger Generation
E05B	Patricia Barney	Use of Common Sense Makes Dogs Acceptable
E05C	Anonymous	The Malady Lingers On
E06	Joseph E. Choate	The American Boating Scene
E07	Paul Larson and Gordon Odegard	How to Design Your Interlocking Frame
E08	Don Francisco	Formulas and Math Every Hot Rodder Should Know
E09A	Don McMahan	The Week at Ben White Raceway
E09B	Edith Shaw	The Picture at Del Mar
E10	Larry Koller	The New Guns of 61
E11	Idwal Jones	Santa Cruz Run
E12	Julia Newman	Travel and Camera USA
E13	Robert Deardorff	Step by Step through Istanbul
E14	Ann Carnahan	Nick Manero's Cook-out Barbecue Book
E15A	Anonymous	Pottery from Old Molds
E15B	Anonymous	Knitting Knacks
E16	Hal Kelly	Build Hotei
E17A	Anonymous	This is the Vacation Cottage You Can Build
E17B	Patrick K. Snook	Care and Basic Use of the Drill Press
E18A	Lura W. Watkins	The Bridge Over the Merrimac
E18B	Boyd B. Stutler	Veteran Philippi Bridge

E19	Booth Hemingway and Stuart H. Brown	How to Own a Pool and Like It.
E20	Anonymous	What You Should Know About Air Conditioning
E21	Richard McCosh	Recreation Site Selection
E22A	Roy Harris	Roy Harris Salutes Serge Prokofieff
E22B	Helen Havener	A 30 Years War
E23	Norman Kent	The Watercolor Art of Roy M. Mason
E24	Bonnie Prudden	The Dancer & the Gymnast
E25	Walter Ho Buchsbaum	Advances in Medical Electronics
E26	Bern Dibner	Oersted & the Discovery of Electromagnetism
E27A	Mike Bay	What Can Additives Do for Ruminants?
E27B	James S. Boyd	Which Feed Bunk for You
E28	John R. Sargent	Where to Aim Your Planning
E29	Edward A. Walton	On Education for the Interior Designer
E30	Anonymous	The Attack on Employee Services
E31A	Jim Dee	Expanding Horizons
E31B	George Laycock	The Challenge
E32	E. J. Tangerman	Which Way Up. Technical or Management?
E33A	Robert Gray	Fifty Houses, One Tank
E33B	Chet Cunningham	Truck Talk
E34	Anonymous	The New Look in Signs
E35	Anonymous	The Industrial Revolution in Housing
E36	Ethel Norling	Renting a Car in Europe
F01	Rosemary Blackmon	How Much Do You Tell When You Talk?
F02	Glenn Infield	America's Secret Poison Gas Tragedy
F03	Nathan Rapport	I've Been Here Before
F04	Ruth F. Rosevear	North Country School Cares for the Whole Child
F05	Richard S. Allen	When Fogg Flew the Mail
F06A	Alice Ho Austin	Let's Discuss Retirement
F06B	Harold P. Winchester	What It Means to be Creative
F07A	Marvin Sentnor and Stephen Hult	How to Have a Successful Honeymoon
F07B	Ho Walter Yoder	Attitudes Toward Nudity
F08	Philip Reaves	Who Rules the Marriage Bed?
F09A	David Martinson	Fantastic Life & Death of the Golden Prostitute.
F09B	Isel D. Rugget	When It Comes to Carpets
F10	Jack Kaplan	Therapy by Witchcraft
F11	Lillian Pompian	Tooth-Straightening Today

F12	Marian Neater	New Methods of Parapsychology.
F13	Orlin J. Scoville	Part-time Farming
F14	Harold Rosenberg	The Trial and Eichmann
F15	John A. O'Brien	Let's Take Birth Control Out of Politics
F16	James Boylan	Mutiny
F17	John Harnsberger and Robert P. Wilkins	Transportation on the Northern Plains
F18	Bell I. Willy	Home Letters of Johnny Reb & Billy Yank
Fl9	Tristram P. Coffin	Folklore in the American Twentieth Century
F20	Kenneth Allsop	The Bootleggers and Their Era
F21A	Joseph Bernstein	Giant Waves
F21B	L. Don Leet	Introduction
F21C	L. Don Leet	The Restless Earth and Its Interiors
F22	Booton Herndon	From Custer to Korea, the 7th Cavalry
F23	Barry Goldwater	A Foreign Policy for America
F24	Peter J. White	Report on Laos
F25	David Boroff	Jewish Teen-Age Culture
F26	Amy Lathrop	Pioneer Remedies from Western Kansas
F27	Creighton Churchill	A Notebook for the Wines of France
F28	Frank O. Gatell	Doctor Palfrey Frees His Slaves
F29	Douglass Cater	The Kennedy Look in the Arts
F30	Frederic A Birmingham	The Ivy League Today
F31	Edward Do Radin	Lizzie Borden: The Untold Story
F32	Florence M. Read	The Story of Spelman College
F33	James Be Conant	Slurs and Suburbs
F34	Frederic R. Senti and W. Dayton Maclay	Age-old uses of Seeds and Some New Ones
F35	Ramon F. Adams	The Old-time Cowhand
F36	Robert Easton and Mackenzie Brown	Lord of Beasts
F37	Samuel M. Cavert	On the Road to Christian Unity
F38	Robert Smith	Baseball in America
F39	Clark E. Vincent	Unmarried mothers
F40	William Greenleaf	Monopoly on Wheels
F41	George W. Oakes	Turn Right at the Fountain
F42	James Baldwin	Nobody Knows My Name
F43	Frank Getlein and Harold C. Gardiner	Movies, Morals, and Art
F44	Gibson Winter	The Suburban Captivity of the Churches
F45	Paul C. Phillips	The Fur Trade
F46	Russell Baker	An American in Washington
F47	Clara L. Simerville	Home Visits Abroad

F48	Paul Ramsey	Christian Ethics & the Sit-In
G01	Edward P. Lawton	Northern Liberals & Southern Bourbons
G02	Arthur S. Miller	Toward a Concept of National Responsibility
G03	Peter Wyden	The Chances of Accidental War
G04	Eugene Burdick	The Invisible Aborigine
G05	Terence O'Donnell	Evenings at the Bridge
G06A	Ruth Berges	William Steinberg, Pittsburgh's Dynamic Conductor
G06B	Henry W. Koller	German Youth Looks to the Future
G07	Richard B. Morris	Seven Who Set Our Destiny
G08	Frank Murphy	New Southern Fiction: Urban or Agrarian?
G09	Selma J. Cohen	Avant-Garde Choreography
G10	Clarence Streit	How the Civil War Kept You Sovereign
G11	Frank Oppenheimer	Science and Fear
G12	Tom F. Driver	Beckett by the Madeleine
G13	Charles Glicksberg	Sex in Contemporary Literature
G14	Helen H. Santmeyer	There Were Fences
G15	Howard Nemerov	Themes and Methods: Early Storie of Thomas Mann
G16	John F. Hayward	Mimesis & Symbol in the Arts
G17	Randall Stewart	A Little History, a Little Honesty
G18	Charles W. Stork	Verner von Heidenstam
Gl9	R. F. Shaw	The Private Eye
G20	Dan McLachlan, Jr.	Communication Networks & Monitoring
G21	Brainard Cheney	Christianity & the Tragic Vision
G22	Kenneth Reiner	Coping with Runaway Technology
G23	William C. Smith	Why Fear Ideas
G24	Sanchia Thayer	Personality & Moral Leadership
G25	Stanley Parry	The Restoration of Tradition
G26	Selma Fraiberg	Two Modern Incest Heroes
G27	Matthew Josephson	Jean Hélion. The Return from Abstract Art
G28	Arlin Turner	William Faulkner, Southern Novelist
G29	Anonymous	References for the Good Society
G30	Norwood R. Hanson	Copernican & Keplerian Astronomy
G31	Irving Fineman	Woman of Valor: Life of Henrietta Szold 1860-1945
G32	Finis Farr	Frank Lloyd Wright
G33	Virgilia Peterson	A Matter of Life and Death

G34	Harry Golden	Carl Sandburg
G35	Dwight D. Eisenhower	Peace With Justice
G36	DeWitt Copp & Marshall Peck	Betrayal at the UN
G37	Gordon L. Hall	Golden Boats from Burma
G38	Bertrand A. Goldgar	The Curse of Party
G39	Edward Jablonski	Harold Arlen Happy with the Blues
G40	Gene Fowler	Skyline: A Reporter's Reminiscences of the 1920s.
G41	Lillian R. Parka and Frances S. Leighton	My Thirty Years Backstairs at the White House
G42	Harold D. Lasswell	Epilogue
G43	Robert E. Lane	The Liberties of Wit
G44	Newton Stallknecht	Ideas and Literature
G45	W. A. Swanberg	Citizen Hearst: A Biography of W. R. Hearst
G46	Henry R. Winkler	George Macaulay Trevelyan
G47	Carry Davis	The World Is My Country
G48	Francis F. McKinney	Education in Violence
G49	Paul van K. Thomson	Francis Thompson, a Critical Biography
G50	Curtis C. Davis	The King's Chevalier
G51	Ilka Chase	The Carthaginian Rose
G52	Robert L. Duncan	Reluctant General
G53	Bertram Lippincott	Indians, Privateers, and High Society
G54	Mabel W. Wheaton & LeGette Blythe	Thomas Wolfe & His Family
G55	Ralph E. Flanders	Senator from Vermont,. 112
G56	Keith F. McKean	The Moral Measure of Literature
G57	Robin M. Williams, Jr.	Values & Modern Education in the United States
G58	North Callahan	Daniel Morgan
G59	Esther R. Clifford	A Knight of Great Renown
G60	Gertrude Berg & Cherney Berg	Molly and Me
G61	Donald A. White	Litus Saxonicum
G62	C. H. Cramer	Newton D. Baker
G63	George Steiner	The Death of Tragedy
G64	Mark Eccles	Shakespeare in Warwickshire
G65	Timothy P. Donovan	Henry Adams & Brooks Adams
G66	Van Wyck Brooks	From the Shadow of the Mountain
G67	Mark Schorer	Sinclair Lewis: An American Life
G68	Harris F. Fletcher	The Intellectual Development of John Milton
G69	Mark R. Hillegas	Dystopian Science Fiction
G70	Joseph W. Krutch	If You Don't Mind My Saying So

G71	Joseph Frank	André Malraux: The Image of Man
G72	J W. Fulbright	For a Concert of Free
G73	Carolyn See	The Jazz Musician as Patchen's Hero
G74	John McCormick	The Confessions of Jean Jacques Krim
G75	George Garrett	A Wreath for Garibaldi
H01	U. S Dep't of Commerce	Handbook of Federal Aids to Communities
H02	U. S. Dep't of State	An Act for International Development
H03	U. S. 87th Congress	House Document No. 487
H04	R. I. Legislative Council	State Automobiles & Travel Allowances
H05	R. I. Leglelative Council	Taxing of Movable Tangible Property
H06	R. I. Development Council	Annual Report, 1960
H07	R. I. Legislative Council	linlform Fiscal Year for Municipalities
H08	John A. Notte, Jr.	R. I. Governor's Proclamations
H09	U. S. 87th Congress	Public Laws 295, 300, 347
H10	U. S. Dep't of Defense	Medicine in National Defense
H11	U. S. Dep't of Commerce	1961 Reaearch Highlights, Nat'1 Bureau of Standards
H12	U. S. 87th Congress	Legislation on Foreign Rels
H13	U. S. 87th Congreas	Congressional Record: Extension of Remarks. May 2, 1961
H14	U. S. Dep't of Health, Education & Welfare	Grants-in-Aid and Other Financial Assistance Programs
H15	U. S. Office of Civil and Defence Mobilization	The Family Fallout Shelter
H16	U. S. Reports	Cases AdJudged in the Supreme Court, October Tenm 1960
H17	U. S. Reports	Cases AdJudged in the Supreme Court, October Tenm 1959-60
H18	Dean Rusk	The Department of State
Hl9	Peace Corps	Fact Book
H20	U. S. Dep't of Agriculture	Development Program for the National Forests
H21	Dwight D. Eisenhower	Public Papers, 1960-61
H22	U. S. Dep't of State	U. S. Treatiea and Other International Agreements
H23	U. S. Federal Communications Commiasion	Pederal Communications Commission Reports
H24	U. S. Tresaury Dep't	Your Federal Income Tax
H25	Guggenheim Foundation	Report of the Secretary Gen'1
H26	Anonymous	A Brief Background of Brown & Sharpe
H27A	Robert Leeson	Leesona Corporation President's

		Report
H27B	Leesona Corporation	More Efficient Production for Expanding Textile Markets
H28	Carleton College	Carleton College Bulletin
H29	Sprague Electric Company	Sprague Log
H30	Carnegie Foundation	Annual Report of Year Ending June 30, 1961
J01	Cornell H. Mayer	Radio. Emission of the Moon and Planets
J02	R. C. Binder et al.	1961 Heat Transfer & Fluid Mechanics Institute
J03	Harry H. Hull	Normal Forces & Their Thermodynamic Significance
J04	James A. Ibers et al.	Proton Magnetic Resonance Study
J05	John R. Van Wazer, ed.	Phosphorus and Its Compounds
J06	Francis J. Johnston & John E. Willard	Exchange Reaction Between C12 and CC14
J07	J. F. Vedder	Micrometeorites
J08	LeRoy Fothergill	Biological Warfare
J09	M. Yokayama et al	Chemical & Serological Characteristics
J10	B. J. D. Meeuse	The Story of Pollination
J11	Clifford H Pope	The Ciant Snakes
J12	Richard F McLaughlin et al.	A Study of the Subgross Pulmonary Anatomy
J13	S. Idell Pyle et al.	Onsets, Completions & Spans
J14	Jacob Robbins et al.	The Thyroid-Stimulating Hormone
J15	J. W. C. Hagstrom et. al.	Debilitating Muscular Weakness
Jl6	A. N. Nagaraj & L. M. Black	Localization of Wound-Tumor Virus Antigen
J17	E. Gellhorn	Prolegomena to a Theory of the Emotions
J18	Kenneth Hoffman & Ray Kunze	Linear Algebra
Jl9	Frederick Mosteller et al.	Probability with Statistical Applications
J20	R. P. Jerrard	Inscribed Squares in Plane Curves
J21	C. R. Wylie, Jr.	Line Involutions in S3
J22	Max F. Millikan & Donald L. Blackmer, editors	The Emerging Nations
J23	Joyce O. Hertzler	American Social Institutions
J24	Howard J. Parad	Preventive Casework: Problems & Implications
J25	Sister Claire M. Sawyer	Some Aspects of Fertility of a Tri-Racial Isolate
J26	Frank Lorimer	Demographic Information on Tropical Africa

J27	Dale L. Womble	Functional Marriage Course for the Already Married
J28	William H. Ittelson & Samuel B. Kutash, editors	Perceptual Changes in Psychopathology
J29	Jesse W. Grimes & Wesley Allinsmith	Compulsivity, Anxiety & School Achievement
J30	Raymond J. Corsini	Roleplaying in Business & Industry
J31	Harold Searles	Schizophrenic Communication
J32	Hugh Kelly & Ted Ziehe	Glossary Lookup Made Easy
J33	Ralph Bc Long	The Sentence & Tts Parts
J34	H.A. Cleason	Review of *African Language Studies*
J35	A. T. Kroeber	Semantic Contribution of Lexicostatistics
J36	D. F. Fleming	The Cold War & Its Origins
J37	Douglas Ashford	Elections' in Morocco: Progress or Confusion
J38	Committee for Economic Development	Distressed Areas in a Growing Economy
J39	William O'Connor	Stocks, Wheat & Pharaohs
J40	James J. O'Leary	The outlook for Interest Rates in 1961
J41	Allan J. Braff & Roger F. Miller	Wage-Price Policies Under Public Pressure
J42	Morton A. Kaplan ~ Nicholas Katzenbach	The Political Foundation of Internationa1 Law
J43	Wallace Mendelson	Justices Black & Frankfurter
J44	J. Mitchell Reese, Jr,	Reorganization Transfers
J45	Albert Schreiber et al.	Defense Procurement & Small Business
J46	Irving Perluss	Agricultural Labor Disputes in California 1960
J47	William S. Ragan	Teaching America's Children.
J48	Paul Cooke	Desegregated Education in the Middle-South Region
J49	Robert J. Havighurst	Social-Class Influences on American Education
J50	James C. Bonbright	Principles of Public Utility Rates
J51	Irving L. Horowitz	Philosophy, Science & the Sociology of Knowledge
J52	Brand Blanshard	The Emotive Theory
J53	William S. Haymond	Is Distance an Original Factor in Vision?
J54	Chester G. Starr	The Origins of Greek Civilization 1100-650 B. C
J55	Jim B. Pearson	The Maxwell Land Grant
J56	Edwin L. Bigelow & Nancy H. Otis	Manchester, Vermont, A Pleasant Land

J57	J. H. Hexter	Thomas More: on the Margins of Modernity
J58	John M, Ray	Rhode Island's Reactions to John Brown's Raid
J59	Clement Greenberg	Collage
J60	Robert A. Futterman	The Future of Our Cities
J61	Allyn Cox	Completing & Restoring the Capitol Frescos
J62	Jimmy Ernst	A Letter to Artists of the Soviet Union
J63	John H. Schaar	Escape from Authority, Perspectives of Erich Fromm
J64	Katherine G. McDonald	Figures of Rebellion
J65	Samuel Hynes	The Pattern of Hardy's Poetry
J66	Kenneth Rexroth	Disengagament: The Art of the Beat Generation
J67	William Whallon	The Diction of Beowulf
J68	Charles R. Forker	The Language of Hands in Great Expectations
J69	I. B. M. Corporation	IBM 7070, Autocoder Reference Manual
J70	Ross E. McKinney & Howard Edde	Aerated Lagoon for Suburban Sewage Disposal
J71	Thomas D. McGrath	Submarine Defense
J72	Mellon Institute	Annual Report; 1960, Independent Research
J73	Nat'l Research Council	Directory of Continuing Numerical Data Projects
J74	Harlan W. Nelson	Food Preservation by Ionizing Radiation
J75	W. K. Asbeck	Forces in Coatings Removal Knife Cutting Method
J76	Joel Frados, editor	Survey of Foamed Plastics
J77	William D. Appel, editor	1961 Technical Manual of American Ass'n of Textile Chemists & Colorists
J78	Paul J. Dolon & Wilfrid F. Niklas	Gain & Resolution of Fiber Optic Intensifier
J79	Rutherford Aris	The O'ptim.A1 Design of Chemical Reactors
J80	C. J. Savant Jr. & R. C. Howard	Principles of Inertial Navigation
K01	Christopher Davis	First Family
K02	Clayton C. Barbeau	The Ikon
K03	Tristram Coffin	Not to the Swift
K04	W. E. B. Du Bois	Worlds of Color
K05	David Stacton	The Judges of the Secret Court

K06	Louis Zara	Dark Rider
K07	Francis Pollini	Night
K08	Guy Endore	Voltaire! Voltaire!
K09	Howard Fast	April Morning
K10	Gladys H. Barr	The Master of & Geneva
K11	Robert Penn Warren	Wilderness
K12	Gerald Green	The Heartless Light
Kl3	William Maxwell	The Chateau
K14	Irving Stone	The Agony & the Ecstasy
K15	Ann Hebson	The Lattimer Legend
K16	Stephen Longstreet	Eagles Where I Walk
Kl7	Leon Uris	Mila 8
K18	John Dos Passos	Midcentury
K19	Robert J Duncan	The Voice of Strangers
K20	Guy Bolton	The Olympians
K21	Bruce Palmer	My Brother's Keeper
K22	John Cheever	The Brigadier & the Golf Widow
K23	Prieda Arkin	The Tight of the Sea
K24	W. H. Gass	The Pedersen Kid
K25	Arthur Miller	The Prophecy
K26	Jane G. Rushing	Against the Moon
K27	E. Lucas Myers	The Vindication of Dr. Nestor
K28	Sallie Bingham	Moving Day
K29	Marvin Schiller	The Sheep's in the Meadow.
L01	Winfred Van Atta	Shock Treatment
L02	A. A. Fair	Bachelors Get Lonely
L03	Amber Dean	Encounter With Evil
L04	David Alexander	Bloodstain
L05	Brett Halliday	The Careless Corpse
L06	Thomas B. Dewey	Hunter at Large
L07	Genevieve Golden	Deadlier Than the Male
L08	Dell Shannon	The Ace of Spades
L09	Mignon G. Eberhart	The Cup, the Blade or the Swords
L10	Harry Bleaker	Impact
L11	Hampton Stone	The Man Who Looked Death in the Eye
L12	Whit Masterson	Evil Come, Evil Go
L13	Dolores Hitchens	Footsteps in the Night
Ll4	Frances & Richard Lockridge	Murder Has Its Points
L15	Doris M. Disney	Mrs. Meeker's Money
L16	Alex Gordon	The Cipher
L17	Brent James	Night of the Kill
L18	George H. Coxe	Error of Judgment

L19	Brad Williams	Make a Killing
L20	Ed Lacy	Death by the Numbers
L21	Helen McCloy	The Black Disk
L22	S. L. M. Barlow	Monologue of Murder
L23	J. W. Rose	Try My Sample Murders
L24	Fredric Brown	The Murders
M01	Robert Heinlein	Stranger in a Strange Land
M02	Philip J. Farmer	The Lovers
M03	James Blish	The Star Dwellers
M04	Jim Harmon	The Planet with No Nightmare
M05	Anne McCaffrey	The Ship Who Sang
M06	Cordwainer Smith	A Planet Named Shayol
N0l	Wayne D. Overholser	The Killer Marshall
N02	Clifford Irving	The Valley
N03	Cliff Farrell	The Trail of the Tattered Star
N04	James D. Horan	The Shadow Catcher
N05	Richard Ferber	Bitter Valley
N06	Thomas Anderson	Here Comes Pete Now
N07	Todhunter Ballard	The Night Riders
N08	Mary Savage	Just for Tonight
N09	Jim Thompson	The Transgressors
N10	Joseph Chadwick	No Land Is Free
N11	Gene Caesar	Rifle for Rent
N12	Edwin Booth	Outlaw Town
N13	Martha F. McKeown	Mountains Ahead
N14	Peter Field	Rattlesnake Ridge
N15	Donald J. Plantz	Sweeney Squadron
N16	Ralph J. Salisbury	On the Old Sante Fe Trail to Siberia
N17	Richard S. Prather	The Bawdy Beautiful
N18	Peter Bains	With Women Education Pays off
Nl9	David Jackson	The English Gardens
N20	T. C. McClary	The Flooded Dearest
N21	C. T. Sommers	The Beautiful Mankillers of Eromonga
N22	Gordon Johnson	A Matter of Curiosity.
N23	Wheeler Hall	Always Shoot to Kill
N24	T. K. Brown III	The Fifteenth Station
N25	Wesley Newton	Aid & Comfort to the Enemy
N26	Paul Brock	Toughest Lawman in the Old West
N27	James Hines & James Morris	Just Any Girl
N28	Ralph Grimshaw	Mrs. Hacksaw, New Orleans Society Killer
N29	Harlan Ellison	Riding the Dark Train Out

P01	Octavia Waldo	A Cup of the Sun
P02	Ann Ritner	Seize a Nettle
P03	Clark McMeekin	The Fairbrothers
P04	B. J. Chute	The Moon & the Thorn
P05	Allan R. Bosworth	The Crows of Edwina Hill
P06	Richard Tiernan	Land of the Silver Dollar
P07	Vina Delmar	The Big Family
P08	R. Leslie Course	With Gall & Honey
P09	Jesse Hill Ford	Mountains of Gilead
P10	Jay Williams	The Forger
P11	Bessie Breuer	Take Care of My Roses
P12	Morley Callaghan	A Passion in Rome
P13	Frank B. Hanes	The Fleet Rabble
P14	Livingston Biddle, Jr.	Sam Bentley's Island
P15	Loretta Burrough	The Open Door
P16	Margery F. Brown	A Secret Between Friends
P17	Al Hine	The Huntress
P18	Anonymous	No Room in My Heart to For Give
Pl9	Anonymous	This Cancer Victim May Ruin My Life
P20	Spencer Norris	Dirty Dog Inn
P21	Elizabeth Spencer	The White Azalea
P22	Anonymous	A Husband Stealer from Way Back
P23	Barbara Robinson	Something Very Much in Common
P24	Samuel Elkin	The Ball Player
P25	William Butler	The Pool at Ryusenji
P26	Ervin D. Krause	The Snake
P27	Lee McGiffin	Measure of a Man
P28	Carol Hoover	The Shorts on the Bedroom Floor
P29	Robert Carson	My Hero
R01	Anita Loos	No Mother to Guide Her
R02	Jean Mercier	Whatever You Do, Don't Panic
R03	Patrick Dennis	Little Me
R04	Edward Streeter	The Chairman of the Bored
R05	Evan Esar	Humorous English
R06	James Thurber	The Future, If Any, of Comedy
R07	John H. Wildman	Take It Off
R08A	Leo Lemon	Catch Up With
R08B	Leo Lemon	Something to Talk About
R09	S. J. Perelman	The Rising Gorge