Universidade Federal de Pernambuco
Centro de Informática

Graduação em Engenharia da Computação

An Unsupervised Multi-view Approach for Authorship Attribution of Early Modern Plays

Trabalho de Graduação

Autora: Andréa Brandão Duque
Orientador: Renato Vimieiro
Área: Mineração de Dados

Recife
Dezembro de 2017

To my mother and father, for supporting my education in joy and misfortune. To my grandmother, who is now in heaven watching down on me.

Acknowledgements

I am grateful to my advisor Prof. Renato Vimieiro for his guidance and patience, for being a genuine teacher.

I would like to thank Prof. Edna Barros and the other excellent professors from the Center of Informatics, for being essential in my formation as an engineer and individual.

Finally, I acknowledge the remarkable people I worked and studied with this final graduation year, as I wouldn't be able to finish my studies alone.

And seeing ignorance is the curse of God,
Knowledge the wing wherewith we fly to heaven.

—WILLIAM SHAKESPEARE (Henry VI, Part 2)

Resumo

Muitas obras de literatura inglesa do período Early Modern English, que ocorreu entre 1500 a 1700, tem a autoria disputada ou desconhecida. É um desafio atribuir um estilo de escrita para cada escritor do período pois seu estilo muda ao longo da carreira. Além disso, autores às vezes trabalharam juntos. Outra dificuldade está no fato de existirem diferentes características estilométricas que podem ser extraídas de um texto. Portanto, métodos que consideram múltiplas visões dos dados do problema podem melhorar a tarefa de atribuição da autoria, onde cada visão é um conjunto de atributos diferente. Este trabalho investiga diferentes características estilométricas extraídas de obras históricas de literatura inglesa. Nós exploramos afinidades de estilo de escrita entre as obras e a autoria de obras anônimas considerando múltiplas visões dos dados textuais.

Palavras-chave: Atribuição de Autores, Métodos Não supervisionados, Múltiplas Visões, Literatura Inglesa

Abstract

Many literary works from the Early Modern English period, which dates from 1500 to 1700, have disputed or unknown authorship. It is a challenge to assign a writing style to each writer from the period because the style changed through time. Furthermore, authors sometimes worked together. There are also many stylometric features which can be extracted from a text. Therefore, methods which can consider the multiple views of the problem data might perform the authorship attribution task better. This work investigates different stylometric characteristics extracted from literary works. We explore authorship affinities between works and authorship of anonymous plays considering multiple views from the text data.

Keywords: Authorship Attribution, Unsupervised Methods, Multi-view, Early Modern Plays

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Many literary works from the Early Modern English period [25], which dates from 1500 to 1700, have disputed or unknown authorship. The task of authorship attribution of documents from this era is a difficult task for several reasons. Authors sometimes worked together, collaborating in writing. Sometimes there was no collaboration, but they copied texts from each other. Also, as authors matured, their style changed [22]. Thus, an author writing style is difficult to define. The question of whether a particular book is a collaboration and which authors collaborated in the book is a matter of discussion throughout the years.

Computer-based authorship attribution methods have arisen to assist in the literary discussion, given an improvement in available statistical and computational techniques [4, 16]. The task of classifying text based on the author also has diverse applications besides addressing literary issues, such as forensic applications and plagiarism detection [26]. The authorship attribution field has become more relevant with the growth of the web and consequently the number of available texts.

Research in the area initially appeared with the attempt to define features for quantifying writing style, considering aspects from the text such as lexical, syntactic and semantic aspects. This research field is called stylometry [26]. The most traditional feature used to quantify author style is the lexical feature based on the most frequent words in the text. Word n-grams have also been proposed in an attempt to consider word order. Part-of-speech (POS) and POS n-grams are used to extract syntactic features. POS taggers assign a tag of morpho-syntactic information to words from the text. However, POS tags fail to provide deep structural analysis [26].

More recently, studies show that human language can be modeled as complex networks [7, 28]. These networks are graphs in which words are the nodes, and adjacent words are connected by an edge. Networks measurements have been used to model writing styles [3, 2] and in authorship attribution tasks [18]. These graphs have also been previously used to investigate authorship attribution in Early Modern Plays [10].

It is a challenge to select which features to use in authorship attribution task since writing style depends on many aspects. Writing styles may be more complicated than the traditional text representation techniques, and the use of multiple features may produce better results [16].

Arefin et al. [4] investigated the authorship problem using word frequencies as features and an unsupervised clustering method. Word frequencies were extracted from the literary documents and used to compute a distance matrix which is the input of their clustering method. The choice of clustering allowed the exploration of how the plays relate to each other. Works have a close relationship in the clusters results mainly because of authorship similarity, but also by topic similarity.

In this work, we study the combination of word frequencies, POS-tag and network measures employed in the authorship attribution task of selected plays from the Early Modern English Period. We investigate the hypothesis that multiple views shed the light of different perspectives of an author's writing style. Our goal is to identify more homogeneous groups of texts, which in turn facilitates the identification of the authorship of disputed or unknown works.

Initially, we introduce necessary concepts to the understanding of this work and review previous works similar to ours (Chapter 2). We show how we built a corpus of Early Modern works, containing plays of sole, unknown and disputed authorship and how we extracted from these plays different features (Chapter 3). We investigate clusters composition utilizing each feature alone and compare to the results using multi-view clustering algorithms, combining the features extracted (Chapter 4, Section 4.1). Then, we use a multi-view clustering algorithm to cluster our plays in groups of same writing style. Obtaining these results, we analyze authorial and genre affinities in our clusters, also performing the authorship attribution task of disputed and unknown works. (Chapter 4, Section 4.2).

CHAPTER 2

Background and Related Works

In this chapter, we introduce the area of our research. We review necessary concepts to comprehend this work and present relevant previous works.

2.1 Text Mining

Text mining is a cross-disciplinary research field. This area benefits from advances in fields such as data mining, machine learning, natural language processing (NLP) and information retrieval. The interest in the area is growing in recent years with advances in web-based systems. Overall, the amount of text data available online is increasing along with the desire to understand it.

The primary purpose of text mining is to analyze textual information to discover patterns, which may be trends or outliers, and assist in decision making. In Figure 2.1, we present a general text mining flow that we adopt in this work consisting of three phases: preprocessing, representation and analysis. Given a text corpus, which in our work is a collection of Early Modern plays, we preprocess text data to make the natural language documents more consistent. Thereby, it is possible to build better text representation models, which are input to data analysis algorithms [12, 1].

In this work, we focus on the authorship attribution task. Authorship attribution is the task of identifying the author of a particular textual document. This problem is perceived as a matter of attribute extraction and data mining [16]. To assign the authorship of documents, we convert the text to numerical attributes which can represent the writing style of the authors, such as words frequencies. After the extraction of the numerical features, it is often applied supervised or unsupervised machine learning techniques for the attribution analysis.

2.1.1 Text Preprocessing

The first preprocessing task usually applied is *tokenization*. In this task, a stream of text data is split into smaller units, named tokens, while discarding punctuation and other characters. These units can be sentences, paragraphs or words. In this work, we consider a token is a word.

A common problem that occurs while analyzing historical texts is word spelling variety. English texts from the sixteenth century displayed a great deal of word spelling variation, and the establishment of the English language to its modern spelling would only happen further in history. Therefore, a *standardization* step is necessary when there is orthographic variety within a document collection to be analyzed. This task maps variant spelling to a standard

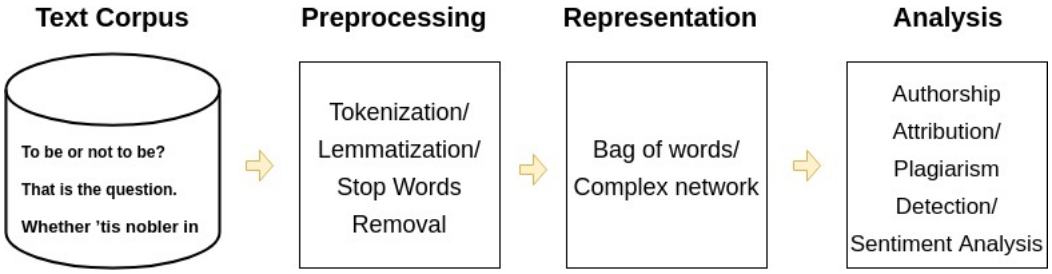

Figure 2.1 Example of a generic text mining process, consisting of three phases applied consecutively to a text corpus (i.e. collection of documents). The first phase is preprocessing. This phase comprises tasks such as *tokenization, lemmatization and stop words removal* that can be applied to data to build a more concise representation. Then, it is possible to discover knowledge from the data, performing authorship analysis, sentiment analysis, and several other applications.

form.

Another problem when dealing with text data is that there may be a lot of unique words in the corpus. This problem increases complexity in calculating some representational models, such as networks, and consequently demands more computational resources. Therefore, reducing the feature size is possible with a *lemmatization* task. The goal of the *lemmatization* process is to reduce inflectional endings of a word to a common base form or dictionary form of a word, which is known as the lemma. A lemmatizer does this process analyzing words morphologically.

It is also possible to reduce the word space with *stop words removal*. Stop words are words which often appear in a document collection. The argument to remove stopwords is that they do not distinguish each document well enough because they appear in all the collection.

We also can use preprocessing tasks to extract other types of information besides words. It is possible to extract *Part of Speech (POS)-tags* to capture morpho-syntactic information. These tags classify words from the text to simple tags such verbs, adverbs, preposition, articles, nouns [17].

2.1.2 Text Representation Models

We build a representational model from a corpus transforming each document into a numeric vector, thus obtaining a matrix. This matrix is the set of features input to data analysis algorithms, as we already mentioned in the previous section. The extraction of features from text which can represent the semantics is a challenge in Text Mining because it affects the quality of the analyzes [1]. Given the potentially large number of words and phrases that can be extracted from the text, a document has different levels of representation [12, 1]. Therefore, it is difficult to select the best representation for each document. For example, lexical features or syntactic features may represent a document. A traditional model for representing lexical features is the bag-of-words model, where the word order is disregarded. Then, we can represent each doc-

ument as a vector of its word frequencies, where each position is the number of occurrences for each word. Considering syntactic features, part-of-speech (POS) tags can be extracted for each word, and POS-tag frequencies vector can be calculated. In Figure 2.2, we demonstrate an example of these representations.

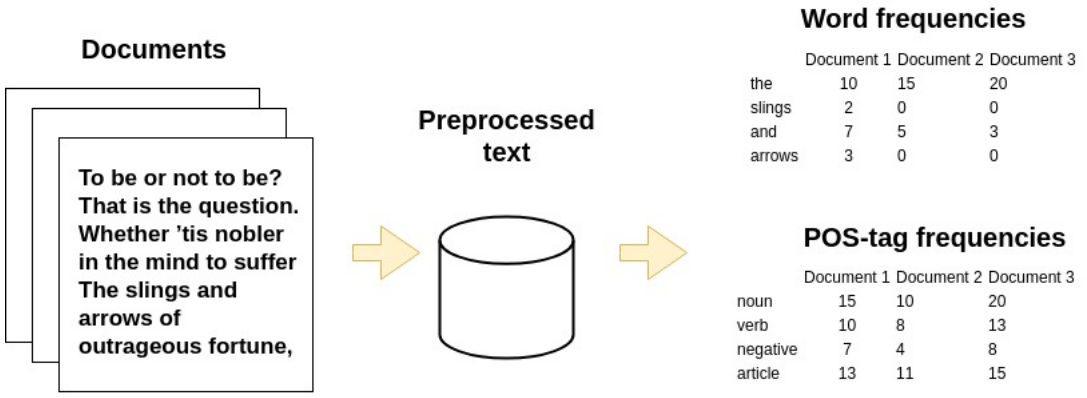

Figure 2.2 Example of how representational matrices can be built for a corpus, after the preprocessing task, from words or POS-tags. Each position (i, j) of the matrix is the frequency of term i in document j. For a corpus with D documents and m unique terms, the matrix will have $m \times D$ dimensions

A more recent model to represent human language is the complex network model [7]. These networks are graphs with attributes that are found neither in regular networks neither random networks. The advantage of modeling documents as a complex network is performing better structural analysis than with traditional representations.

One of the network models that can represent text documents and from which stylistic attributes can be extracted is the co-occurrence or adjacency network model [7, 3]. In this model, each different word is a node of the graph, and adjacent words are connected by an edge. Other characters such as punctuation are not considered to build the network. In Figure 2.3, an example of co-occurrence network is observed.

Many attributes from complex networks have been extracted and employed to perform the authorship attribution task. Amancio [2] used traditional stylistic attributes such as word frequency and topology measures from word adjacency networks such as degrees (the number of edges connected to a node) and betweenness, a measure of centrality considering a node is relevant if it is much accessed by shortest paths. Amancio [2] achieved the best results by combining traditional and network attributes in a hybrid approach. Three-node motifs frequencies were also used as features to identify authors [18]. Network motifs are recurrent sub-graphs in complex networks. In this approach, each document is described as a vector of thirteen positions, each value is the frequency of a motif from the complex network model of the document. In Figure 2.4 we illustrate the thirteen possible motifs. A variation of the motifs model is labeled motifs. These motifs were also used to identify authorship, in an attempt to combine word frequencies and three-node motifs. [19].

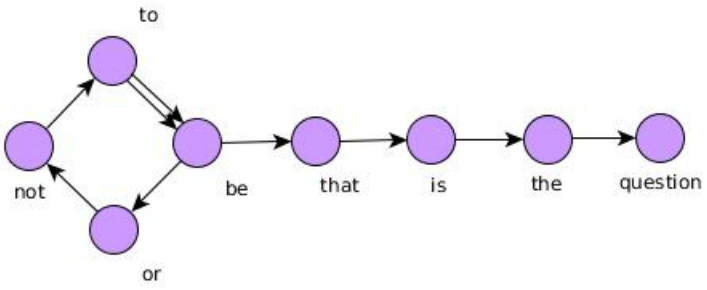

Figure 2.3 Example of co-occurrence or adjacency network for text extract "To be or not to be? That is the question." Each word from the phrase is a node and adjacent words are connected by an edge. The edge direction is the natural reading order. Note that punctuation and capital letters were not considered.

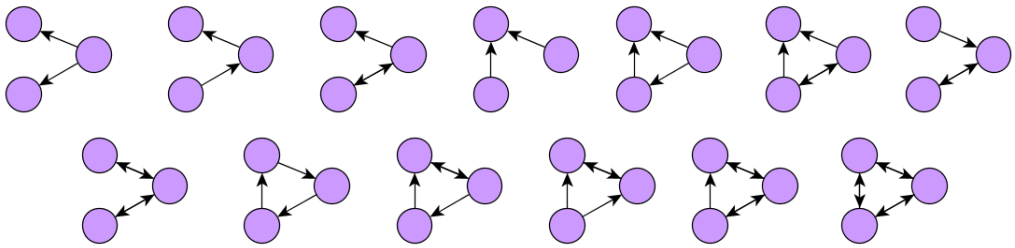

Figure 2.4 Motifs are isomorphic subgraphs which can be found in complex networks. There are thirteen three-node connected motifs.

For each document vector described by labeled motifs, each entry n of the vector is calculated by

$$n = \sum_{m=1}^{13} \frac{n_{i,m}}{n_m} \tag{2.1}$$

where n_m is the total number of occurrences of motif m and $n_{i,m}$ the total number of occurrences of term i in motif m. The term i is from the set of most frequent words from the corpus. For example, in the matrix represented in Figure 2.5, each entry is calculated using Equation 2.1 for a particular term and document. Therefore, considering the 100 most frequent words from the corpus, it is possible to build a matrix $100 \times D$ for a corpus with D documents, where each column represents a feature vector for a particular document.

2.1.3 Unsupervised Authorship Analysis

It is not necessary to know the data labels previously to perform unsupervised analysis, which is an advantage for data exploration tasks. One of the most popular unsupervised methods is clustering [1]. Clustering algorithms find homogeneous groups in the data. Thus, it is expected that the works of the same author, or authors who have a similar writing style, are grouped in

Labeled motifs frequencies

	Document 1	Document 2	Document 3
the	n11	n21	n31
slings	n12	n22	n32
and	n13	n23	n33
arrows	n14	n24	n34

Figure 2.5 In this feature matrix, each column is a document vector, and each line is a different word from the corpus. Each position n of a document vector is the sum of proportions that a word appears in each of the thirteen motifs.

the same cluster. These groupings allow inferring the authorship of unknown or disputed works since the specialist will have a much smaller scope of authors to make the decision.

We focus on distance-based clustering algorithms, which receive as input a dissimilarity matrix. A similarity function computes the distance between documents utilizing text characteristics such as a word frequencies matrix. Clustering algorithms group the data according to the similarity measure. Books from the same cluster should be similar to each other, and clusters should be distinct from each other.

In this work, we utilize single view hierarchical and hard partitioning clustering methods, which are the most popular techniques and multi-view hard partitioning methods. Single view methods receive as input one dissimilarity matrix, multi-view methods receive multiple matrices.

Partitioning methods construct a fixed number of partitions given an input parameter k. Each partition optimizes an objective function. Partitioning methods can be divided into hard clustering and fuzzy clustering [29]. Hard clustering assigns each data point to exactly one cluster. In fuzzy clustering, each data point can belong to more than one group.

A simple partition clustering is k-means. The goal is to calculate a centroid (i.e. means) for each cluster based on the data points and assign data points to clusters based on the distance to the centroid. This algorithm is sensitive to noisy data and outliers, so more robust strategies were developed, such as k-medoids. This variation uses medoids instead of the mean, which are representative objects for each cluster. The goal is to minimize the sum of distances of objects within a group, for arbitrary distance functions. The most common k-medoids clustering method is the PAM algorithm (Partitioning Around Medoids). PAM uses a greedy search to optimize the partitions.

Hierarchical clustering generates a nested sequence of partitions of the input data. These methods can be agglomerative (bottom-up) or divisive (top-down). Agglomerative methods start in a configuration where each document is a cluster and successively agglomerates pairs of clusters until all elements are connected in a hierarchy. A divisive method begins with all items in a single group and performs a splitting procedure until a stopping criterion is met. Ward's method is a standard criterion applied to agglomerative hierarchical clustering. At each step, the algorithm finds the pair of clusters that leads to a minimum increase in total within-

cluster variance after merging.

Another type of hierarchical clustering is graph-based methods. A weighted graph can be built from a dissimilarity matrix, where each work from a corpus is a node in the graph. The edge weights correspond to a pair-wise dissimilarity between two documents. Graph-based clustering methods calculate subgraphs from the complete corpus graph, removing edges according to a criterion and the final graph is the result of clustering output. A state-of-the-art graph technique utilized previously for authorship attribution is the MST-kNN [4]. This clustering method calculates a minimum spanning tree (MST) and a k-nearest neighbor graph (kNN) for the corpus graph and recursively removes edges from MST, which for each edge, both nodes do not share any of the k nearest neighbors. The number k of neighbors is given by $k = \lfloor \ln n \rfloor$, where n is the number of nodes in each tree of the MST. Initially, n is equal to the number of works in the corpus, but as the algorithm continues recursively, inspecting the subtrees, k is adjusted. Therefore, MST-kNN is non-parametric because the number of clusters is automatically determined.

The multi-view clustering methods we utilize to group our plays are dynamic hard clustering algorithms based on multiple dissimilarity matrices (MRDCA) [9]. This algorithm computes best prototypes for each partition such that it optimizes an objective function measuring the fit between the clusters and their prototypes. The number of prototypes and number of partitions are parameters to be chosen by the user. There are two variations of this algorithm, MRDCA-RWG and MRDCA-RWL. The former estimates relevance weight for each dissimilarity matrix globally (i.e. for all clusters) and the other locally (i.e. for each cluster).

In the next two subsections, we show how to calculate dissimilarity matrices for feature data extracted from a corpus and then we discuss cluster metrics we utilized to evaluate results.

2.1.3.1 Dissimilarity matrices

From a feature matrix $m \times n$ extracted from the text documents of m features and n documents, we can compute a dissimilarity matrix $n \times n$ containing the distances, taken pairwise, between each document. The similarity functions we use in this work to calculate distance matrices are the Jessen-Shannon divergence (JSD) and the cosine distance. We chose these metrics because the cosine similarity function is one of the most used to calculate similarities in the text domain and the JSD metric was previously shown to unveil authorship affinities [4].

The JSD metric measures the dissimilarity between two probability distributions P and Q. Arefin et al. [4] have shown that documents can be interpreted as distributions of probability of occurrence of terms, and therefore can be compared by this metric. JSD can be computed as follows:

$$JSD(P,Q) = \sqrt{H\left(\frac{P+Q}{2}\right) - \frac{H(P)+H(Q)}{2}} \tag{2.2}$$

where, H(X) is Shannon information entropy for probability distribution X and is calculated as

$$H(X) = -\sum_{x_i \varepsilon X} x_i \log_2 x_i \tag{2.3}$$

x_i is the probability of a text feature i in document X, which in our case can be lemmas, POS-tags or a network measurement.

Another interpretation for documents is to consider each document a vector in a normalized vector space. Thus, it is possible to measure the similarity between a pair of documents calculating the cosine of the angle between two vectors. This calculation measures the difference in orientation of both vectors. Therefore, the cosine distance can be calculated as follows:

$$COS(A,B) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} \tag{2.4}$$

A and B are two numerical vectors, each vector representing a document.

2.1.3.2 Cluster metrics

We use two well-known clustering similarity measures to evaluate the quality of our clusters: Normalized Mutual Information (NMI) [27] and Adjusted Rand Index (ARI) [14]. These two measures evaluate the similarity between the true labels of a ground-truth dataset and the labels assigned by the clustering algorithm to be qualified.

Mutual Information (MI) measures the reduction in the entropy of class labels given the cluster labels. It is possible to compare cluster results with MI if it is normalized. For two different cluster solutions A and B, we normalize mutual information (MI) to compare both solutions as follows:

$$NMI_{sqrt} = \frac{I(A;B)}{\sqrt{H(A)H(B)}} \tag{2.5}$$

Where $I(X,Y)$ is MI measure and $H(X)$ is Shannon information entropy defined in equation 2.3 for a cluster solution distribution. For this metric, 0 is no mutual information and 1 is perfect agreement.

ARI score is a variation of the Rand Index (RI) in which it is adjusted for the chance grouping of elements. RI represents the frequency of occurrence of agreements over class label and cluster solution. The similarity score of the ARI metric is between -1.0 and 1.0.

2.2 Related Works

Clustering algorithms were previously used to model the writing style. Arefin et. al [4] used cluster analysis to provide a practical guide to the authorship of disputed Early Modern Plays. A distance matrix was computed from word frequencies using an information theoretic measure, the Jensen-Shannon divergence. This matrix is input to a clustering algorithm based on a graph partitioning algorithm, MST-kNN.

More recently, Naeni et. al [23] also utilized a graph-based clustering to identify groups of documents which reflect authorial affinities in a corpus of Early Modern Plays, showing remarkable results in comparison to other state-of-art clustering algorithms.

Outside the literary domain, Iqbal et. al [15] used clustering to identify authors of criminal e-mails. Their proposed method first clusters anonymous e-mails based on the text features such as word frequencies and then extracts writing styles from each resultant cluster utilizing frequent patterns.

CHAPTER 3

Materials and Methods

3.1 Text Corpus

We utilized a text corpus containing 230 plays from the Early Modern English period. We selected works from the Shakespeare His Contemporaries Corpus [21] and the Folger Shakespeare Corpus [20]. The plays creation date range from 1583 to 1659. From the 230 plays, 218 are considered texts of sole authorship (i.e. undisputed), 8 are plays of anonymous authorship, and four plays are collaborations which have uncertain authors participation. The authorship information, creation date and genre from the plays was compiled from the Database of Early English Playbooks (DEEP) [11]. In Table 3.1, we describe the authors in the corpus and the number of individual plays of each author.

Author	**Number of Sole Plays**
Shirley, James	31
Shakespeare, William	29
Heywood, Thomas	19
Chapman, George	14
Fletcher, John	14
Jonson, Ben	14
Middleton, Thomas	14
Brome, Richard	13
Massinger, Philip	13
Dekker, Thomas	9
Ford, John	9
Lyly, John	8
Marston, John	8
Greene, Robert	6
Marlowe, Christopher	6
Peele, George	6
Rowley, William	5

Table 3.1 List of authors and the number of plays attributed to them from the 218 plays of sole authorship present in the corpus.

In Table 3.2, we present the anonymous and uncertain plays present in our corpus, of which we investigate the authorship previously by other authorship studies. [25, 4, 22, 10, 8, 5, 13].

Play	Category	Possible Authors
Arden of Faversham	Anonymous	Kyd, Thomas Marlowe, Christopher Shakespeare, William
Fair Em	Anonymous	Shakespeare William Wilson, Robert
King Leir and his Three Daughters	Anonymous	Kyd, Thomas Greene, Robert Peele, George Lodge, Thomas Munday, Anthony Shakespeare, William
A Knack to Know a Knave	Anonymous	Greene, Robert Lodge, Thomas
1 Troublesome Reign of King John	Anonymous	Marlowe, Christopher Greene, Robert Lodge, Thomas Peele, George
2 Troublesome Reign of King John	Anonymous	Marlowe, Christopher Greene, Robert Lodge, Thomas Peele, George
The Wars of Cyrus	Anonymous	Marlowe, Christopher Shakespeare, William
The Reign of King Edward the Third	Anonymous	Shakespeare, William Marlowe, Christopher
Eastward Ho	Uncertain	Chapman, George Jonson, Ben Marston, John
The Spanish Gypsy	Uncertain	Dekker, Thomas Ford, John Middleton, Thomas Rowley, William
The Bloody Brother (Rollo, Duke of Normandy)	Uncertain	Fletcher, John Massinger, Philip Chapman, George Field, Nathan Jonson, Ben
The Fair Maid of the Inn	Uncertain	Fletcher, John Ford, John Massinger, Philip Webster, John

Table 3.2 Anonymous and uncertain plays present in the corpus and the writers which have been proposed as authors solely or collaboratively

A natural language processing software MorphAdorner [6] preprocessed and annotated each play from the corpus with morphological information, such as part of speech and lemma. The software is specialized in processing Early Modern Plays, solving the problem of the spelling variation of the historical texts by performing standardization. In Table 3.1, we show spelling variation of "advance". After the preprocessing tasks, the corpus works are stored in annotated XML documents. We display in Figure 3.1 an excerpt of *Hamlet*, written by William Shakespeare in the XML version. We extract lemmas and POS-tags from plays by parsing these XML documents.

aduauce	advance
aduauced	advanced
aduauceing	advancing
aduaucement	advancement
aduauceth	advanceth
aduaucing	advancing
aduaucyng	advancing
aduaucynge	advancing
aduaunc'd	advanced

Table 3.3 The first column is the variant spelling and the second column is after the standardization preprocesing task. Source: [6]

```
<w xml:id="fs-ham-0271850" n="3.1.64" lemma="to" ana="#acp-cs">To</w>
<c> </c>
<w xml:id="fs-ham-0271870" n="3.1.64" lemma="be" ana="#vvi">be</w>
<c> </c>
<w xml:id="fs-ham-0271890" n="3.1.64" lemma="or" ana="#cc">or</w>
<c> </c>
<w xml:id="fs-ham-0271910" n="3.1.64" lemma="not" ana="#xx">not</w>
<c> </c>
<w xml:id="fs-ham-0271930" n="3.1.64" lemma="to" ana="#acp-cs">to</w>
<c> </c>
<w xml:id="fs-ham-0271950" n="3.1.64" lemma="be" ana="#vvi">be</w>
<pc xml:id="fs-ham-0271960" n="3.1.64">-</pc>
<w xml:id="fs-ham-0271970" n="3.1.64" lemma="that" ana="#d">that</w>
<c> </c>
<w xml:id="fs-ham-0271990" n="3.1.64" lemma="be" ana="#vvz">is</w>
<c> </c>
<w xml:id="fs-ham-0272010" n="3.1.64" lemma="the" ana="#d">the</w>
<c> </c>
<w xml:id="fs-ham-0272030" n="3.1.64" lemma="question" ana="#n1">question</w>
```

Figure 3.1 Text extracted from Hamlet in the XML annotated version. The attribute *lemma* from the XML represents the lemmatized version of the token and the attribute *ana* the POS-tag.

The original text tokens appear in the natural reading order in the XML. Some tags enclose the tokens. The tag <w> encloses the original token text, <pc> encloses punctuation and <c>

encloses whitespace. We retrieved the XML attributes "ana" and "lemma" from the <w> tags. They represent POS-tags and lemmas, respectively. We also computed features by removing stop words. In Table 3.4, we display an original text extract and the features computed from the text.

Original Text	To be, or not to be? That is the question
Lemma	to be or not to be that be the question
Pos-Tag	#acp-cs #vvi #cc #xx #acp-cs #vvi #d #vvz #d #n1
Lemma Without Stopwords	question

Table 3.4 Original text extract from *Hamlet* and features extracted from the text.

We show some of the POS-tags retrieved from the XML and their meanings in Table 3.5. These tags are predefined by the NLP software and tag the words with syntatic information. The full set of tags and their meanings and can be found in the software documentation [6].

POS-tag	Classification	Token
acp-cs	subordinating conjunction	to
vvi	infinitive, verb	be
cc	coordinating conjunction	or
xx	negative	not
d	determiner	that
vvz	3rd singular present, verb	is
n1	singular, noun	question

Table 3.5 POS tags and their respective morpho-syntactic information.

3.2 Feature Extraction

We convert lemmas, POS-tags, and lemmas without stop words, extracted from XML documents to numerical values. These values are feature data we later convert to relational data, calculating dissimilarity matrices. We obtain different values from the text features depending on the text representation model chosen, as we described in Chapter 2. We considered two characterizations of the documents: a more traditional representation, where each document is a vector of frequencies of text items extracted; a complex network model, where each document is a complex network, upon which we computed motifs frequencies and labeled motifs frequencies.

We obtain eight different feature data in total:

- *Traditional* model - three different matrices are computed using this model. Considering i an item from the set of lemmas, POS-tags or lemmas without stopwords, each different feature data extracted is a matrix which an entry (d,i) is the number of occurrences of item i in document d.

- *Complex network* model - we obtain 5 different matrices considering this model. Each play is modeled as three different word adjacency networks. A network is built from a set of lemmas, another from a set of POS-tags and other from a set of lemmas without stopwords.

    1. We calculate a matrix for each of the three networks, representing each play (i.e. row) as a vector of 13 connected three-node motif frequencies, obtaining 3 matrices.
    2. We calculate a matrix for each of the lemma-based network and POS-tag network. Each play is a vector where each position of the vector is the sum of labeled motifs frequencies for a frequent item. We use the 100 most frequent items, POS-tag or lemma to calculate the matrices. This number was the maximum number of items we could use without increasing calculation time too much. We do not remove stop words using this type of feature, because we utilize the most frequent items in the calculation.

We compute several distance matrices, i. e. views, from the feature data we extracted from the corpus. These views represent different relational data between plays in the corpus. Thereby, we can cluster the plays and analyze authorial affinities utilizing the matrices. A view from the corpus is a dissimilarity matrix in which each entry is a pairwise distance between two plays. We computed 16 views for the corpus. We used two distance metrics, *Jessen-Shannon Divergence* (JS) and *cosine* distance, applied on the eight different features extracted previously. In the Table 3.6, we list the views calculated and properties.

View	Distance metric	Feature set
jsLemma cosLemma	JSD Cosine	Lemma
jsLemmaStop cosLemmaStop	JSD Cosine	Lemma without stopwords
jsLemmaStopMotifs cosLemmaStopMotifs	JSD Cosine	Lemma without stopwords motifs
jsLemmaMotifs cosLemmaMotifs	JSD Cosine	Lemma motifs
jsLemmaLabeled cosLemmaLabeled	JSD Cosine	Lemma labeled motifs
jsPostag cosPostag	JSD Cosine	POS-tag
jsPostagMotifs cosPostagMotifs	JSD Cosine	POS-tag motifs
jsPostagLabelled cosPostagLabelled	JSD Cosine	POS-tag labelled motifs

Table 3.6 Distance metric and feature data used for each view extraction

CHAPTER 4

Experiments and Discussion

This chapter is divided in two sections. The first section is the Ground Truth Experiments and the other is Authorship Attribution Analysis.

Since we extract several views from different corpus representations, built using different feature combinations, we do not know which views might represent well writing style similarity between two works. Therefore, we first compute experiments using only the historically non-disputed plays, which we call ground truth experiments because each play has a single author. We remove the uncertain and anonymous plays from the corpus, compute the views and cluster the plays with state-of-the-art algorithms, extracting cluster metrics, analyzing the results and discussing the performance of the views individually. Then, we utilize multi-view clustering algorithms and compare the results of both single and multi-view clusters.

In the last section, we use the multi-view cluster on the whole corpus and analyse authorial and genre affinities in our clusters.

4.1 Ground Truth Experiments

In this section, we initially evaluate each view computed from the 218 plays of sole authorship using single view clustering algorithms. We utilize MST-kNN, previously used in authorship attribution tasks and two classical clustering algorithms, PAM and Ward.

MST-kNN chooses automatically the number of clusters. For PAM and Ward we chose the number of clusters equal 17, which are the number of authors in the reduced corpus.

Then, we avaliate if the views combined using multi-view clustering algorithms, MRDCA, MRDCA-RWG and MRDCA-RWL can cluster plays by authorial affinities. We chose ARI and NMI metrics to compare the clusters.

4.1.0.0.1 Single View Cluster Results The results are summarized in the Table 4.1. The best results are for views using JSD as a distance metric. Good results were obtained for views with labeled motifs, as good as with item frequencies, but the results with simple motifs were close to random results. The feature POS-tag also presented good performance. The best results for each clustering algorithm is in bold.

4.1.0.0.2 Multi-View Cluster Results We chose the views jsLemmaLabelled, jsPostag, jsLemmaStop and jsPostagLabelled to use in multi-view clustering because they had the best performance amongst all views, i.e. the concordance between clusters and authors were the highest in our experiments. We utilize the same parameters in all multi-view clustering algorithms,

View	Algorithm	Number of Clusters	ARI	NMI
jsLemmaStop	MST-kNN	8	**0.26**	0.57
	PAM	17	0.26	0.57
	Ward	17	0.53	0.74
jsPostagLabelled	MST-kNN	8	**0.26**	0.54
	PAM	17	0.43	0.60
	Ward	17	0.50	0.65
jsLemma	MST-kNN	9	0.23	**0.59**
	PAM	17	0.25	0.55
	Ward	17	**0.65**	**0.78**
jsPostag	MST-kNN	6	0.21	0.59
	PAM	17	**0.50**	**0.67**
	Ward	17	0.58	0.73
cosPostag	MST-kNN	13	0.21	0.56
	PAM	17	0.41	0.63
	Ward	17	0.40	0.63
jsLemmaStopMotifs	MST-kNN	10	0.13	0.34
	PAM	17	0.15	0.42
	Ward	17	0.16	0.42
jsLemmaMotifs	MST-kNN	10	0.13	0.34
	PAM	17	0.15	0.42
	Ward	17	0.16	0.42
cosLemma	MST-kNN	5	0.12	0.44
	PAM	17	0.23	0.48
	Ward	17	0.36	0.61
cosPostagLabelled	MST-kNN	5	0.07	0.43
	PAM	17	0.24	0.51
	Ward	17	0.48	0.61
cosPostagMotifs	MST-kNN	14	0.09	0.33
	PAM	17	0.08	0.35
	Ward	17	0.10	0.36
cosLemmaMotifs	MST-kNN	15	0.08	0.33
	PAM	17	0.13	0.41
	Ward	17	0.20	0.44
cosLemmaStopMotifs	MST-kNN	15	0.08	0.33
	PAM	17	0.13	0.41
	Ward	17	0.20	0.44
jsLemmaLabelled	MST-kNN	2	0.02	0.25
	PAM	17	0.39	0.63
	Ward	17	0.53	0.70
cosLemmaStop	MST-kNN	2	0.01	0.16
	PAM	17	0.03	0.25
	Ward	17	0.29	0.55
cosLemmaLabelled	MST-kNN	4	0.00	0.18
	PAM	17	0.27	0.55
	Ward	17	0.27	0.54
jsPostagMotifs	MST-kNN	4	-0.01	0.14
	PAM	17	0.12	0.36
	Ward	17	0.11	0.38

Table 4.1 Cluster results for each view extracted.

MRDCA, MRDCA-RWG and MRDCA-RWL. We chose the number of clusters parameter equals 17, and three prototypes to be chosen for each cluster, since the author with less plays in the reduced corpus has five plays. We show the results for our experiments in Table 4.2. In Table 4.3, we show the weights the MRDC-RWG computed for each view globally to calculate the clusters. The most relevant view according to the weights was the jsPostag.

Algorithm	K	ARI	NMI	Number Prototypes
MRDCA	17	**0.63** (0.05)	**0.85** (0.02)	3
MRDCA-RWG	17	0.60 (0.05)	0.85 (0.02)	3
MRDCA-RWL	17	0.59 (0.09)	0.83 (0.03)	3

Table 4.2 Results for multi-view clusters for ground truth dataset. K is the number of clusters. Ten experiments were executed, the results are the mean for ARI and NMI and standard deviation (in parenthesis). The best result is denoted in bold.

View	Weights
jsPostag	2.007728
jsPostagLabelled	1.460023
jsLemmaLabelled	1.127170
jsLemmaStop	0.302654

Table 4.3 The weights for each view MRDC-RWG computed in a experiment. Each weight view is global to all the partitions.

The algorithm with the best performance was MRDCA, which does not assign weights to each view and considers all the views equally important. The results for multi-view clustering outperformed most of the results utilizing single views.

4.2 Authorship Attribution Analysis

In this section, we utilize the MRDCA algorithm to cluster all the 230 plays using the same parameters from the experiments from the previous section obtaining seventeen clusters. We analyze the results looking for authorial and genre affinities in the plays and analyze the authorship of anonymous and uncertain plays. The complete cluster solution can be found in the Appendix.

Several clusters were formed by homogeneous author contributions. Cluster 7 grouped all the plays by William Shakespeare present in our corpus, and in Cluster 15 only one Shirley James play from 31 plays is not present.

Other groups were composed by different authors, but authors were known for writing style similarity and for collaborating. For example, Cluster 2 was formed by 22 comedies (CO), and 2 tragicomedies (TC). The major contributors are Richard Brome and Ben Jonson. Jonson popularized dramatic comedy and influenced the writing style of many playwriters, which are called *Sons of Ben* [8]. Richard Brome was one, profoundly influenced by Jonson's comedies. Cluster 4 was formed mostly from George Chapman tragedies, and two tragedies by

Jonson. From 8 plays of the cluster, 5 are attributed to George Chapman. Two plays are Roman tragedies, attributed to Ben Jonson. However, the play *Sejanus His Fall* attributed to Jonson has a history of authorship controversies, in which Chapman may have collaborated to write the play [5]. In Cluster 16, there are six plays written by John Lyly in this cluster, and a play by George Peele *The Arraignment of Paris*. This particular play has strong Lyly influences [8].

Another cluster which contains collaborations is the Cluster 12. The major contributors of this cluster are George Chapman and Ben Jonson. The collaborative play *Eastward Ho* written by Chapman, Jonson and John Marston is present and is a prototype of the cluster together with a play from Jonson and a play from Chapman.

Some authors formed more than one cluster and had the plays separated by genre affinities. For example, all plays in Cluster 3 are city comedies by Thomas Middleton, a subgenre of comedy which Middleton helped popularize. Thomas Middleton tragedies present in the corpus were not assigned to this cluster.

The anonymous plays present in the corpus were assigned to Cluster 5 and Cluster 17. The major contributors of Cluster 5 were called the *University Wits*, highly educated late 16th-century writers. The anonymous play *The Reign of King Edward the Third* is often attributed to William Shakespeare, but some researchers claim another author contributed to the play, probably Marlowe or Thomas Kyd. Kyd is nor present in our corpus. Our results suggest Robert Greene, Christopher Marlowe or George Peele contributed to the play, solely or together. [10, 22]. We also support this theory for *Troublesome Reign of King John*, which has also been previously attributed to the three authors, and *The Wars of Cyrus*, which is attributed to Shakespeare and Marlowe. The play by John Ford *Perkin Warbeck* is probably in this cluster by affinities of the historical genre.

CHAPTER 5

Conclusions

The task of authorship attribution of Early Modern Plays is a difficult task. Experts speculate whether books are collaborations and which authors collaborated in a particular book. This discussion takes place because an author writing style is difficult to define. To extract writing style, we convert the text to numerical attributes which can represent the writing style of the authors. However, there are many ways to structure text data, so the authorship attribution task is possible. Lexical features may quantify writing style. We disregard word order and represent each document as a vector of its word frequencies. Or we can represent text with syntactic features. More recently, studies showed complex networks model documents and that we can extract stylistic attributes from these networks.

In this work, we built a corpus of historical documents. We extracted different features from the texts, combining preprocessing tasks and representation models. We calculated several dissimilarity matrices from feature data we extracted. We analyzed which matrices represent authors similarity better. We show that combining those views, we group books by authorial affinities. Then, we utilized a combination of views to cluster books in our corpus. This way, we obtained several clusters with writing style affinities and investigated authorship of anonymous plays present in our corpus. We identified collaborations of different authors and genre similarity in our clusters.

There are plenty of other possible investigations based on this work. One possibility is to divide each play into acts. This way it is possible to investigate authorship for different book parts. Another possibility is to investigate other stylometric features we have not approached, such as semantic features.

APPENDIX A

Multi-view Cluster Solution

In this section, we present all the seventeen clusters obtained utilizing MRDCA. Each cluster is organized in a table, where each row is a play. The columns for each table inform the author, year that the play was written, genre and title.

Heywood, Thomas	1611	TC	The Brazen Age
Heywood, Thomas	1594	TC	The Four Prentices of London
Heywood, Thomas	1610	TC	**The Golden Age, or The Lives of Jupiter and Saturn**
Heywood, Thomas	1612	TC	**1 The Iron Age**
Heywood, Thomas	1612	TC	**2 The Iron Age**
Heywood, Thomas	1607	TR	The Rape of Lucrece
Marston, John	1605	TR	The Wonder of Women, or Sophonisba

Table A.1 Cluster 1 - The play *The Wonder of Women* may be in the cluster by genre affinities.

Brome, Richard	1640	CO	**The Antipodes**
Brome, Richard	1632	CO	The Northern Lass
Brome, Richard	1635	CO	**The Sparagus Garden (Tom Hoydon o' Tanton Deane)**
Brome, Richard	1641	CO	A Jovial Crew, or The Merry Beggars
Brome, Richard	1657	CO	The Queen's Exchange (The Royal Exchange)
Brome, Richard	1632	CO	A Mad Couple Well Matched
Brome, Richard	1639	CO	The Novella
Brome, Richard	1632	CO	The English Moor, or The Mock Marriage
Brome, Richard	1635	TC	**The Lovesick Court, or The Ambitious Politic**
Brome, Richard	1635	CO	The Weeding of the Covent Garden
Brome, Richard	1659	CO	The New Academy or The New Exchange
Brome, Richard	1659	TC	The Queen and Concubine
Chapman, George	1604	CO	The Widow's Tears
Heywood, Thomas	1627	CO	The English Traveller
Heywood, Thomas	1631	CO	1 The Fair Maid of the West
Jonson, Ben	1607	CO	Volpone
Jonson, Ben	1612	CO	The Alchemist
Jonson, Ben	1614	CO	Bartholomew Fair
Jonson, Ben	1614	CO	The Staple of News
Jonson, Ben	1629	CO	The New Inn
Jonson, Ben	1616	CO	The Devil Is an Ass
Middleton, Thomas	1621	CO	Anything for a Quiet Life
Shirley, James	1625	CO	The School of Compliment
Rowley, William	1611	CO	A New Wonder, A Woman Never Vexed

Table A.2 Cluster 2 - principal authors in this cluster are Richard Brome and Ben Jonson

Middleton, Thomas	1613	CO	A Chaste Maid in Cheapside
Middleton, Thomas	1606	CO	A Mad World, My Masters
Middleton, Thomas	1606	CO	**Michaelmas Term**
Middleton, Thomas	1607	CO	**The Phoenix**
Middleton, Thomas	1605	CO	**A Trick to Catch the Old One**
Middleton, Thomas	1607	CO	The Puritan, or The Widow of Watling Street
Middleton, Thomas	1616	CO	The Widow
Middleton, Thomas	1607	CO	Your Five Gallants

Table A.3 Cluster 3 - City comedy plays by Middleton

Chapman, George	1604	TR	**Bussy D'Ambois**
Chapman, George	1608	TR	The Conspiracy of Charles Duke of Byron
Chapman, George	1608	TR	**The Tragedy of Charles Duke of Byron**
Chapman, George	1610	TR	**The Revenge of Bussy D'Ambois**
Chapman, George	1605	TR	Caesar and Pompey (The Wars of Caesar and Pompey)
Fletcher, John	1608	TC	The Faithful Shepherdess
Jonson, Ben	1604	TR	Sejanus His Fall
Jonson, Ben	1611	TR	Catiline His Conspiracy

Table A.4 Cluster 4 - Chapman and Jonsonian Tragedies

Ford, John	1633	HI	Perkin Warbeck
Greene, Robert	1587	CO	Alphonsus, King of Aragon
Greene, Robert	1591	HI	Orlando Furioso
Greene, Robert	1589	CO	Friar Bacon and Friar Bongay
Greene, Robert	1591	TR	The Tragical Reign of Selimus
Marlowe, Christopher	1593	TR	The Massacre at Paris
Marlowe, Christopher	1587	TR	1 Tamburlaine
Marlowe, Christopher	1587	TR	**2 Tamburlaine**
Peele, George	1588	HI	The Battle of Alcazar
Peele, George	1591	HI	Edward I
Peele, George	1594	TR	King David and Fair Bathsheba
anon.	1591	HI	**1 Troublesome Reign of King John**
anon.	1591	HI	2 Troublesome Reign of King John
anon.	1594	TR	The Wars of Cyrus
anon.	1596	HI	**The Reign of King Edward the Third**

Table A.5 Cluster 5 - Four anonymous plays are in this cluster.

Chapman, George	1602	CO	The Gentleman Usher
Ford, John	1630	TR	The Broken Heart
Ford, John	1635	CO	The Fancies Chaste and Noble
Ford, John	1638	CO	The Lady's Trial
Ford, John	1628	CO	**The Lover's Melancholy**
Ford, John	1632	TR	**Love's Sacrifice**
Ford, John	1631	TR	'Tis Pity She's a Whore
Ford, John	1619	TC	The Laws of Candy
Ford, John	1628	TC	**The Queen**
Marston, John	1604	CO	Parasitaster, or The Fawn
Rowley, William	1619	TR	All's Lost by Lust
Dekker, Thomas; Ford, John; Middleton, Thomas; Rowley, William	1623	CO	The Spanish Gypsy

Table A.6 Cluster 6 - Collaborative play The Spanish Gypsy is present in this cluster

Shakespeare, William	1606	TR	Antony and Cleopatra
Shakespeare, William	1598	CO	As You Like It
Shakespeare, William	1611	CO	The Tempest
Shakespeare, William	1603	CO	All's Well That Ends Well
Shakespeare, William	1592	CO	The Comedy of Errors
Shakespeare, William	1595	TR	Romeo and Juliet
Shakespeare, William	1599	HI	Henry V
Shakespeare, William	1609	CO	The Winter's Tale
Shakespeare, William	1591	HI	Richard III
Shakespeare, William	1608	TR	Coriolanus
Shakespeare, William	1590	CO	The Taming of the Shrew
Shakespeare, William	1591	HI	3 Henry VI
Shakespeare, William	1600	TR	**Hamlet**
Shakespeare, William	1610	TR	**Cymbeline**
Shakespeare, William	1591	HI	2 Henry VI
Shakespeare, William	1597	CO	The Merry Wives of Windsor
Shakespeare, William	1595	CO	A Midsummer Night's Dream
Shakespeare, William	1590	CO	The Two Gentlemen of Verona
Shakespeare, William	1597	HI	2 Henry IV
Shakespeare, William	1595	HI	Richard II
Shakespeare, William	1599	TR	Julius Caesar
Shakespeare, William	1605	TR	**King Lear**
Shakespeare, William	1596	CO	The Merchant of Venice
Shakespeare, William	1597	HI	1 Henry IV
Shakespeare, William	1598	CO	Much Ado About Nothing
Shakespeare, William	1602	TR	Troilus and Cressida
Shakespeare, William	1601	CO	Twelfth Night
Shakespeare, William	1594	CO	Love's Labours Lost
Shakespeare, William	1603	TR	Othello

Table A.7 Cluster 7 - Shakespeare plays

Marston, John	1599	TC	**Antonio and Mellida**
Marston, John	1600	TR	**Antonio's Revenge**
Marston, John	1601	CO	**What You Will**
Marston, John	1600	CO	Jack Drum's Entertainment

Table A.8 Cluster 8 - Four plays by John Marston

Brome, Richard	1637	CO	**The City Wit, or The Woman Wears the Breeches**
Fletcher, John	1614	TR	**Valentinian**
Marston, John	1610	CO	**Histriomastix, or The Player Whipped**

Table A.9 Cluster 9

Dekker, Thomas	1601	CO	Blurt, Master Constable
Dekker, Thomas	1634	TR	The Noble Spanish Soldier
Dekker, Thomas	1605	CO	**2 The Honest Whore**
Dekker, Thomas	1611	CO	**If It Be Not Good, the Devil Is in It**
Dekker, Thomas	1599	CO	Old Fortunatus
Dekker, Thomas	1601	CO	Satiromastix, or The Untrussing of the Humorous Poet
Dekker, Thomas	1611	CO	Match Me in London
Dekker, Thomas	1631	CO	**The Wonder of A Kingdom**

Table A.10 Cluster 10 - Thomas Dekker plays

Middleton, Thomas	1624	TC	A Game at Chess
Middleton, Thomas	1618	CO	**The Mayor of Quinborough**
Middleton, Thomas	1615	CO	**More Dissemblers Beside Women**
Middleton, Thomas	1621	TR	**Women Beware Women**

Table A.11 Cluster 11 - Four plays by Middleton

Chapman, George	1602	CO	Sir Giles Goosecap
Chapman, George	1601	CO	All Fools
Chapman, George	1602	CO	**May Day**
Chapman, George	1597	CO	An Humorous Day's Mirth
Chapman, George	1619	CO	Two Wise Men and All the Rest Fools
Chapman, George	1605	CO	Monsieur D'Olive
Heywood, Thomas	1605	HI	2 If You Know Not Me You Know Nobody
Heywood, Thomas	1604	CO	The Wise Woman of Hogsdon
Heywood, Thomas	1602	CO	How a Man May Choose a Good Wife
Jonson, Ben	1598	CO	Every Man in His Humour
Jonson, Ben	1600	CO	Cynthia's Revels
Jonson, Ben	1601	CO	Poetaster
Jonson, Ben	1597	CO	The Case Is Altered
Jonson, Ben	1610	CO	Epicoene, or The Silent Woman
Jonson, Ben	1599	CO	**Every Man Out of His Humour**
Lyly, John	1591	CO	Mother Bombie
Marston, John	1605	CO	The Dutch Courtesan
Middleton, Thomas	1608	CO	The Family of Love
Rowley, William	1622	CO	A Match at Midnight
Chapman, George; Jonson, Ben; Marston, John	1605	CO	**Eastward Ho**

Table A.12 Cluster 12 - A Chapman and Jonson cluster with a known collaboration by both authors, *Eastward Ho*

Fletcher, John	1624	CO	Rule a Wife and Have a Wife
Fletcher, John	1617	TC	The Mad Lover
Fletcher, John	1617	CO	The Chances
Fletcher, John	1618	TC	**The Loyal Subject**
Fletcher, John	1621	TC	The Island Princess
Fletcher, John	1619	CO	**The Humorous Lieutenant**
Fletcher, John	1613	TR	Bonduca
Fletcher, John	1621	CO	The Pilgrim
Fletcher, John	1611	CO	The Woman's Prize, or The Tamer Tamed
Fletcher, John	1620	TC	**Women Pleased**
Fletcher, John	1624	TC	A Wife for a Month
Fletcher, John	1621	CO	The Wild Goose Chase

Table A.13 Cluster 13 grouped works by Fletcher.

Heywood, Thomas	1635	CO	A Challenge for Beauty
Heywood, Thomas	1631	CO	2 The Fair Maid of the West
Massinger, Philip	1623	CO	The Bondman
Massinger, Philip	1621	TR	The Duke of Milan
Massinger, Philip	1631	TC	The Emperor of the East
Massinger, Philip	1627	TC	The Great Duke of Florence
Massinger, Philip	1621	CO	The Maid of Honour
Massinger, Philip	1626	CO	A New Way to Pay Old Debts
Massinger, Philip	1629	TC	The Picture
Massinger, Philip	1624	CO	**The Renegado**
Massinger, Philip	1626	TR	The Roman Actor
Massinger, Philip	1624	TR	**The Unnatural Combat**
Massinger, Philip	1632	CO	The City Madam
Massinger, Philip	1636	TC	**The Bashful Lober**
Massinger, Philip	1633	CO	The Guardian
Fletcher, John; Massinger, Philip; Chapman, George (?); Field, Nathan (?); Jonson, Ben (?)	1617	TR	The Bloody Brother
Fletcher, John; Ford, John; Massinger, Philip; Webster, John (?)	1626	CO	The Fair Maid of the Inn

Table A.14 Cluster 14 - The major contributor is Massinger.

Shirley, James	1633	CO	The Bird in a Cage (The Beauties)
Shirley, James	1632	CO	Changes, or Love in a Maze
Shirley, James	1638	CO	The Constant Maid
Shirley, James	1631	TC	The Contention for Honor and Riches (Honoria and Mammon)
Shirley, James	1635	CO	The Coronation
Shirley, James	1636	CO	**The Duke's Mistress**
Shirley, James	1634	CO	**The Example**
Shirley, James	1633	CO	The Gamester
Shirley, James	1629	CO	The Grateful Servant
Shirley, James	1632	CO	Hyde Park
Shirley, James	1631	CO	The Humorous Courtier
Shirley, James	1635	CO	The Lady of Pleasure
Shirley, James	1631	TR	Love's Cruelty
Shirley, James	1626	TR	The Maid's Revenge
Shirley, James	1634	TR	The Opportunity
Shirley, James	1640	TC	The Arcadia
Shirley, James	1637	CO	The Royal Master
Shirley, James	1639	CO	1 Saint Patrick for Ireland
Shirley, James	1631	TR	The Traitor
Shirley, James	1626	CO	The Wedding
Shirley, James	1628	CO	The Witty Fair One
Shirley, James	1632	CO	The Ball
Shirley, James	1639	CO	The Gentleman of Venice
Shirley, James	1639	TR	The Politician
Shirley, James	1640	TC	**The Impostor**
Shirley, James	1642	TC	The Court Secret
Shirley, James	1652	CO	The Brothers
Shirley, James	1641	TR	The Cardinal
Shirley, James	1638	TC	The Doubtful Heir
Shirley, James	1642	CO	The Sisters

Table A.15 Cluster 15 - Shirley James plays only

Lyly, John	1583	CO	Campaspe (Alexander, Campaspe, and Diogenes)
Lyly, John	1591	CO	**Endymion**
Lyly, John	1584	CO	**Gallathea**
Lyly, John	1590	TC	Love's Metamorphosis
Lyly, John	1589	CO	Midas
Lyly, John	1584	CO	**Sappho and Phao**
Peele, George	1584	TC	The Arraignment of Paris

Table A.16 Cluster 16 - Five Lyly plays and one play by Peele with strong Lyly influences

Chapman, George	1596	CO	**The Blind Beggar of Alexandria**
Dekker, Thomas	1599	CO	The Shoemaker's Holiday
Greene, Robert	1591	CO	George a Green, the Pinner of Wakefield
Greene, Robert	1590	HI	The Scottish History of James IV
Heywood, Thomas	1602	CO	The Fair Maid of the Exchange
Heywood, Thomas	1604	HI	1 If You Know Not Me You Know Nobody
Heywood, Thomas	1599	HI	**1 Edward IV**
Heywood, Thomas	1599	HI	**2 Edward IV**
Heywood, Thomas	1602	TC	The Royal King and the Loyal Subject
Heywood, Thomas	1603	TR	A Woman Killed with Kindness
Lyly, John	1593	CO	The Woman in the Moon
Marlowe, Christopher	1589	TR	The Jew of Malta
Marlowe, Christopher	1592	TR	Dr. Faustus
Marlowe, Christopher	1592	HI	Edward II
Peele, George	1590	CO	The Old Wives Tale
Peele, George	1599	HI	Clyomon and Clamydes
Rowley, William	1638	CO	A Shoemaker a Gentleman
Rowley, William	1608	TC	The birth of Merlin
anon.	1591	TR	Arden of Faversham
anon.	1590	CO	Fair Em
anon.	1589	TR	King Leir and his Three Daughters
anon.	1592	CO	A Knack to Know a Knave

Table A.17 Cluster 17 - several authors contribute to this cluster and four anonymous plays are present.

Bibliography

[1] C. C. Aggarwal and C. Zhai. *Mining text data*. Springer Science & Business Media, 2012. 2.1, 2.1.2, 2.1.3

[2] D. R. Amancio. A complex network approach to stylometry. *PloS one*, 10(8):e0136076, 2015. 1, 2.1.2

[3] D. R. Amancio, O. N. Oliveira Jr, and L. da Fontoura Costa. Identification of literary movements using complex networks to represent texts. *New Journal of Physics*, 14(4):043029, 2012. 1, 2.1.2

[4] A. S. Arefin, R. Vimieiro, C. Riveros, H. Craig, and P. Moscato. An information theoretic clustering approach for unveiling authorship affinities in shakespearean era plays and poems. *PloS one*, 9(10):e111445, 2014. 1, 2.1.3, 2.1.3.1, 2.2, 3.1

[5] A. Barton. *Ben Jonson: Dramatist*. Cambridge University Press, 1984. 3.1, 4.2

[6] P. R. Burns. Morphadorner v2: A java library for the morphological adornment of english language texts. http://morphadorner.northwestern.edu/morphadorner/, 2013. Accessed: 2017-08-28. (document), 3.1, 3.3, 3.1

[7] J. Cong and H. Liu. Approaching human language with complex networks. *Physics of life reviews*, 11(4):598–618, 2014. 1, 2.1.2

[8] D. Daiches. *Critical History of English Literature*, volume 1. Allied Publishers, 1969. 3.1, 4.2

[9] F. D. A. De Carvalho, Y. Lechevallier, and F. M. De Melo. Partitioning hard clustering algorithms based on multiple dissimilarity matrices. *Pattern Recognition*, 45(1):447–464, 2012. 2.1.3

[10] M. Eisen, S. Segarra, G. Egan, and A. Ribeiro. Stylometric analysis of early modern period english plays. *arXiv preprint arXiv:1610.05670*, 2016. 1, 3.1, 4.2

[11] A. B. Farmer and Z. Lesser. *DEEP: Database of Early English Playbooks*. 2007. 3.1

[12] R. Feldman and J. Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007. 2.1, 2.1.2

[13] C. Hoy. The Shares of Fletcher and His Collaborators in the Beaumont and Fletcher Canon (I). *Studies in Bibliography*, 8:129–146, 1956. 3.1

[14] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985. 2.1.3.2

[15] F. Iqbal, H. Binsalleeh, B. C. Fung, and M. Debbabi. Mining writeprints from anonymous e-mails for forensic investigation. *digital investigation*, 7(1):56–64, 2010. 2.2

[16] P. Juola et al. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1(3):233–334, 2008. 1, 2.1

[17] D. Jurafsky. Speech and language processing: An introduction to natural language processing. *Computational linguistics, and speech recognition*, 2000. 2.1.1

[18] V. Q. Marinho, G. Hirst, and D. R. Amancio. Authorship attribution via network motifs identification. In *Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on*, pages 355–360. IEEE, 2016. 1, 2.1.2

[19] V. Q. Marinho, G. Hirst, and D. R. Amancio. Labelled network motifs reveal stylistic subtleties in written texts. *arXiv preprint arXiv:1705.00545*, 2017. 2.1.2

[20] B. Mowat, P. Werstine, M. Poston, and R. Niles. Shakespeare's plays, sonnets and poems from folger digital texts. `www.folgerdigitaltexts.org`, 2017. Accessed: 2017-08-28. 3.1

[21] M. Mueller. Shakespeare his contemporaries. `https://shc.earlyprint.org/`, 2017. Accessed: 2017-08-28. 3.1

[22] D. N. Murphy. *The Marlowe-Shakespeare Continuum: Christopher Marlowe, Thomas Nashe, and the Authorship of Early Shakespeare and Anonymous Plays*. Cambridge Scholars Publishing, 2013. 1, 3.1, 4.2

[23] L. M. Naeni, H. Craig, R. Berretta, and P. Moscato. A novel clustering methodology based on modularity optimisation for detecting authorship affinities in shakespearean era plays. *PloS one*, 11(8):e0157988, 2016. 2.2

[24] T. Nevalainen. *Introduction to Early Modern English*. Edinburgh University Press, 2006.

[25] T. Nevalainen and H. Raumolin-Brunberg. Early modern english. *History*, 1514:1518, 1993. 1, 3.1

[26] E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*, 60(3):538–556, 2009. 1

[27] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010. 2.1.3.2

[28] G. A. Wachs-Lopes and P. S. Rodrigues. Analyzing natural human language from the point of view of dynamic of a complex network. *Expert Systems with Applications*, 45:8–22, 2016. 1

[29] R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005. 2.1.3

This volume has been typeset in LaTeXwith a modified UFPEThesis class (www.cin.ufpe.br/~paguso/ufpethesis) by the author (abd@cin.ufpe.br).