DHQ: Digital Humanities Quarterly
2009
Volume 3 Number 2

Gender, Race, and Nationality in Black Drama, 1950-2006: Mining Differences in Language Use in Authors and their Characters

Shlomo Argamon <argamon_at_iit_dot_edu>, Linguistic Cognition Lab, Dept. of Computer Science, Illinois Institute of Technology, Chicago
Charles Cooney <chu_dot_cooney_at_gmail_dot_com>, ARTFL Project, University of Chicago
Russell Horton <russ_at_diderot_dot_uchicago_dot_edu>, Digital Library Development Center, University of Chicago
Mark Olsen <markymaypo57_at_gmail_dot_com>, ARTFL Project, University of Chicago
Sterling Stein <scubed2_at_gmail_dot_com>, Linguistic Cognition Lab, Dept. of Computer Science, Illinois Institute of Technology, Chicago
Robert Voyer <rlvoyer_at_diderot_dot_uchicago_dot_edu>, Powerset

Abstract

Machine learning and text mining offer new models for text analysis in the humanities by searching for meaningful patterns across many hundreds or thousands of documents. In this study, we apply comparative text mining to a large database of 20th century Black Drama in an effort to examine linguistic distinctiveness of gender, race, and nationality. We first run tests on the plays of American versus non-American playwrights using a variety of learning techniques to classify these works, identifying those which are incorrectly classified and the features which distinguish the plays. We achieve a significant degree of performance in this cross-classification task and find features that may provide interpretative insights. Turning our attention to the question of gendered writing, we classify plays by male and female authors as well as the male and female characters depicted in these works. We again achieve significant results which provide a variety of feature lists clearly distinguishing the lexical choices made by male and female playwrights. While classification tasks such as these are successful and may be illuminating, they also raise several critical issues. The most successful classifications for author and character genders were accomplished by normalizing the data in various ways. Doing so creates a kind of distance from the text as originally composed, which may limit the interpretive utility of classification tools. By framing the classification tasks as binary oppositions (male/female, etc), the possibility arises of stereotypical or “lowest common denominator” results which may gloss over important critical elements, and may also reflect the experimental design. Text mining opens new avenues of textual and literary research by looking for patterns in large collections of documents, but should be employed with close attention to its methodological and critical limitations.

Introduction

The Black stage has been an important locus for exploring the evolution of Black identity and self-representation in North America, Africa, and many African diaspora countries. African-American playwrights have examined many of the most contentious issues in American history since emancipation — such as migration, exploitation, racial relations, racial violence, and civil rights activism — while writers outside of the United States have treated similar themes arising from the history of colonialism, slavery, and apartheid. Alexander Street Press (ASP), in collaboration with the ARTFL Project, has developed a database of over 1,200 plays written from the mid-1800s to the present by more than 180 playwrights from North America, as well as English-speaking Africa, the Caribbean, and other nations. The Black Drama collection is remarkable for the wealth of information provided for each play as well as for the authors, characters, and performance histories of these works. While such extensive metadata permits sophisticated search and analysis of the collection, it also provides an environment that lends itself well to experiments in machine learning and text mining. 1

Using the Black Drama data, we examine the degree to which machine learning can isolate stylistic or content 2

characteristics of authors and/or characters having particular attributes — gender, race, and nationality — and the degree to which pairs of author/character attributes interact. We attempt to discover if lexical style or content markers could be found which reliably distinguish plays or speeches broken down by a particular characteristic, such as gender of character. A positive result would constitute strong evidence for distinctive, in this example male and female, character voices in the sample of plays. If distinctiveness could be shown, we then sought some "characterization" of the differences found, in terms of well defined grammatical or semantic classes.[1] Secondarily, we attempt to see whether and how plausibly comparative data mining tools could aid scholars doing literary critical work. Beyond the statistical results, we examine features the various algorithms used to classify texts and try to determine the degree to which they illuminate or deepen our understanding of the texts in the database.

3 We find that comparative tools doing supervised learning are quite good at classifying plays by American versus non-American authors. Even the outliers, or misclassified plays, demonstrate that these algorithms are able to identify American writers by language usage with remarkable reliability. We are slightly less successful when trying to distinguish the gender of author and/or character. This gender experiment, however, does reveal differences in the ways male and female authors and characters use language.

4 Equally important to the relative abilities of individual tools to classify texts, our experiments have alerted us to potential concerns about data mining as a critical or interpretive endeavor. Comparative classifiers quite powerfully lump together texts and objects that they find "similar" or "dissimilar." However, as we go forward, we have to try to develop results analysis that does not rely on simple binary opposition. Framing comparative tasks based on existing, often binary categories, can lead to results which have a distinctly stereotypical or "lowest common denominator" feel. Application of these new technologies in humanistic research requires that that we understand not only how the tools work, but that we also bring to bear critical evaluation of the questions we ask, the tasks we ask the tools to perform, and the results obtained.

The Black Drama Database

5 The Black Drama collection developed by Alexander Street Press has, at the time of this work, over 1,200 plays by 181 primary authors containing 13.3 million words, written from the middle of the 19th century to the present, including many previously unpublished works [BLDR 2005]. As might be expected of a collection of texts from a particular literary genre, this collection cannot in any way be considered a "random" or statistically representative sample of Black writing in the 20th century. Rather, it reflects editorial decisions and includes key works from a number of American and non-American artistic movements, such as the Harlem Renaissance, Black Arts Movement, and Township Theatre, as well as the principle works by many critically acclaimed authors whom the editors consider, not unreasonably, the most important playwrights. The database contains 963 works by 128 male playwrights (10.8 million words) and 243 pieces by 53 female playwrights (2.5 million words). The most important authors in the collection include Langston Hughes (49 plays), Ed Bullins (47), OyamO (43), and Willis Richardson (41). Plays by Americans dominate the collection (831 titles), with the remaining 375 titles representing the works of African and Caribbean authors. The database contains 317,000 speeches by 8,392 male characters and 192,000 speeches by 4,162 female characters. There are 336,000 speeches by 7,067 black characters and 55,000 by 1,834 white characters with a smattering of speeches by other racial groups. As would be expected, the predominance of American authors is reflected in the nationalities of speakers in the plays. 272,000 speeches are by American characters and 71,000 by speakers from a variety of African nations.

6 Like other Alexander Street Press data sets, the Black Drama collection is remarkable for its detailed encoding and amount of metadata associated with authors, titles, acts/scenes, performances, and characters. Of particular interest for this study are the data available for authors and characters which are stored as "stand-off mark-up" data tables. The character table, for example, contains some 13,360 records with 30 fields including name(s), race, age, gender, nationality, ethnicity, occupation, sexual orientation, performers, if a real person, and type. Even more extensive information is available for authors and titles. The character data are joined to each character speech, giving 562,000 objects that can be queried by the full range of character attributes.

7 The ARTFL search system, PhiloLogic [PhiloLogic 2007], allows joining of object attribute searches, forming a matrix of

author/title/character searching. For example, one can search for words in speeches by female, black, American characters depicted by male, non-American authors in comedies first published during the first half of the 20th century. Given this architecture, designed to support traditional digital humanities research tasks, user-initiated full-text word searches on author and character attributes can help scholars examine specific questions about language use and character depiction. Initial work, for example, on racial epithets in this collection reveals striking differences in the use of such language between male and female authors and characters as well as characters of different races. The plays feature extensive use of racial epithets, especially the frequently appearing “n-word” and its variants. The 5,116 (3.8/10000 words) occurrences of this slur appear in just under 1 in 100 dialogue shifts in the collection and in almost half of all the plays (515). Its extensive use by a substantial majority (119/181) of playwrights in this collection suggests that it has had an important role in the representation of Black experience in the past century.[2] An examination of frequencies suggests that its use is marked by a variety of social factors, such as gender of author and gender/race of character. Male playwrights use the “n-word” twice as frequently as female authors (4.2 vs 2.1/10000 words). Similarly, male characters overall are depicted using this slur almost twice as frequently as female characters (9.5 vs 5.0/1000 speeches). However, factoring author gender into the equation changes the rates somewhat. While male playwrights still represent the genders using it at roughly a 2 to 1, male to female ratio (10.3 vs 5.4/1000 speeches), female authors depict female characters using it at a rate more closely equal to that of males (4.8 male vs 3.5 female/1000 speeches). This leveling of comparative rate may be an artifact of the moderate preference of female authors to represent female characters, as just over a third (34.7%) of speeches in plays by male authors are female characters, while female playwrights allocate slight more than half (52.1%) of speeches to female characters. Similar gender distinctions are also apparent in representation of character race. White characters comprise 14% of speeches (10.5% in female authors). Male authors represent white characters using this racial slur at just under twice the rate as black characters (15.1 vs 8.6/1000 speeches), female authors represent this distinction at a 5 to 4 ratio (5.4 vs 4.3/1000 speeches).

8 While illustrative, such “micro-studies” based on standard full-text searches for specific words or patterns can do little more than hint at larger discursive and representation issues, such as differences between male and female writing. We believe that the new generation of machine learning tools and text data mining approaches have the potential to reveal more general variations in language use because they look for patterns of differential language use broken down by various combinations of author and character attributes.

Philomine: Machine Learning and Text Mining

9 Machine learning and text mining techniques are commonly used to detect patterns in large numbers of documents. These tools are often used to classify documents based on a training sample and apply the resulting model to unseen data. A spam filter, for example, is trained on samples of junk mail and real mail, and then used to classify incoming mail based on the differences it found in the training set. In our application, we already know the classifications of interest, such as author gender and race, for the entire collection. Thus, we apply supervised learning techniques[3] with three rather different objectives: to test whether we can build a training model to classify texts accurately, to identify individual texts which are incorrectly identified after training, and to isolate a set of features[4] which are strongly associated with a particular class. To test the accuracy of a training model, we use n-fold cross-validation, which trains on some of the sample and then attempts to predict the remaining portion of the data set.[5] The machine learning algorithms employed for this study, several implementations of Multinomial Naive Bayes and Support Vector Machine learners, build models by assigning weights to features.[6] Features with high weights, positive or negative, are the most influential in identifying a text as belonging to one class or another. Thus, “viagra” would be assigned a high feature weight in a model underlying a spam filter. In a classification task that results in a high success rate, misclassified documents draw particular interest because these failures of the model often point to literary or linguistic elements which distinguish the outliers from the mass of correctly classified documents.

10 To support text mining experimentation, we have implemented a set of machine learning extensions to PhiloLogic, our full text analysis system, called PhiloMine.[7] PhiloMine allows the interactive construction of machine learning and text mining tasks, and provides links back to documents and features which facilitate further examination. The interactive environment allows users to select and modify the features used for any task in a variety of ways. Feature selection and

manipulation may be the most important aspects of text mining, since the accuracy and validity of classification tasks arise from careful feature selection. For example, basing a strong classification of American and non-American texts on orthographic variations (color/colour) is certainly effective, but of little use for exploring more interesting differences. In addition to PhiloMine, we also use stand-alone text mining applications in cases where we require more sophisticated text sampling or other functions not supported in the interactive environment.

Black Nationality and the Diaspora

11 The diverse Black Drama corpus is a useful collection for examining the stage throughout the Anglophone black diaspora, allowing specific focus on the impact of colonialism and independence, and for comparing African-American plays with works from other cultures. The collection contains 394 plays by American and 303 plays by non-American playwrights written during the period we are studying, 1950-2006. In this experiment, we tested the degree to which we could distinguish between American and non-American plays. To this end, we generated an excluded feature list of words with common spelling differences — such as “color/colour,” “center/centre,” etc — that would have had an impact on results. We further excluded words or names that might appear frequently in a small number of texts, limiting classification on features present in less than 80% but in more than 10% of the documents. The resulting feature list numbered approximately 4200 surface forms of words.

12 For this preliminary experiment, we achieved accuracy rates ranging 85% to 92% depending on the algorithm selected. Specific classifiers yielded slightly different accuracy rates, partly because they weight features and function differently. Using the parameters described above with PhiloMine, the Multinomial Naive Bayesian (MNB) classifier generally had high rates of success distinguishing American and non-American plays, 88.8 percent correct with 84.4 percent correct on cross-validation. Other classifiers achieved similar performance rates for this task. The Weka (Weka3, [Weka3]) implementation of MNB attained 91.4% training and 84.7% cross-validated accuracy and the Weka SMO[8] achieved 94.0% cross-validated accuracy. To ensure that results were not artifacts of the classifiers themselves, we ran random falsification tests, deliberately randomizing the document instances to confuse the classifier [Ruiz and López-de-Teruel 1998]. As would be expected, random falsification generally gave accuracy rates around 50%, such as 49.2 percent using MNB. Our tests revealed that there are significant differences between American and non-American plays, and various learning algorithms can identify them reliably.

13 Splitting the time period in 1984, we found some indication that American and non-American plays might be slightly less distinguishable over the past twenty years. Table One shows the differences in accuracy rates for the earlier period (236/178 American/non-American) and the later period (158/125).

	MNB	**WekaSMO**	**WekaNB**
	raw/xval	xval	raw/xval
1950–1984	93.7/87.6	95.7	96.9/90.8
1985–2006	87.6/81.0	88.0	96.1/82.3

Table 1. Percentage Accuracy by Period/Test

The somewhat lower accuracy for the classification of plays in the later period suggests that these discourses might be merging to an extent. But one should not exclude the possibility that smaller sample sizes may have had some impact on this task.

14 The efficacy of this classification task is matched, to some degree, by the less than startling features most strongly associated with the two bodies of texts. Appendix One shows the top 200 features most predictive of works by American and non-American playwrights. The American plays are marked by their use of slang, references to some place names, and orthographical renderings of speech. The state names suggest a Southern, rural backdrop to many plays, but the terms “hallway” and “downtown” have a decidedly urban feel. The top features of non-American authors had very few slang words and comparatively fewer words that reflect spoken language. Many, in fact, belong to a traditional social

sphere or reveal more formal attitudes toward government, as noted by terms like “crown,” “palace,” “politicians” and “corruption.” The features assigned the highest probabilities in this classification task strike the casual observer as both expected and reasonable.

15 While an extended discussion of the features isolated in this experiment is beyond the scope of this paper, a couple of observations are warranted. First, is the number of features that can be selected for a successful classification task is surprisingly small. We selected the top 30 features from the lists for American and non-American playwrights (a standard PhiloMine function):

```
abroad alabama blasted buddy cattle chief chop colored compound corruption county
custom don downtown
    dude eh elders forbid funky gal georgia git goat goats gods gon' gonna hallway
hmm hunger jive jones
    lawd learnt mississippi momma mon na naw nothin outta palace pat politicians
priest princess professor
    punk quarrel rubbish rude runnin' sho tryin' warriors whiskey wives y' ya' yo'
```

16 Based on this tiny feature set, we achieved similar cross-validated classification performance: MNB 90%; Weka Bayes 85.5%; and Weka SMO 88.8%. While effective on a technical level, few critics would find this list to be a sufficient way to characterize the differences between American and non-American Black drama. The second observation is that different classifiers appear to give weights to remarkably different features. For example, the Weka Naive Bayesian classifier generates a single list of features that looks very little like those above. One sees no place names, no “speech” words, and few terms that stand out on their own as necessarily African, traditional, urban, or Southern:

```
thinks, greeting, aid, million, taken, facts, serve, obviously, mighty, guitar,
    gray, medicine, twenty, tied, shown, practical, matters, corn, luther,
interview
```

17 The Weka SMO classifier identifies features that are more comparable to the list from the MNB. The 20 most heavily weighted features of each corpus were, for American authors:

```
practically, folks, ought, besides, hung, screamed, negro, range, asses,
traveling, finishing,
    underneath, toward, negroes, barely, surround, humor, reaches, bound, ending,
sounds
```

and for non-American authors:

18

```
eh, quiet, pause, rubbish, expose, taxi, hides, concrete, noise, behave, food,
sympathy,
    although, remind, usual, useless, visiting, trousers, towards, leaves, sleep
```

19 While all three of these algorithms produced strong results classifying authors as American and non-American, the different features they rely on raise questions about their usefulness for literary scholars studying text collections.

20 In the case of these three tests, the MNB classifier's feature sets are the easiest to grasp immediately. They point to differences in speech patterns between Americans and non-Americans, different social structures, and different environments. The feature set of the Weka Naive Bayesian classifier makes very little intuitive sense. Why is it that the word “thinks,” either by itself or in conjunction with other terms, most effectively distinguishes a play as American or non-American? To a lesser degree, the Weka SMO classifier's feature set suffers from the same liability. These last classifiers do not give the easy access into the plays that the MNB does. The user can look at the feature set of that classifier and, almost immediately, begin thinking about issues like the migration of black Americans from the rural South to Northern cities. The drawback to the immediacy of this feature set is that it also has a “stereotypical” feel. Plays by non-American black authors, following a simple-minded reading of the MNB result set, might be thought to take place in villages where chiefs and elders buy wives with goats and cattle. In this sense, the features tend to obscure any nuance in either corpus because the algorithm itself tends to be very selective as it generates a model.

21 Beyond feature lists, an examination of incorrectly classified documents demonstrates the utility of comparative

classification. In most PhiloMine tasks, the system will identify documents that are not correctly classified. As might be expected, some non-American subgroups were classified better than others. Thus, British, Australian, and Canadian authors were more frequently misclassified in this task than African or Caribbean playwrights. Inspecting the outliers was instructive. For example, eight plays by Kia Corthron were consistently misclassified as American. Upon further review, the problem arose not because of the algorithm, but the metadata we provided it. The editors incorrectly identified this American playwright as a Canadian. Because of the lexical features it found, the classifier decided that her plays, in fact, should not be grouped with those of the non-American authors, as the metadata would have had her, but with the American authors. This misclassification was actually a correct classification that alerted us to the error in the metadata. Other outliers also bear further inspection, which may raise further critical questions. The plays of Montserratian Edgar White were correct classified except for one, *When Night Turn Day*, which is set in the Bronx. Similarly, American Joseph Walker was correctly classified in 9 of 10 cases, except for *The Lion is a Soul Brother*, set in "rural/jungle West Africa." OyamO's *Three Red Falcons?*, set in rural Nigeria, was incorrectly classified, but so too was his 1995 play, *Dancing on the Brink*, which was set in Palo Alto. Examination of outliers in a generally successful classification task may provoke new questions about why a particular text was not classified properly.

Comparative text mining on the nationality of Black playwrights in the second half of the 20th century shows clear distinctions between American and non-American authors. The algorithms consistently achieve high levels of performance and generate distinguishing feature lists that are of potential interest in helping to characterize topics, themes, and styles that mark each group. Finally, the outliers or incorrectly classified authors and plays may lead to further critical scrutiny, allowing for variations within the general classification. 22

It may be objected that our classification task itself is a trivial case, attempting to confirm a distinction that is all too obvious to be of significant literary interest. A task like this, however, provides a good test case for verifying how well classifiers are able to function. We were able to check the accuracy of our results easily through the bibliographic data. In the case of the outlier, Corthron, a web search confirmed that she is, in fact, American. Of course, features that reveal certain linguistic and dramatic differences between the American and non-American plays might not be particularly surprising. Authors tend to place characters in localities and use idiomatic language they themselves know. For example, an American playwright is more likely to set a play in Mississippi and depict characters using a Southern vernacular. However, the features are often able to bring out further, less obvious stylistic and creative choices in operation across text corpora. It is beyond the scope of this paper to address just what these could be. A scholar with a deeper knowledge of this body of works could potentially examine the feature sets and find lexical patterns that point to larger tendencies. In the end, the algorithms can only rely on and propose what they find to be statistically significant. The scholar, as always, must decide the meaning of the results. 23

Mining Author and Character Gender

Classification tasks based on the gender of authors, characters, and a combination of the two provides a more challenging test than classification by author nationality. Previous work suggests that gender classification tasks are somewhat less accurate and that the feature sets they generate are less obviously distinctive [Argamon et al. 2003], [Koppel et al. 2002]. The nature of these documents, plays with male and female characters and with a significant amount of generic language found in stage directions and other textual objects, increases the difficulty of automatic classification. The issue of male authors writing female characters and vice versa makes the task of classifying by gender more difficult, but also introduces the ability to test the degree to which authors of one gender can effectively write characters of the other gender. 24

The Black Drama collection contains 573 (82%) plays by male and 124 (18%) by female playwrights written between 1950-2006. An initial classification using our default feature set selection is deceptively accurate, with MNB returning 79.4% cross-validated accuracy and Weka SMO 85.1% cross-validated. The Weka "confusion matrix" indicates that a significant majority of male authors are correctly classified while performance on female authors is less than 50%: 25

a	b	← classified as
680	45	a = male
85	64	b = female

Table 2.

The classifier performed only slightly better than if it had simply guessed “male” for every document (82%). Therefore, to determine whether the system was in fact finding a gender difference, we had to balance corpus sizes. PhiloMine supports a standard corpus balancing option that randomly selects documents from the larger sample until it has found a number equal to the smaller class. Since this is a random selection, effectively comparing all of the female plays against a set of male documents, one needs to perform this task numerous times to determine an average accuracy rate. Using the Weka NB and SMO functions five times each, cross-validated results indicate a fair degree of success:

```
WekaSMO (cross-validated): 70.2%, 78.2%, 79.0%, 73.0%, 73.8%

WekaNB (cross-validated): 65.7%, 74.2%, 68.1%, 70.2%, 69.8%
```

Classifying author gender of randomly balanced test sets arrived at an accuracy of 70% for Naive Bayesian and 75% for Support Vector Machine. Similar performance of high 60%/low 70% accuracy was also generated by the MNB function. Applying a random falsification test to one of these tasks we found an accuracy rate of 49%. This suggests that, while it is not as strong as in the nationality classification, a signal can be found that is not an artifact of the selected classifier or other influences.

26 The classification tasks to this point have been looking at entire plays with no attempt to control for expected skewing factors, such as stage directions, which would not be expected to have as strong a gendered writing style; and characterization, where authors depict characters of the other gender. Controlling for obvious skewing influences, particularly as a control following more generic experiments, tends to provide more coherent results, but at the cost of creating composite documents which do not reflect the structure and organization of the plays as individual works.

27 For this experiment, we rebuilt the Black Drama collection under PhiloLogic/PhiloMine to behave as a database of some 13,000 distinct objects (characters) containing 8.9 million words. We eliminated stage directions, cast lists, and other such apparatus containing some 4.5 million words, while combining individual speeches of each character into one object for faster processing. This changed the unit of analysis from documents to composite character speeches with associated metadata, including author attributes. For the period 1950-2006, there are 4,228 characters by male authors and 865 characters by female playwrights with more than 200 words. For the same period, there are 3,226 male characters and 1,742 female characters.

28 Classifying all the characters by author gender using the PhiloMine Bayesian function resulted in 75.5% cross-validated accuracy. This general result was confirmed on 5 runs using the random document balancing function (865 characters), with cross-validated accuracy rates of 72.8%, 72.1%, 71.0%, 72.2%, and 71.6%. Deliberately shuffling the character instances of randomly balanced characters returned results approximating the expected 50% accuracy: 47.6%, 51.2%, and 50.8%. Similar results were obtained for classifications by character gender. Overall, for all of the characters, MNB accurately classified 73.2% of the instances. Five runs of randomly balanced characters (1,742) resulted in accuracy rates of 72.5%, 71.1%, 72.3%, 72.2%, and 71.4%. Random falsification tests again approximate the expected 50% accuracy: 48.4%, 48.1%, and 50.4%. Classification of gender or author and character on composite character objects showed modest improvement in accuracy. Further, we found that classifying on more major characters (total words greater than 1,000 and 2,000) again resulted in modest increases in accuracy.

29 This experiment suggests that, as we balance our authors and characters more rigorously, essentially testing on a more and more abstract representation of the texts, our success rates improve. We first extracted all speeches with character gender attributes from the corpus, splitting them into tokenized word frequency vectors for all authors, all characters, male authors, female authors, male characters, and female characters. For each of these, we used SVM-Light[9] to build a model to identify authors and characters by gender. As indicated in Table Three, the system correctly identified 88.2%

of the authors' gender and 77.4% of the speakers' gender.

Given:	Female Author	Female Speaker	Male Author	Male Speaker	Full Sample	Full Sample
Find gender of:	***Speaker***	***Author***	***Speaker***	***Author***	***Speaker***	***Author***
Accuracy	69.7%	83.1%	78.9%	88.6%	77.4%	88.2%
Majority Class	54.5%	74.5%	69.4%	84.7%	66.6%	8.13%

Table 3. Accuracy over 10-fold cross-validation

Performance varied when examining subsets of the corpus, from 86.6% for gender of author in male characters to 69.7% for gender of speaker in female authors. The differences in accuracy in male and female author/character might, however, result from the fact that male authors tend to include fewer female characters, as we show in the Majority Class of Table Three, which indicates the rate of male instances for each assessment. For female authors, male characters constitute 54.5% of the speakers.

We then equalized a test sample for class by discarding instances in the majority classes until we had a balanced set. As part of this process, we further corrected for number of words in each character class by selecting character instances to balance the word frequencies overall. As shown in Table Four, this test sample produced a balanced dataset by number of instances and number of average words spoken by each character.

Name	Female Author	Female Speaker	Male Author	Male Speaker	Author	Speaker
Given:	**Female Author**	**Female Speaker**	**Male Author**	**Male Speaker**		
Find gender of:	***Speaker***	***Author***	***Speaker***	***Author***	***Speaker***	***Author***
Accuracy	71.9%	82.8%	80.3%	83.7%	79.9%	86.4%
			Average words/document			
Male	770.1	877.8	853.4	645.2	859.4	750.8
Female	777.9	877.9	853.0	645.0	859.4	750.8

Table 4. Accuracy over 10-fold cross-validation

The creation of a more controlled data set discriminated author and character gender effectively, achieving cross-validated accuracies of between 72 and 86 percent. Author gender was identified more accurately and consistently than speaker gender. Table Five demonstrates that male authors/speakers are correctly identified somewhat more often than are female authors in five of the six cases, with the sole exception of characters in female-authored plays, where identification rates are roughly the same for males and females.

Name	Female Author	Female Speaker	Male Author	Male Speaker	Author	Speaker
Given:	**Female Author**	**Female Speaker**	**Male Author**	**Male Speaker**		
Find gender of:	***Speaker***	***Author***	***Speaker***	***Author***	***Speaker***	***Author***
Male Accuracy	71.8%	85.7%	81.7%	86.6%	81.1%	88.9%
Female Accuracy	72.1%	79.9%	78.8%	80.8%	78.6%	83.9%

Table 5.

This suggests that male language in Black Drama is somewhat more marked than female language, both at the level of authors and at the level of language represented by how authors write characters.

Machine learning systems are clearly able to identify gender of playwrights and their characters with impressive accuracy. We have found that on raw running texts, the systems can reliably identify author gender at rates between the high 60s and mid-70s percent accuracy. And, as the plays are processed in various ways to eliminate potential skewing factors — such as unbalanced subsets, extraneous textual data (e.g. stage directions), and differences in raw word counts — classification performance increases. This increase in performance comes, however, at the cost of increasing the distance from the text themselves. 31

Given the ability to classify author and character genders, we will now return to the texts briefly to examine the features most characteristic of gendered writing in Black Drama. Appendix Two shows the top 200 features as measure by Bayesian probability ratios, broken down by male and female playwrights without respect to character gender. The features suggest a rather traditional set of gender distinctions. Male authors tend to focus on legal/criminal issues (officer, gang, pistol, jail, etc.); numerous obscenities and slurs (bullshit, nigger(s), goddamn, fuck, shit); music (band, drum, leader, drums, spiritual, player, jazz); and money (dollars, price, cents, cash, etc.). Female playwrights of this period tend to privilege issues concerning family/home (child, stories, hug, mama, girls, birth); emotive states (smiling, imagine, memories, memory, happiness, happy); descriptions (handsome, lovely, grace, cute, courage, loving, ugly); and women (herself, girls, she, female, lady, women, her). The representation of traditional gender roles are most notable in the characterization of non-American male authors. As shown in Appendix Three, male characters are given very clear public roles, with the top eight words by probability ratios being “chief, order, government, lead, power, position, country, land” while the female characters are limited to the private realm (“husband, dress, shame, marry, doctor, married, please, parents”). 32

Gender characterization among American authors writing between 1950-2006 (Appendices 4 and 5) provides evidence that men and women depict characters slightly differently. The feature sets are rather similar: both contain numbers (men generally do counting) and share terms as varied as sir, american, power, bitch, country, and killed. But the male list is noticeably coarser, with more profanity, the term “nigger(s),” and more references to law enforcement and violence. With only a few exceptions, the female character feature lists have basically the same domestic tenor. In contrast to male characters, female characters apparently use very little profanity and seem to be much less involved in struggles with public authorities. Of course, these lists only reveal generalities. The features are differential frequencies. There might in fact be foul-mouthed female characters in the corpus who are active in the public sphere. But those characters would probably be exceptions. The degree to which these lists reveal true differences among black American male and female authors is a matter for discussion. The important thing is that the mining algorithm gives fuel to the discussion and serves as a starting point for closer textual study. 33

This same character gender classification test on non-American authors yields feature sets suggesting even more disparate depictions of the sexes than among American authors. Appendices 6 and 7 show that, for authors of both 34

sexes, male characters inhabit the public sphere, their discourse deals with leadership, and they are more likely to use grammatically precise terms like “which” and “whom.” Female characters' language, again, primarily centers on domestic concerns. Distinctions between male and female authors' use of language in the depiction of their characters are few. One of the striking differences, however, is that only the male characters written by female authors in this data set use scatalogical language at a significant rate. And comparing the results from the American and non-American tests highlights the different concerns for these characters who inhabit different cultures.

Classification of texts by gender of playwrights, characters, and the combination of the two is a more difficult test than by nationality. Results ranging from high-60 percent accuracy for plays as they are found to the mid-80s in carefully constructed samples extracted from the texts. It is also clear that the features used to construct classifier models might be of interest to literary researchers in that they identify themes and language use that characterize gender distinctions. Of course, men talk more of wives than women and only women tend to call other women “hussies,” so it is hardly surprising that male and female authors/characters speak of different things in somewhat different ways. The features suggest, however, that we are finding “lowest common denominators” which distinguish male from female, but which may also privilege particular stereotypes. The unhappy relationship of Black American men with the criminal justice system or the importance of family matters to women are both certainly themes raised in these plays. The experimental design itself, using classifiers to detect patterns of word usage which most distinguish the genders, may bring to the forefront literary and linguistic elements which play a relatively minor role in the texts themselves. 35

Conclusion

We have found that, although algorithms can in fact detect differences in lexical usage to a striking degree and output feature sets that characterize differences between corpora of data, human scholars must still do the work of scrutinizing results and, more importantly, decide how best to develop these tools for humanities research. Fundamentally, automatic classifiers deal with general features and common traits. In contrast, in recent decades, literary criticism has focused on the peripheries, often looking at the ways understudied works by authors from underrepresented groups work within a larger cultural context. Literary critics have tried to nuance general understanding about what the mainstream is and how it works. As we mentioned above, a danger in framing comparative tasks based on binary oppositions is that doing so can produce simplistic or stereotypical results. Furthermore, given the power of classifiers, we might always be able to prove or detect some binary opposition between two groups of texts. And so the task before us, if we want to develop tools to aid literary criticism, is to try in some way to respond to the values driving scholarship currently and, as those values change, continue to take them into account. We must also keep in mind that measures of success and procedure are often different for computer scientists and literary critics. For example, using only 60 total feature terms, 30 per corpus, we can classify the Black Drama texts as American or non-American with approximately 90% accuracy. Distinguishing differences on such a small number of words is an impressive technical feat, to be sure. But to a literary scholar, such a circumscribed way of thinking about creative work may not be terribly fruitful. As we go forward, we will have to try to bridge gaps such as these. Our success in this endeavor will, of course, depend upon close collaboration between those building the data mining tools and those who will finally use them. 36

Appendix One

American Playwrights	Non-American Playwrights
ya', momma, gon', jones, sho, mississippi, dude, hallway, nothin, georgia, yo', naw, alabama, git, outta, y', downtown, colored, lawd, mon, punk, whiskey, county, tryin', runnin', jive, buddy, gal, gonna, funky, louis, busted, piano, banks, folks, huh, talkin', ol', stack, rip, washington, exiting, hisself, lyin', kin, tellin', blues, callin', preacher, porch, tom, luther, an', buck, stairs, lookin', pops, dime, holler, nothin', lotta, workin', puttin', doin', negro, sittin', somethin', johnson, chocolate, thinkin', chicago, dope, uh, neighborhood, humor, negroes, crosses, c', ain', sayin', pop, askin', should', reverend, bein', oughta, yep, givin', basement, gray, mule, smith, bitches, bill, closet, freak, figured, makin', feelin', havin', clay, hammer, livin', ta, gut, upstage, ass, avenue, lee, sidewalk, waitin', reckon, wanna, rap, dig, ma', hop, takin', singers, someplace, lincoln, cats, june, gotta, playin', niggers, asses, mama', gettin', comin', walkin', goin', em, satan, cute, intense, nigger, ole, harlem, cept, daddy', lord', righteous, nerve, sofa, awhile, jazz, bullshit, somebody', apartment, punches, toward, um, da, summer, liable, aunt, grace, paul, could', rag, shit, dammit, blackness, hooked, joint, southern, frank, gotten, upstairs, sung, holiday, honey, baby, downstairs, messing, con, grins, hung, dreamed, would', d, kinda, yea, juice, cause, traveling, fuck, bar, downstage, butt, drift, name', guys, favorite, colors, sadness, screw, congregation, dollar	na, learnt, don, goat, rubbish, eh, chief, elders, compound, custom, rude, blasted, quarrel, chop, wives, professor, goats, pat, corruption, cattle, hmm, priest, hunger, palace, forbid, warriors, princess, gods, abroad, politicians, dey, boot, harvest, ancestors, mate, idiot, d', trace, witch, nearer, frighten, bloody, economic, messenger, native, palm, government, royal, crown, greetings, properly, addressing, tradition, visitors, madam, strangers, development, patience, drumming, mere, village, citizen, rejected, husbands, youth, fourth, fathers, disturb, proceed, shall, salute, surrender, dem, forest, hen, port, duties, british, prisoners, panting, gate, prime, flood, accused, rejoice, politics, x, excitedly, trousers, taxi, senior, reject, branches, towards, officer, seas, tribe, political, wailing, accompanied, fetch, whom, sand, official, england, arrives, disgrace, beads, reply, market, guards, hut, oil, 1st, sacrifice, wisdom, o, boss, disaster, assistant, obey, corpse, behave, arrival, improve, cannot, stream, expose, council, iron, matters, tax, democracy, london, wealth, leadership, women', 2nd, warrior, kneel, obeys, appeal, thief, soldier, greet, burial, advise, deaf, alarmed, throne, arrive, therefore, judgement, warn, petty, powers, haste, stinking, medical, await, armed, instructions, sergeant, fails, estate, worship, soil, madness, land, foreign, breed, military, beg, rushed, liberty, graves, secretary, absence, greeting, cock, enemies, court, politician, object, lake, urge, oath, greets, fruits, disease, gathering, insult, sons, shield, bundle, succeed, idle, useless, message, thirst

Table 6. American vs non-American Playwrights, 1950-2006: MultiNomial Naive Bayesian top 200 features by probability ratio

Appendix Two

Male Playwrights	Female Playwrights
buddy, joe, jack, ol', union, johnson, sho, officer, gang, john, jive, pistol, leaps, slight, drag, band, shadow, drum, leader, shouts, bullshit, drums, image, everytime, major, jail, gradually, nigger, thunder, naw, mister, shooting, crowd, heavily, post, niggers, blind, stairs, judge, dollars, price, goddamn, wow, de, cops, flash, dig, lawyer, strike, cents, cash, git, million, master, silent, dogs, uncle, bones, hill, rule, spiritual, jim, player, kid, hall, fires, boy, brothers, preparing, devil, thousand, enemy, ma, gun, pig, hundred, spit, alright, fades, everything', lotta, west, gentlemen, square, joint, ha, smoke, main, robe, da, shine, numbers, mountain, rent, double, ghost, however, bus, stunned, jazz, strikes, police, hip, guilty, simple, sir, rear, gimme, wooden, fuck, fish, fifty, spite, struck, lap, march, record, rises, law, knife, example, reality, pauses, oughta, intend, precious, cat, fire, christ, build, created, ass, sittin', chain, forces, suffer, shit, allowed, approach, ta, president, figures, soul, pause, led, glory, shock, raises, somehow, nothin', moon, stick, murder, passes, lead, takin', starts, steel, killed, prison, jesus, hey, honor, steal, startled, bone, bless, party, dollar, people', evil, song, blow, slavery, silently, happening, voice, huh, row, seeing, ignorant, fellow, hell, son, couple, reverend, bear, greatest, cost, suddenly, others, damn, doorway, above, goin', death, loose, grab, killing, witness	acts, queen, summer, rain, television, grand, able, grace, response, child, tie, handsome, smiling, tea, doctor, tongue, green, whore, wet, audience, imagine, stories, hug, language, conversation, belly, blue, mama', lovely, towards, fingers, words, memories, mid, dirt, note, colors, day, bill, expected, chorus, wrapped, expensive, thirties, girls, birth, touches, lips, memory, babies, nearly, breast, extra, heart, dress, she', welcome, wearing, worn, dreams, teeth, flying, sat, pregnant, excited, nurse, touching, dim, 6, secret, thoughts, cute, courage, herself, actually, somebody', trees, beauty, apple, yo, discovered, marriage, forehead, thin, smart, happiness, happy, grows, notes, short, female, hair, lady, months, stood, whites, news, smile, responsibility, helped, truly, missing, circle, furniture, books, amused, 5, pop, passion, likes, weather, keeps, needed, fancy, gift, seriously, purse, smell, daddy', aren', wanting, visit, sweat, women, river, spread, forgotten, french, grew, gentle, sisters, enjoy, absolutely, buried, spoken, pain, letters, drawn, reading, cigarette, telephone, period, family, someone, condition, nervous, favorite, promised, legs, slaves, searching, loved, push, lighting, famous, raising, listened, aunt, loving, ugly, barely, thick, spirits, kisses, respond, loves, birthday, rise, cleaning, lover, talks, 30, sweet, pictures, close, bout, awake, christmas, information, dry, harder, brings, touch, members, given, surprised, why', bein', spring, her, covers, she, r, thinks, teacher, breath, blessed, slept, deep, color

Table 7. Black Drama: Male vs Female Playwrights, 1950-2006: MultiNomial Naive Bayesian top 200 features by probability ratio

Appendix Three

Male Characters	Female Characters
chief, order, government, lead, power, position, country, land, hey, fellow, war, question, ourselves, spirit, man', thousand, game, present, party, public, among, part, under, law, cannot, words, most, which, act, secret, may, against, hundred, great, twenty, move, indeed, join, themselves, questions, continue, its, scene, case, human, upon, sir, our, half, white, point, decided, play, sense, simple, shall, an, others, fighting, story, lady, peace, blood, fall, friend, fire, eye, reason, news, fight, problem, sort, hungry, sell, return, idea, given, force, along, rich, world, reach, man, few, hand, yeah, piece, deal, future, earth, clear, best, makes, those, ears, their, learn, wife, many, by, set, ten, we, answer, later, hands, agree, black, catch, death, easy, without, seven, between, end, special, fear, whole, trust, state, beat, also, tree, damn, anyone, everyone, city, car, cut, meeting, unless, shut, watch, real, tomorrow, one, people, being, known, such, killed, has, dog, mine, follow, important, break, six, second, boy, fool, sun, speak, town, himself, other, must, allow, lot, there, read, rather, once, chance, today, dark, comes, beginning, new, hell, from, old, made, less, times, side, eh, instead, four, matter, two, lost, born, king, free, brother, somewhere, means, show, seen, into, running, three, next, whose, own, lives, head, will, full	husband, dress, shame, marry, doctor, married, please, parents, oh, he', o, tired, girl, dear, really, love, gone, hurt, child, school, clothes, sister, glad, clean, happen, room, nice, sick, won', care, him, hate, sweet, couldn', hurry, aren', mother, wake, baby, daughter, didn', sleep, somebody, girls, bed, late, laugh, guess, isn', yourself, looking, miss, suppose, terrible, evening, wouldn', worse, coming, feel, don', dance, inside, stay, wanted, went, imagine, leave, stupid, meet, t, night, he, feeling, touch, door, self, beg, tonight, blame, pass, seeing, going, shouldn', cold, why, children, they', telling, lord, father, sit, sometimes, smell, lie, anybody, help, cry, pick, stop, thank, can', just, early, change, doesn', started, getting, seems, perhaps, mr, family, died, could, wonder, eat, something, it', re, beautiful, you', calling, anything, met, kind, told, crazy, person, begin, ready, friends, except, working, always, so, poor, she, months, house, knew, go, finished, woman, mad, water, son, quite, excuse, sure, till, happy, happening, sorry, want, wasn', used, yesterday, waiting, hope, talk, forget, know, too, funny, morning, home, bad, everybody, quiet, outside, drink, think, least, happened, come, worry, true, things, quick, hear, alone, tell, said, m, say, get, thought, much, everything, done, afraid, wish, none, heart, came, god, about, thing, talking, giving, better

Table 8. Black Drama: Non-American Male Authored Characters, 1950-2006: MultiNomial Naive Bayesian top 200 features by probability ratio

Appendix Four

Male Characters, American female authors	Male Characters, American male authors
hey, office, sir, state, shit, ass, hour, order, son, doin', goin', somethin', wife, straight, eye, self, war, whatever, hell, damn, aw, human, game, respect, yeah, kinda, ought, behind, problem, freedom, round, welcome, brother, n, catch, spend, country, air, paid, man, em, free, plan, sunday, check, running, car, fool, laugh, bitch, gotta, doctor, probably, meet, cause, outta, drive, job, decided, sign, takes, ahead, front, trees, let', money, power, unless, fight, week, sugar, fact, afternoon, next, nobody, number, paper, table, ready, shot, ones, breath, y', peace, dog, york, lady, ten, set, sell, heavy, drink, question, scene, glass, church, man', stuff, place, gettin', upon, miss, food, business, right, pay, face, morning, deal, top, minute, late, hundred, case, thirty, new, herself, finished, start, moment, city, nothin', important, listen, passed, break, feelings, either, now, busy, street, ride, which, caught, buy, standing, any, minutes, american, across, known, crazy, working, actually, keep, conversation, call, show, corner, pick, name, got, tonight, line, kids, against, left, bring, best, seven, ain', expect, person, needed, second, its, school, story, quit, alright, side, killed, may, where, boy, rest, afraid, since, picture, five, fire, shall, none, nor, hard, finish, seen, lay, tomorrow, stand, own, open, high, early, shoes, quite, figure, already, these, okay	dig, naw, hey, sir, gotta, yeah, hundred, shit, state, land, negro, american, gun, game, dollars, war, law, fifty, story, order, problem, nigger, hell, cool, brothers, point, america, york, power, figure, shot, peace, sell, ass, pull, check, damn, line, question, blow, five, niggers, case, couple, great, thousand, city, country, break, ground, white, man, fuck, fair, deal, shoot, twenty, rich, wanna, okay, south, killed, paper, town, number, against, perhaps, worth, whole, man', play, hit, black, wife, fact, here', cut, gonna, six, alright, swear, sun, straight, git, uh, side, lives, thanks, let', jail, changed, stupid, hours, happening, standing, short, bout, everybody, human, also, kill, ah, huh, drop, across, which, ten, three, front, office, running, means, four, lady, got, earth, working, piece, brother, show, job, fight, music, upon, weeks, people, money, fire, chance, pay, hang, far, asked, calling, stuff, somewhere, walk, shut, seen, boy, devil, new, company, talking, police, part, kinda, em, thirty, their, seven, playing, gon', turned, lose, sense, two, run, outside, found, serious, sent, its, us, outta, asking, pretty, alive, world, waiting, lying, bitch, paid, buy, near, went, first, shall, kept, talked, caught, sound, boys, saying, many, death, dead, catch, throw, taken, finish, cause, give, making, right, idea, streets, song, big, down

Table 9. Black Drama: American Female vs Male Authored Characters, 1950-2006: MultiNomial Naive Bayesian top 200 features by probability ratio

Appendix Five

Female Characters, American female authors	Female Characters, American male authors
ugly, husband, tea, babies, child, mama, loved, sing, books, almost, arms, dress, colored, words, poor, soul, loves, born, aunt, thank, young, scared, hadn', negro, seems, pain, read, children, dr, hair, lose, talked, dance, blame, looks, funny, grow, clothes, memory, loving, tree, fall, girls, please, hate, inside, helped, sister, low, kiss, promised, outside, having, between, especially, instead, nice, voice, stories, marry, except, mother, love, wear, gave, music, girl, sat, seem, hurt, oh, lost, waiting, wine, names, used, baby, bed, says, kitchen, doesn', sick, special, beat, hope, thinks, speak, small, gotten, end, changed, carry, him, isn', daddy, knew, feeling, because, feel, lived, god, wanted, fingers, remember, exactly, reach, dream, somewhere, he', felt, streets, others, bit, anymore, today, nothing, kill, said, cry, believe, smell, calling, learned, learn, body, little, dear, forgive, moving, dressed, excuse, gone, live, gets, through, sometimes, grown, eat, given, die, different, pretty, he, couldn', able, year, write, came, knowing, rich, called, month, along, sun, blue, stop, always, listening, laughing, myself, play, anyone, house, going, ya, shouldn', looked, faces, age, six, mad, everyone, meant, makes, something, true, tongue, likes, floor, color, watch, bad, taste, won', died, when, did, idea, yourself, old, mouth, sisters, act, never, being, married, went, really, beautiful	honey, husband, child, lord, dress, silly, dinner, oh, daddy, bed, doctor, girl, children, room, hello, aren', please, dear, poor, girls, father, someone, married, love, age, doesn', alone, hair, kiss, hardly, jesus, shouldn', loved, hurt, herself, mother, isn', thank, marry, does, hospital, nice, sometimes, he', such, brought, wear, sister, floor, yes, sick, glad, anymore, ought, hurry, body, longer, cry, knows, church, fun, suppose, clean, needs, house, yourself, happy, morning, wish, goin', person, feel, wonder, haven', evening, sweet, strong, miss, late, mrs, today, eyes, tired, an', yours, leave, fix, touch, though, beautiful, besides, daughter, drunk, stop, takes, meant, slave, afraid, family, o', hope, mouth, feeling, leaving, taking, myself, comin', wants, should, you', much, sure, eat, always, gave, promise, met, dance, special, blue, seems, having, deep, trust, wouldn', dream, again, woman, door, home, him, going, too, she', excuse, almost, stay, sorry, strange, heart, rest, never, past, living, early, won', mama, anyone, talk, re, mind, why, mad, looks, phone, after, tonight, hear, especially, years, college, table, nothin', worry, ve, tomorrow, anything, inside, gets, clothes, evil, since, better, become, enough, things, help, trouble, knew, important, expect, none, want, forget, used, somethin', school, god, baby, young, write, skin, happen, comes, her, friends, forever, care, soul

Table 10. Black Drama: American Female vs Male Authored Characters, 1950-2006: MultiNomial Naive Bayesian top 200 features by probability ratio

Appendix Six

Male Characters, non-American male authors	Male Characters, non-American female authors
chief, order, government, lead, power, position, country, land, hey, fellow, war, question, ourselves, spirit, man', thousand, game, present, party, public, among, part, under, law, cannot, words, most, which, act, secret, may, against, hundred, great, twenty, move, indeed, join, themselves, questions, continue, its, scene, case, human, upon, sir, our, half, white, point, decided, play, sense, simple, shall, an, others, fighting, story, lady, peace, blood, fall, friend, fire, eye, reason, news, fight, problem, sort, hungry, sell, return, idea, given, force, along, rich, world, reach, man, few, hand, yeah, piece, deal, future, earth, clear, best, makes, those, ears, their, learn, wife, many, by, set, ten, we, answer, later, hands, agree, black, catch, death, easy, without, seven, between, end, special, fear, whole, trust, state, beat, also, tree, damn, anyone, everyone, city, car, cut, meeting, unless, shut, watch, real, tomorrow, one, people, being, known, such, killed, has, dog, mine, follow, important, break, six, second, boy, fool, sun, speak, town, himself, other, must, allow, lot, there, read, rather, once, chance, today, dark, comes, beginning, new, hell, from, old, made, less, times, side, eh, instead, four, matter, two, lost, born, king, free, brother, somewhere, means, show, seen, into, running, three, next, whose, own, lives, head, will, full	fellow, king, war, government, problems, human, e, sacrifice, mr, sir, ancestors, action, peace, quite, agree, public, country, rather, order, follow, law, idea, news, friend, simple, message, belong, dream, indeed, certain, doubt, begin, case, excuse, thinking, skin, boys, possible, line, great, difficult, which, shall, whom, himself, fathers, sense, mad, meeting, worth, village, question, game, wonder, immediately, small, person, might, nothin', met, ways, let', read, offer, chance, true, killed, against, freedom, search, whether, boy, evening, answer, plan, plenty, expect, fire, eye, shut, story, haven', also, truth, respect, present, ourselves, such, shit, rest, man, most, clear, an, themselves, paid, ours, town, return, worse, believe, thank, red, making, few, please, good, sure, taken, step, allow, very, may, them, should, by, set, open, evil, heard, along, his, sent, everybody, big, pass, year, different, given, fine, hell, welcome, point, there, matter, wife, trust, those, straight, round, none, isn', ground, getting, drink, damn, chief, as, act, before, been, o, saw, any, sorry, air, job, tried, called, sold, accept, self, suddenly, more, yes, eh, far, often, make, course, under, ask, able, send, nonsense, mean, second, people, strange, four, catch, ain', white, turned, does, wish, high, become, must, heads, use, at, that', would, right, look, light, everyone, hope, has

Table 11. Black Drama: Non-American Female vs Male Authored Characters, 1950-2006: MultiNomial Naive Bayesian top 200 features by probability ratio

Appendix Seven

Female Characters, non-American male authors	Female Characters, non-American female authors
husband, dress, shame, marry, doctor, married, please, parents, oh, he', o, tired, girl, dear, really, love, gone, hurt, child, school, clothes, sister, glad, clean, happen, room, nice, sick, won', care, him, hate, sweet, couldn', hurry, aren', mother, wake, baby, daughter, didn', sleep, somebody, girls, bed, late, laugh, guess, isn', yourself, looking, miss, suppose, terrible, evening, wouldn', worse, coming, feel, don', dance, inside, stay, wanted, went, imagine, leave, stupid, meet, t, night, he, feeling, touch, door, self, beg, tonight, blame, pass, seeing, going, shouldn', cold, why, children, they', telling, lord, father, sit, sometimes, smell, lie, anybody, help, cry, pick, stop, thank, can', just, early, change, doesn', started, getting, seems, perhaps, mr, family, died, could, wonder, eat, something, it', re, beautiful, you', calling, anything, met, kind, told, crazy, person, begin, ready, friends, except, working, always, so, poor, she, months, house, knew, go, finished, woman, mad, water, son, quite, excuse, sure, till, happy, happening, sorry, want, wasn', used, yesterday, waiting, hope, talk, forget, know, too, funny, morning, home, bad, everybody, quiet, outside, drink, think, least, happened, come, worry, true, things, quick, hear, alone, tell, said, m, say, get, thought, much, everything, done, afraid, wish, none, heart, came, god, about, thing, talking, giving, better	kids, daughters, she', mothers, mama, city, husband, birth, hair, mother', daddy, parents, sitting, married, hurt, child, baby, somewhere, running, dear, mother, pregnant, daughter, tongue, dance, bed, wear, mrs, feet, clean, anymore, leaving, learn, herself, stuff, school, choose, try, takes, nowadays, gets, party, her, she, half, life, gotta, women, went, six, hit, earth, easy, foot, sister, girl, visit, house, door, stay, imagine, hate, hmm, room, own, okay, voice, miss, knew, died, doing, road, nobody, teach, cold, goes, told, gonna, forever, police, maybe, food, d, together, alone, marriage, sons, help, empty, ya, pick, die, marry, future, outside, ones, woman, sell, wait, end, education, save, comes, neither, forget, friends, yeah, it', body, angry, tree, late, father', nice, choice, street, too, touch, still, ll, move, off, fault, says, knows, you', ready, can', won', laugh, certainly, yours, free, hot, anything, bear, son, lost, just, till, anybody, each, bad, water, walk, something, minute, asked, i', whole, why, both, beat, months, long, smell, bought, side, tomorrow, myself, throw, out, m, tired, who', got, finally, day, ve, call, kind, poor, started, much, hear, taking, inside, born, love, having, found, drive, kill, eyes, wanna, hands, exactly, wouldn', seven, front, care, cause, head, gone, going, important, difference, beautiful, market, days

Table 12. Black Drama: Non-American Female vs Male Authored Characters, 1950-2006: MultiNomial Naive Bayesian top 200 features by probability ratio

Notes

[1] The experimental protocol which we have been developing for this purpose, as applied by, e.g. [Argamon et al. 2003], addresses both goals using techniques from machine learning, supplemented by more traditional computer-assisted text analysis.

[2] It is important to mention that this term occurs much more frequently than less offensive terms like negro* (3351) or positive expressions such as afr.* americ.* (230) in this sample.

[3] Supervised machine learning is a process for building a classifier from labeled training data. For example, one might present a learner with two sets of e-mails, identified by human evaluators as either spam or not spam. The learner generates a statistical model which best captures this distinction. Following training, the model generated by the learner is used to classify unseen instances, such as incoming e-mail. Supervised learning is contrasted with unsupervised learning, which is not based on preidentified class labeled training data. For example, clustering algorithms use a variety of measures of similarity to assign instances to groups.

[4] “Features” in this context means the data being used to perform the machine learning task. Each instance, typically a document or part of a document, may have an arbitrary number of features which may include word, lemma or word sequence (n-grams) frequencies as well as other elements of the data which can be computed, such as sentence length or part-of-speech frequencies.

[5] Cross validation is a technique to evaluate the likelihood that the machine learning model being generated by a learner is “over-fitted” to the training data, which would limit the effectiveness of the classifier to properly handle unseen instances. Over-fitting will also tend to weight relatively unrepresentative features too highly. Cross validation is performed by subdividing the training data into random groups (often 10), training on some of these groups and evaluating the predictions on the remainder.

[6] Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) are commonly used techniques in the machine learning community. MNB is based on calculating the prior probabilities of each feature in two or more classes. Thus, “viagra” in an MNB spam filter would be assigned as very high probability of being in a spam e-mail. For an unseen instance, the system calculates the probabilities for each feature

being a member of a particular class, adding up all of the probabilities to assign one or more classifications to the new instance. SVMs are somewhat more complex, as they attempt to divide training data into maximally separated groups by adjusting feature weights.

[7] Open source distribution and demonstrations available at [PhiloMine 2007]. PhiloMine makes use of a number of important open source packages and modules which are listed in the acknowledgments on the site.

[8] The Weka3 system refers to Sequential Minimal Optimization (SMO) which is one of a number of heuristics to allow Support Vector Machines to perform reasonably quickly and effectively.

[9] We conducted these experiments using a stand-alone parallelized implementation of the SVM-Light system[10] with PGPDT [Zanni et al. 2003], [Zanni et al. 2006] on the University of Chicago Teraport.

[10] See [Joachims 1999]. Open source distribution and documentation at http://svmlight.joachims.org/.

Works Cited

Argamon et al. 2003 Argamon, Shlomo, Moshe Koppel, Jonathan Fine and Anat Rachel Shimoni. “Gender, Genre, and Writing Style in Formal Written Texts”. *Text* 23: 3 (August 2003).

BLDR 2005 *Black Drama - 1850 to the Present*. Alexandria: Alexander Street Press, 2005. http://solomon.bld2.alexanderstreet.com/.

Joachims 1999 Joachims, Thorsten. “Making Large-Scale SVM Learning Practical”. In Bernhard Schölkopf Christopher J.C. Burges and Alexander J. Smola, eds., *Advances in Kernel Methods -- Support Vector Learning*. Cambridge, MA: MIT Press, 1999.

Koppel et al. 2002 Koppel, Moshe, Shlomo Argamon and Anat Rachel Shimoni. “Automatically Categorizing Written Texts by Author Gender”. *Literary and Linguistic Computing* 17: 4 (2002), pp. 401-412.

Parallel GPDT SerafiniT., Zanni L. and Zanghirati G. *Parallel GDPT*. http://dm.unife.it/gpdt/.

PhiloLogic 2007 University of Chicago. *PhiloLogic* . http://philologic.uchicago.edu/.

PhiloMine 2007 University of Chicago. *PhiloMine*. http://philologic.uchicago.edu/philomine/.

Ruiz and López-de-Teruel 1998 Ruiz, Alberto, and Pedro E. López-de-Teruel. “Random Falsifiability and Support Vector Machines”. Presented at *Learning '98*. (1998).

Weka3 Eibe, Frank, Hall Mark and Witten Ian. *Weka3*. http://www.cs.waikato.ac.nz/ml/weka/.

Zanni et al. 2003 Zanni, Luca, Thomas Serafini and Gaetano Zanghirati. “A Parallel Solver for Large Quadratic Programs in Training Support Vector Machines”. *Parallel Computing* 29 (2003), pp. 535-551.

Zanni et al. 2006 Zanni, Luca, Thomas Serafini and Gaetano Zanghirati. “Parallel Software for Training Large Scale Support Vector Machines on Multiprocessor Systems”. *JMLR* 7 (2006), pp. 1467-1492.