CS -597

Authorship Analysis

Guided By :

Shlomo Argamon

By:

Himanshu Singh

A20365395

hsingh34@hawk.iit.edu

Content:

- Introduction
- Data Set
- Unmasking Procedure
- Experimental Results
    - Graphs
- Conclusion
- Refrences

Introduction:

Basically, the Authorship attribution consists **of identifying the author of a given text**. Given a set of works from an anonymous author, the main concern would be to find this author by looking into internal text evidences. This task can covet three possible forms:

- **The Author Profiling** problem, where we have no idea of what kind of author has written the text to be studied, which means no set of candidates authors (in this scenario, our task would be an unsupervised learning one),
- **The Needle-in-a-haystack** problem, where we have a big set of candidates with a limited writing sample for each and given this training set we have to determine which one has written the text considered in the study and,
- **The Authorship Verification** problem, where we have one alleged author of a text, we have to verify if he/she is the actual writer(Koppel et. al., 2008).

How to Solve such a problem?

How can we solve efficiently and accurately an authorship attribution problem? Over the last century and more, a great variety of analytical methods have been applied to authorship attribution problems (cf. Juola 2008). They can be classified into three main classes of approaches:

- **the earliest, *unitary invariant,* approach**, in which a single numeric function of a text is determined to discriminate between authors,
- **the *multivariate analysis* approach**, in which statistical multivariate discriminant analysis is applied to word frequencies and related numerical features and,
- **the most recent, *machine learning* approach**, in which modern machine learning methods are applied to sets of training documents to construct classifiers to new anonymous documents.

Nowadays, the **most successful methods used to address an Authorship Attribution have been found to be the modern machine learning algorithms**, since these models present not only a high accuracy rate but also a diversity of works published helping to understand and implement them.

To address our Authorship Verification problem, I implemented machine learning methods to perform an **Unmasking Procedure,** proven to be a robust way to solve this problem. (Koppel et. al., 2007).

Objective:

The objective of this project was to **update an existing Python library** (using scikit-learn) to determine (thanks to the Unmasking Procedure) whether a document is written by a given author or not (i.e address an Authorship Verification problem). The final goal **being to determine if some editorials allegedly written by the famous American Poet Walter "Walt" Whitman (signed anonymously with pseudonyms and dated from 1857) are really products of his clever pen.**

This project could help build a new set of all the documents written by this wonderful Poet.

Data Set:

Who is Walter "walt" Whitman?

Walter "**Walt**" **Whitman** (May 31, 1819 – March 26, 1892) Born in Huntington on Long Island, was an **American poet, essayist, journalist, teacher, government clerk** and accessorily during the American Civil War **a volunteer nurse**.

Whitman is among the most influential poets in the American canon, often called **the father of free verse**. His work was very controversial in its time, particularly his poetry collection ***Leaves of Grass***, which was described as obscene for its overt sexuality.

This collection was first published in **1855** with his own money. He continued expanding and revising it until his death in **1892**.

(Source https://en.wikipedia.org/wiki/Walt_Whitman)

He allegedly has written numerous editorials for **the New York Herald, the Brooklyn Daily Eagle, and the New York Daily Tribune in 1857**, that he signed with pseudonyms. For our procedure, we used a set of texts (a total of **631 articles**) known to have been written by Whitman, and the other set containing the anonymous texts allegedly written by him (a total of **313**, separated by month of publication).

KNOWN editorials written by Whitman:

- https://drive.google.com/drive/folders/0B065CFlVfv6lY2twbF9PX09vcGs

Some unknown editorials:

- From the New York Herald (in 1857)
- https://drive.google.com/drive/folders/0B065CFlVfv6lbDUtTWlIY3lfWVU
- From the Brooklyn Daily Eagle (in 1857)
- https://drive.google.com/drive/folders/0B065CFlVfv6lRGpTLTdWaDRiWTg
- From the New York Daily Tribune (in 1857)
- https://drive.google.com/drive/folders/0B065CFlVfv6lMzhidnphRzdHbTg

Unmasking Procedure:

Why is the Authorship verification more difficult than the attribution?

Let's suppose that we want to determine if a text was written by **Shakespeare or Poe**. It would be sufficient to use their respective known writings, to construct a model distinguishing them, and to test the unknown text against the model (Koppel 2007).

On the other hand, if we need to determine if a text was written by Shakespeare or not, in the same fashion, we would have to build a model distinguishing **Shakespeare and not-Shakespeare** which it is very difficult, if not impossible to do. Simply because, it's hardly possible to assemble an exhaustive, or even representative, sample of not-Shakespeare. First of all, because a single author may consciously vary his or her style from text to text for many reasons or may unconsciously drift stylistically over time.

Thus researchers must learn to somehow distinguish between relatively shallow differences that reflect conscious or unconscious changes in an author's style and deeper differences that reflect styles of different authors.

Consequently, the fact that a particular text is more similar to the known works of Shakespeare than to those of some given set of other authors (non Shakespearian authors) does not irrefutably mean that its was written by Shakespeare. It might rather have been written by some other author not considered at all.

Basically the Authorship Verification is a real nightmare!

What if instead?

What if instead of building a model distinguishing **author A from author not-A or author known from unknown author**, given two example sets, we build two models for each set. Then, we create a method to determine whether these models were generated by the same process (same author) or by two different processes (different authors).

This new method would assess **the depth of difference** between the two example sets, the underlying idea being to test the rate of degradation of the accuracy of learned models as the best features are iteratively dropped from the learning process (Koppel et. al., 2007). This method is the Unmasking Procedure.

Classifiers used:

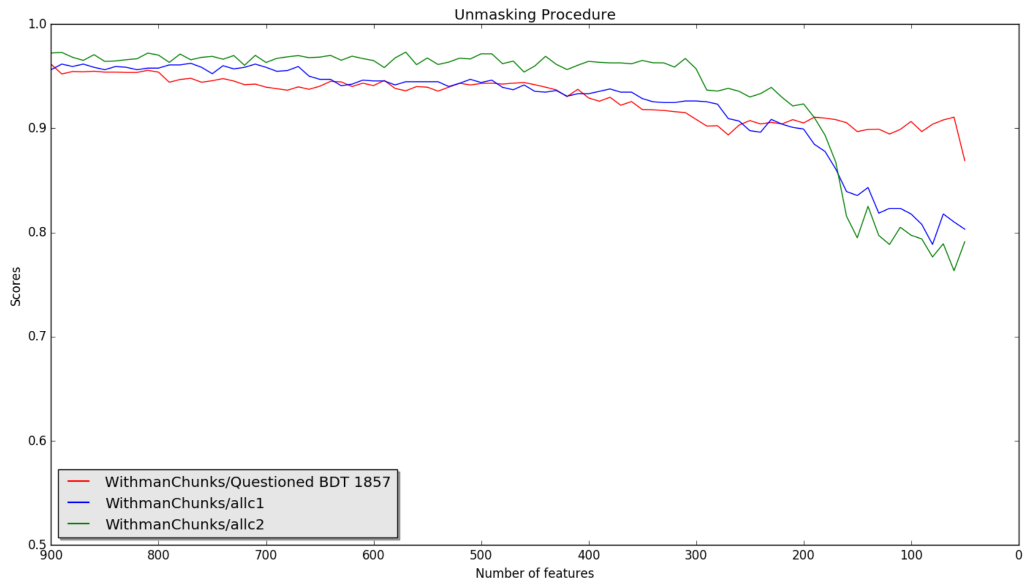
- Multinomial Naïve Bayes Classifier
- Logistic Regression Classifier
- Perceptron Classifier
- Linear Support Vector Classifier
- MLP classifier ((25,25,25,25), activation fun= tanh)
- MLP classifier ((50,50), activation fun= tanh)
- MLP classifier ((100), activation fun= tanh)

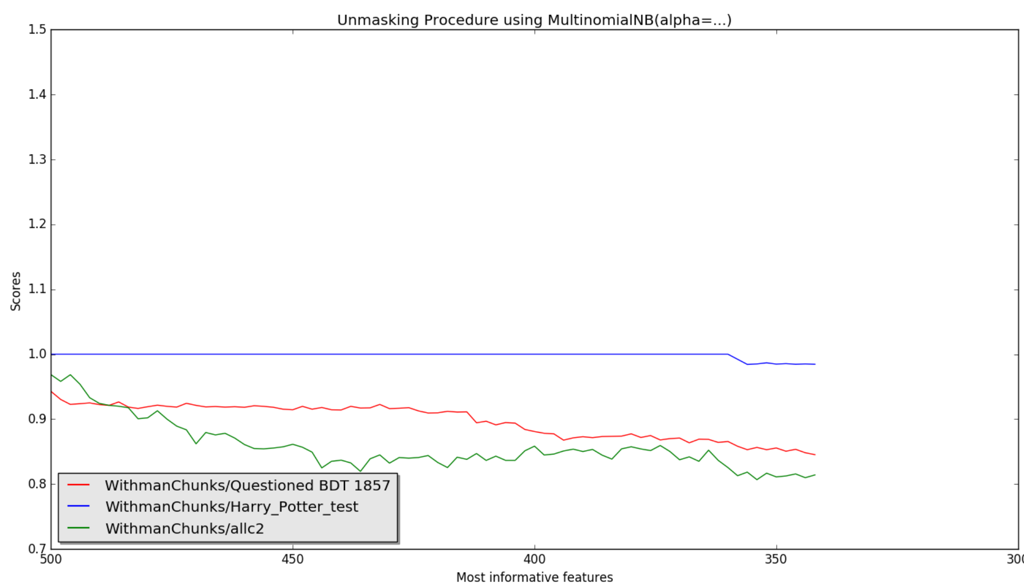
Experimental Results:

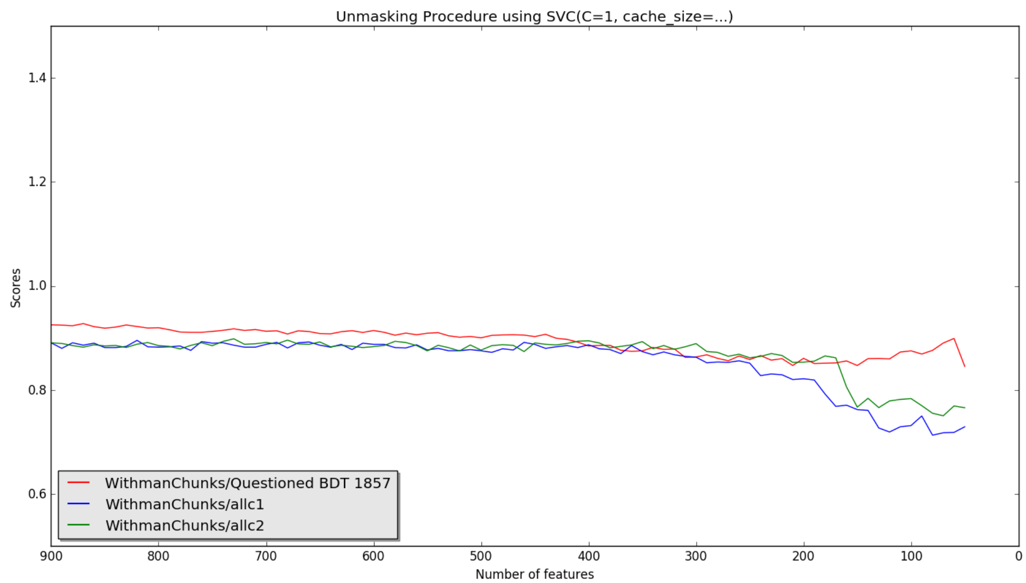
- First, a classifier is trained with the works signed by Whitman and the data set containing his allegedly writings, signed by a pseudonym.
- Second, using the 350 most frequent words as features initially, we compare the two models and assess this comparison to get an accuracy of how well a machine learning algorithm could classify them (the same or different).
- Afterwards, the top 10 most informative features, that more easily distinguish both sets, are removed and the same process is repeated.

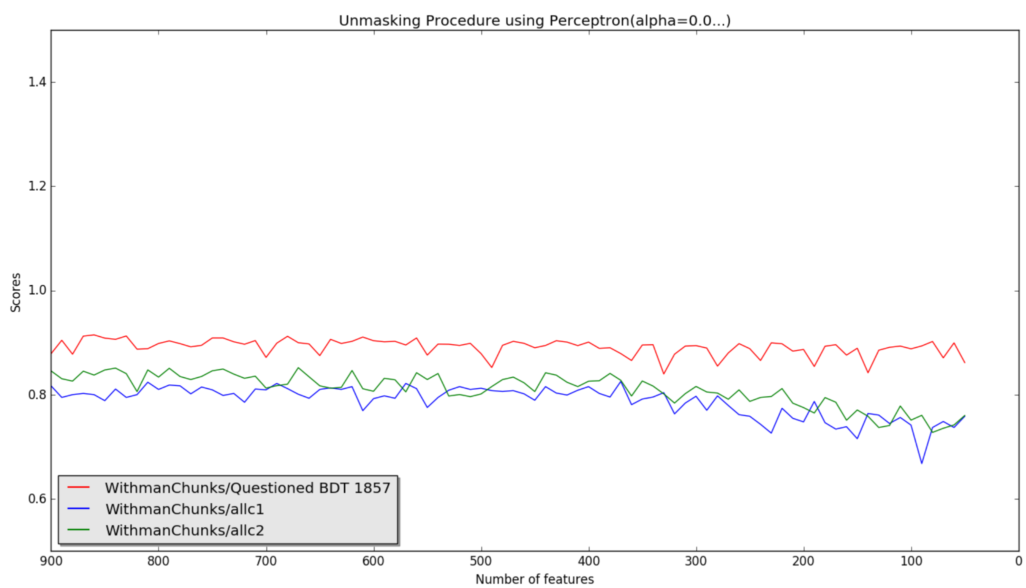
A **simple pattern of the gradual decrement** of the accuracy should point out the writings being made by Whitman or not.

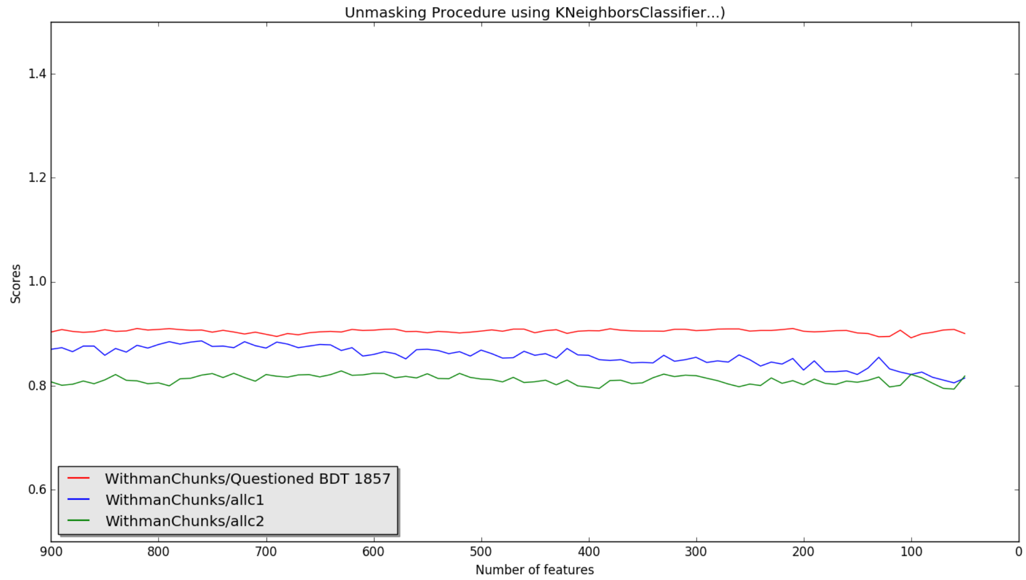
Graphs:

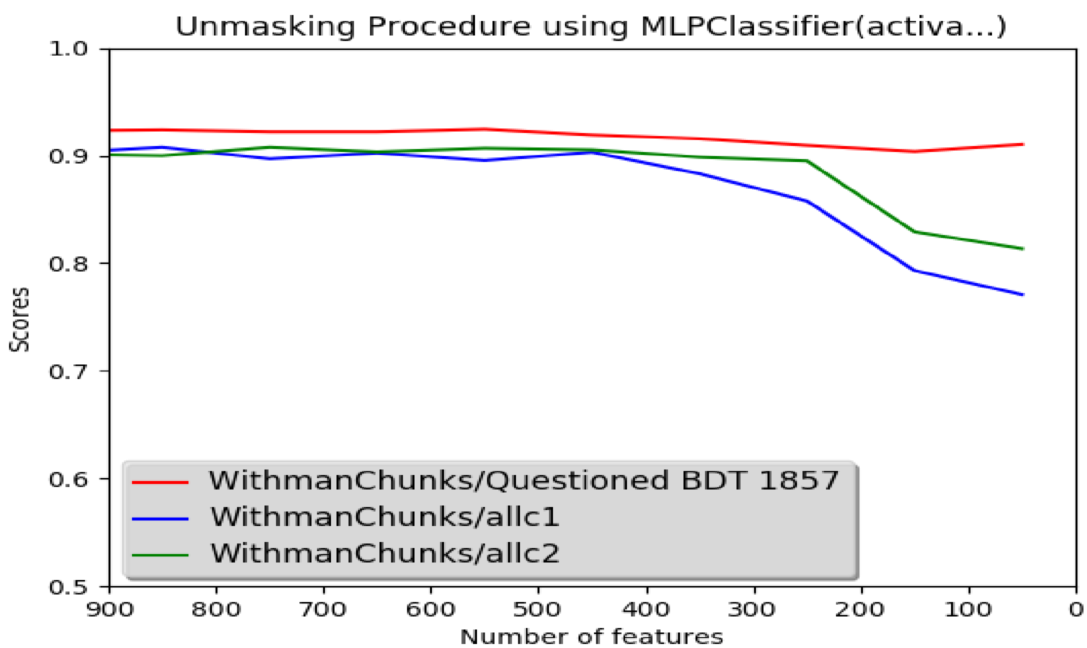

Unmasking Procedure
1.0
0.9
0.8
0.7
0.6
0.5
Scores
900
800
700
600
500
400
300
200
100
0
Number of features
WithmanChunks/Questioned BDT 1857
WithmanChunks/allc1
WithmanChunks/allc2

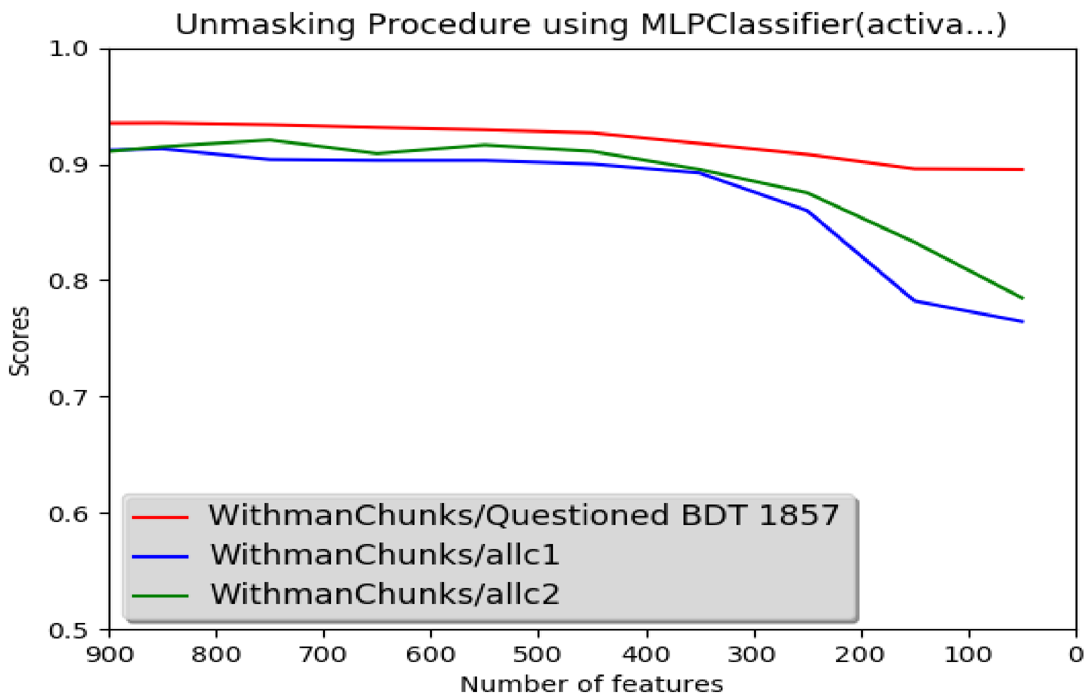

Unmasking Procedure using MultinomialNB(alpha=...)
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
Scores
500
450
400
350
300
Most informative features
WithmanChunks/Questioned BDT 1857
WithmanChunks/Harry_Potter_test
WithmanChunks/allc2

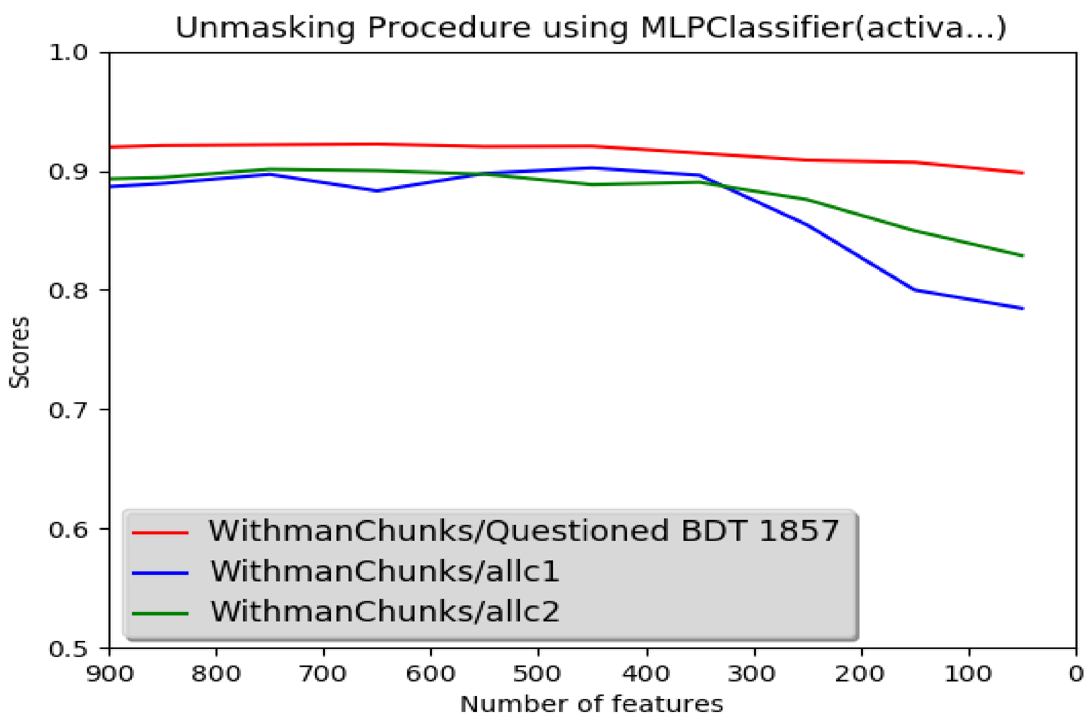

Unmasking Procedure using SVC(C=1, cache_size=...)
Scores
1.4
1.2
1.0
0.8
0.6
900
800
700
600
500
400
300
200
100
0
Number of features
WithmanChunks/Questioned BDT 1857
WithmanChunks/allc1
WithmanChunks/allc2


Unmasking Procedure using Perceptron(alpha=0.0...)
Scores
1.4
1.2
1.0
0.8
0.6
900
800
700
600
500
400
300
200
100
0
Number of features
WithmanChunks/Questioned BDT 1857
WithmanChunks/allc1
WithmanChunks/allc2


Unmasking Procedure using KNeighborsClassifier...)
Scores
Number of features
WithmanChunks/Questioned BDT 1857
WithmanChunks/allc1
WithmanChunks/allc2


Unmasking Procedure using MLPClassifier(activa...)
Scores
Number of features
WithmanChunks/Questioned BDT 1857
WithmanChunks/allc1
WithmanChunks/allc2


Unmasking Procedure using MLPClassifier(activa...)
1.0
0.9
0.8
0.7
0.6
0.5
Scores
WithmanChunks/Questioned BDT 1857
WithmanChunks/allc1
WithmanChunks/allc2
900
800
700
600
500
400
300
200
100
0
Number of features


Unmasking Procedure using MLPClassifier(activa...)
1.0
0.9
0.8
0.7
0.6
0.5
Scores
WithmanChunks/Questioned BDT 1857
WithmanChunks/allc1
WithmanChunks/allc2
900
800
700
600
500
400
300
200
100
0
Number of features

References:

- **Rosa María C., Luis V., Manuel M. & Paolo R., *Authorship Attribution using Word Sequences*, Retrieved from http://users.dsic.upv.es/~prosso/resources/CoyotlEtAl_CIARP06.pdf on 11/22/2016**

- **Moshe K., Jonathan S. & Shlomo A., *Computational Methods in Authorship Attribution*, Dept. of Computer Science, Bar-Ilan University (Ramat-Gan, Israel) and Dept. of Computer Science, Illinois Institute of Technology (Chicago, USA)**

- **Moshe K., Jonathan S. Elisheva B., *Measuring Differentiability: Unmasking Pseudonymous Authors,* Dept. of Computer Science, Bar-Ilan University (Ramat-Gan, Israel)**

- ***Walt Whitman (Article),* Retrieved from *https://en.wikipedia.org/wiki/Walt_Whitman on11/22/2016***

Questions ?