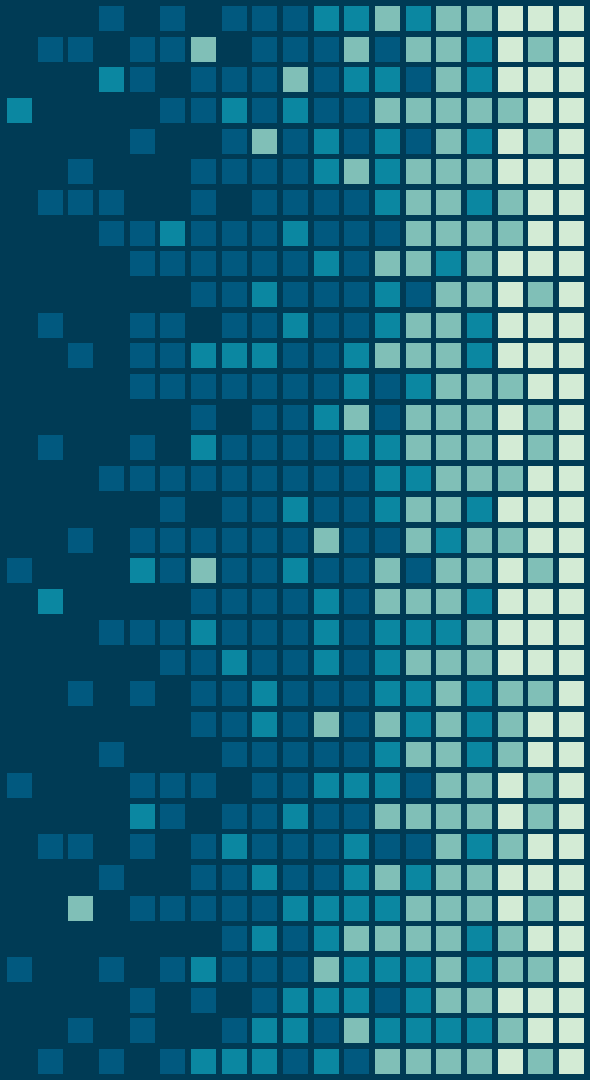
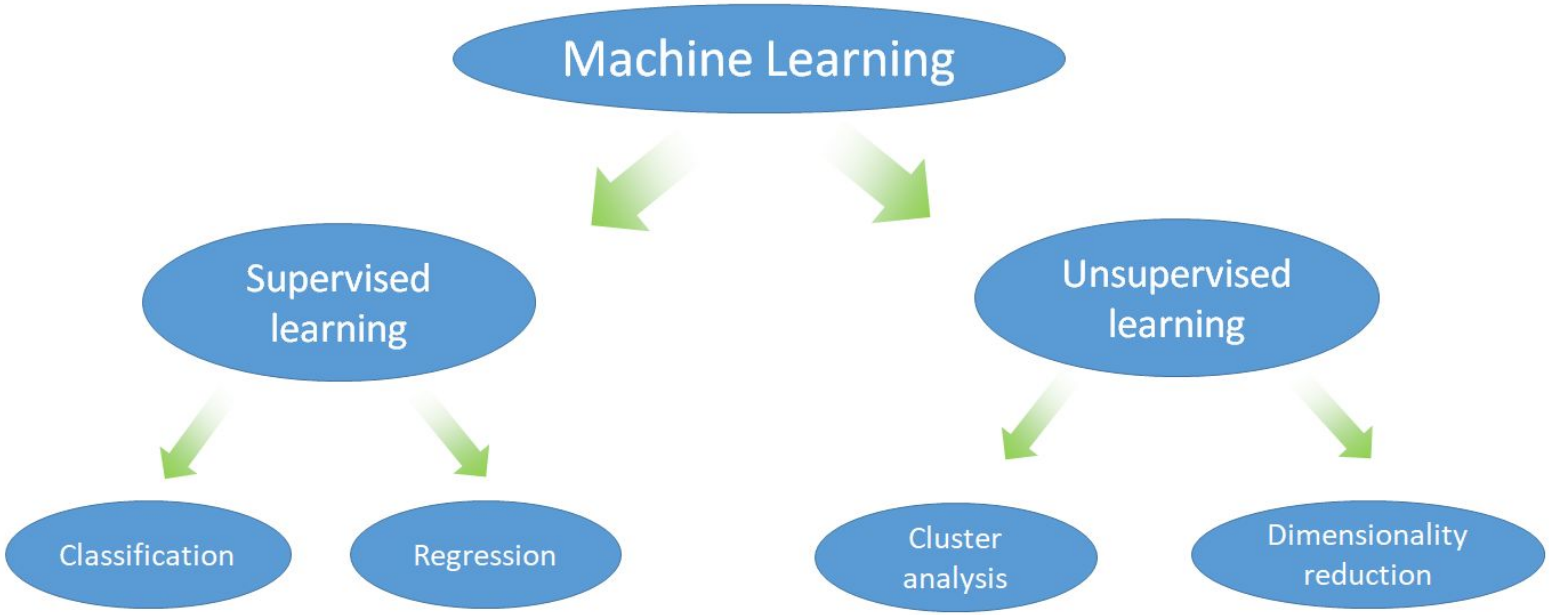
Introduction to classification

Joanna Byszuk & Jeremi Ochab
DHSI 2024
"DIY Computational Text Analysis with R"
Day One



Classification in Machine Learning


Machine Learning
Supervised learning
Unsupervised learning
Classification
Regression
Cluster analysis
Dimensionality reduction

Unsupervised classification

Unsupervised learning

is a set of techniques that allow you to infer models to **extract knowledge of data sets where a priori is unknown**.

E.g.

- Cluster analysis
- Dimensionality reduction (PCA, MDS)

Supervised classification

Supervised learning

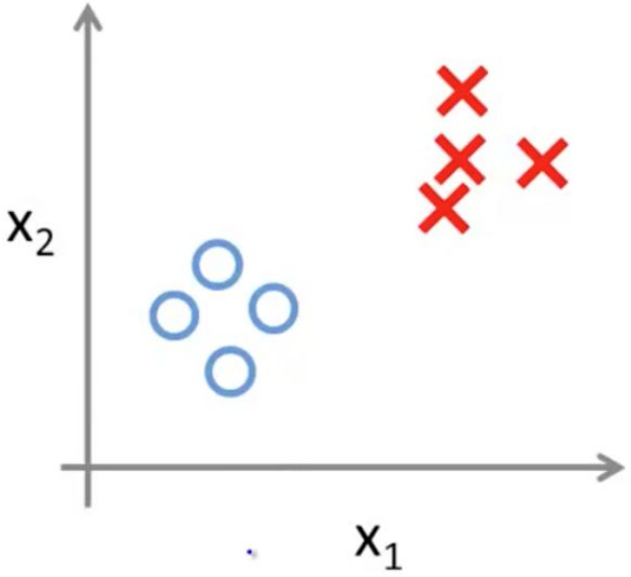
is a set of techniques that allows **future predictions** based on **behaviors or characteristics analyzed in historical data**.

E.g.

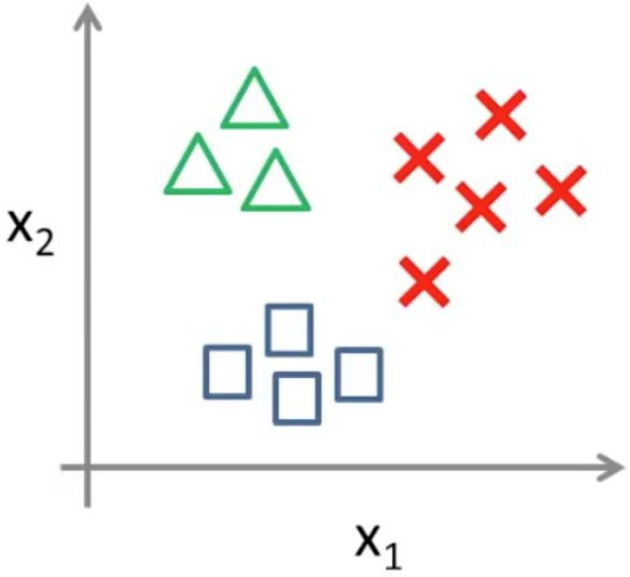
- Regression algorithms (linear regression, neural networks)
- Classification algorithms (logistic regression, Naive Bayes, Support Vector Machines, Random Forest)

Types of classification

Binary classification:



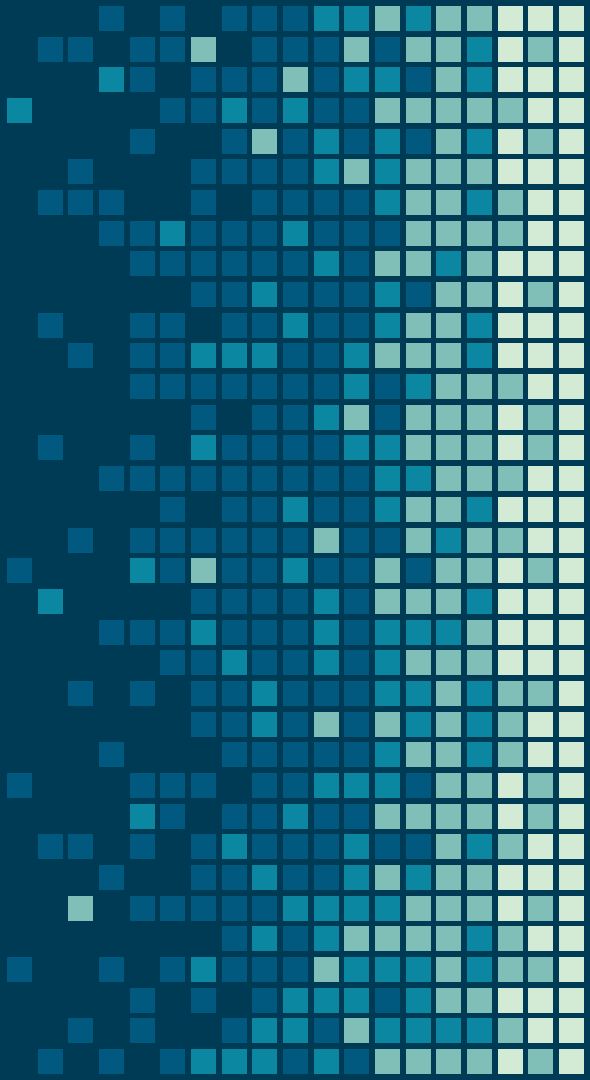
Multi-class classification:



Cross-validation

- Verifying the quality by performing classification on a series of subsets of the main corpus
- E.g. “leave-one-out” = doing classification for all the cases of “corpus – one text”, comparing results

Stylometry in authorship attribution – a classification task



Attribution vs verification

Attribution:

- Determining who of the known candidates from the closed set authored a given work

Verification:

- Determining **IF** one of the known candidates from the open set authored a given work

Classic authorship problems

Federalist papers, JK Rowling

Federalist Papers as an attribution case

- "A series of essays, anonymously published defending the document to the public"

(Lin-Manuel Miranda 2015)

- 85 texts authored by: Alexander Hamilton (51?), James Madison (29?) and John Jay (5)
- 12 letters of disputed authorship determined by stylometry

Mosteller, F., and D. L. Wallace (1964). Applied Bayesian and Classical Inference: The Case of The Federalist Papers. (and numerous other studies)

JK Rowling or Robert Galbraith?

- Who wrote “The Cuckoo's Calling”?
- Study by Patrick Juola (2013)

  “comparing against Rowling's own The Casual Vacancy, Ruth Rendell's The St. Zita Society, P.D. James' The Private Patient and Val McDermid's The Wire in the Blood…. Of the 11 sections of Cuckoo, six were closest (in distribution of word lengths) to Rowling, five to James.”

- Confirmed by the author

Our study

Questions:

1) Who wrote the ending?
2) Did the anonymous writer introduce changes to the rest of the play?

Setting up the experiment

Dataset

Problems we had to face:

- Small availability of Spanish (historical) texts
- Available corpus imbalanced in terms of:
  - author representation
  - gender
  - genre
  - nationality

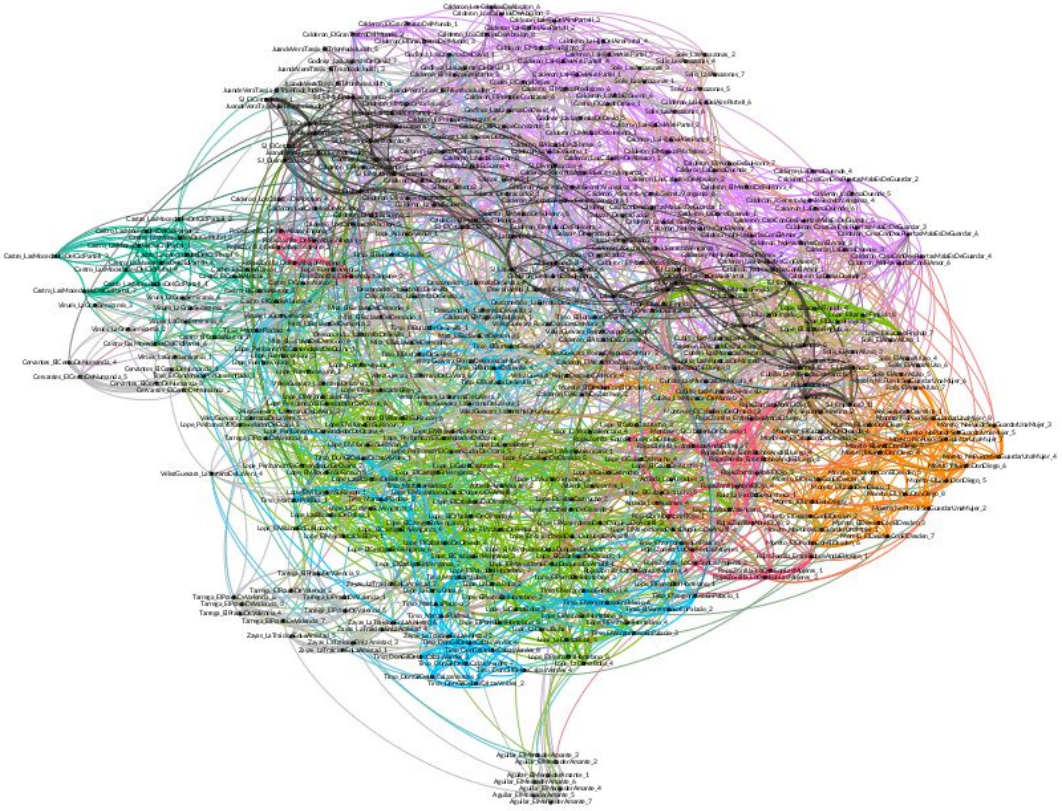
Dataset

Adopted solutions:

- *SC* and SJ's plays extracted from digital editions (Schmidhuber de la Mora, 2016; Cervantes Virtual Library).
- Poor OCR results – > transcription of Salazar's texts.
- Use of Canon-60 corpus (Oleza 2014), but just one genre: "comedia de capa y espada".

Results

Literary landscape



Authorship attribution vs verification

We know the author must be one of a few candidates

Relatively easy

e.g. classify(), rolling.classify()

We don't know if we have the author in our dataset

Still quite difficult

e.g. imposters()

Authorship attribution vs verification – our case

- cross-validated classification with SVM, NSC, Delta
- verification with “Imposters method” (Kestemont et al., 2016; Koppel and Winter, 2014)

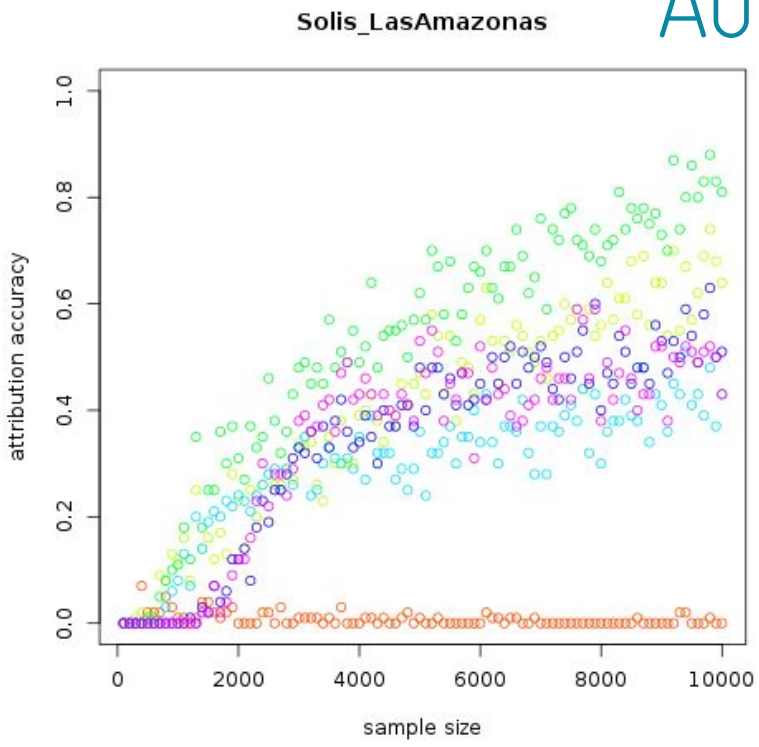
results inconclusive, pointed authors:

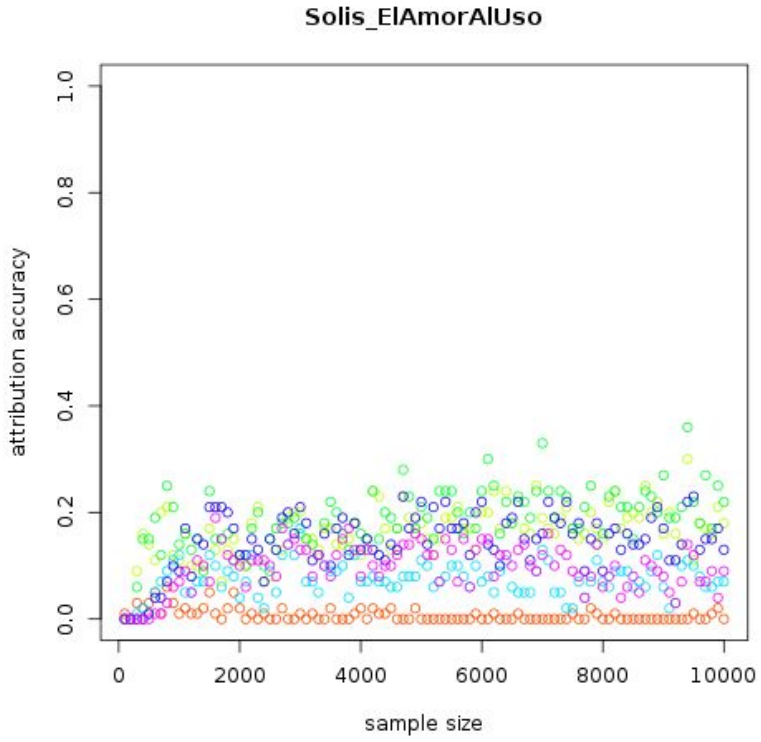
Calderón and Moreto to SJ, Solís and de Vera Tassis.

Adjusting – pruning the corpus

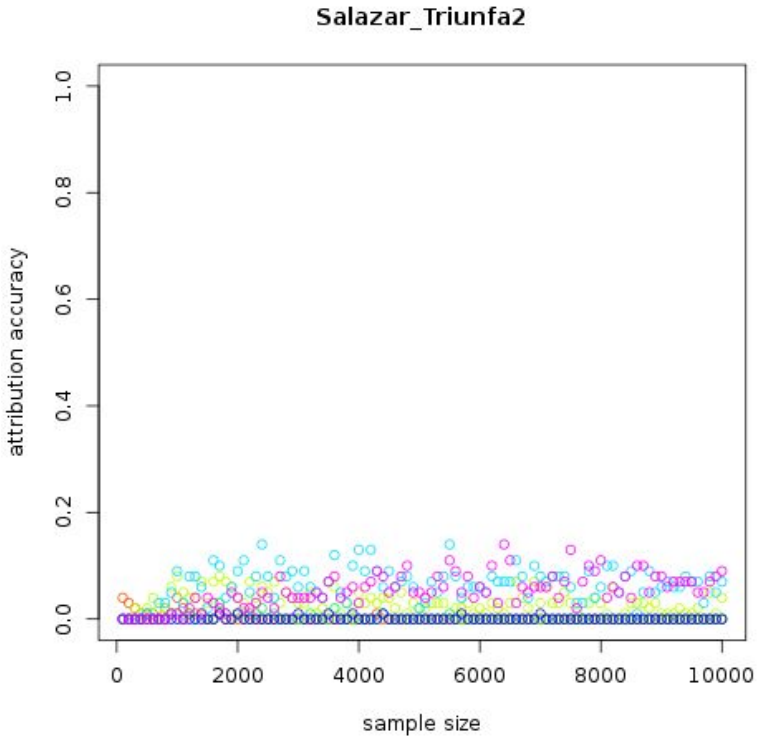
- removing ‘landscape’ authors who could not author the play
- determining strength of authorial signal (Eder 2017)

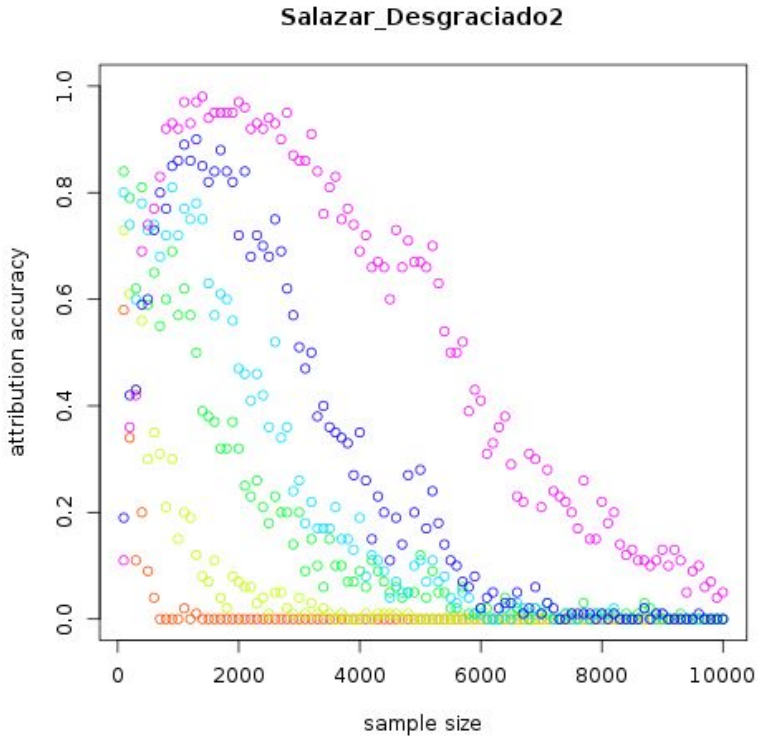
Authorial signal strength



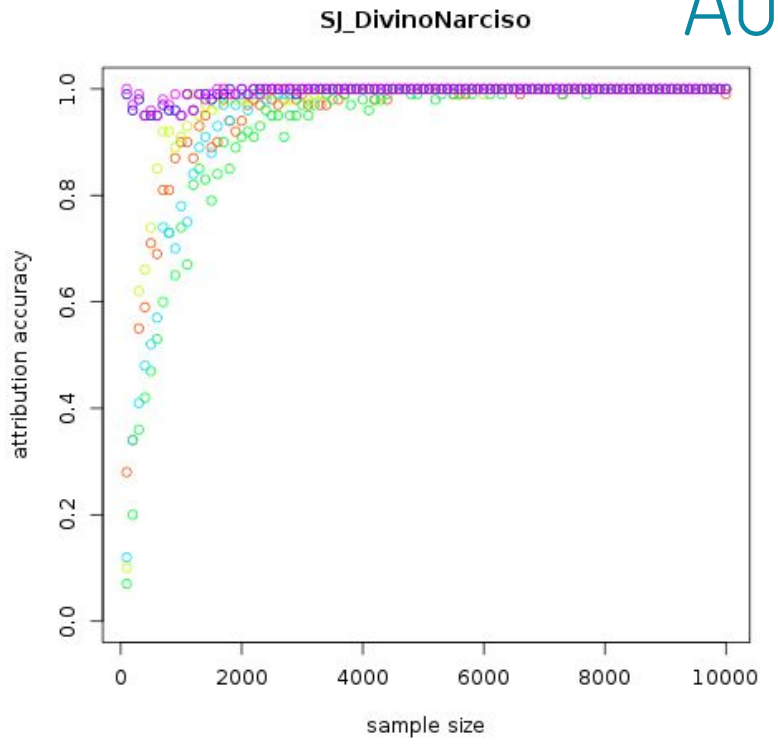


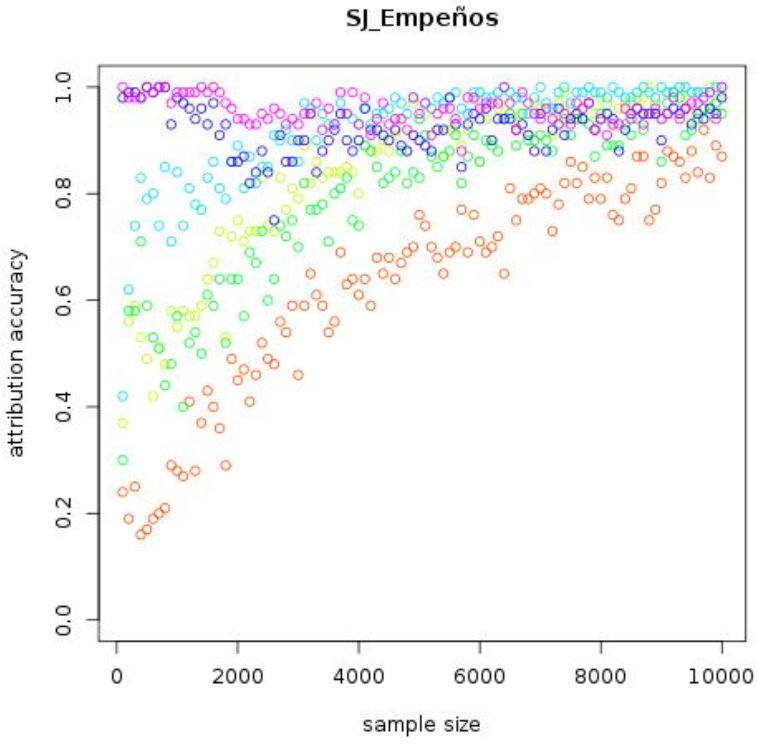
Authorial signal strength





Authorial signal strength





Revised approach

Cross validation & classification on just these three authors:

- SJ attributed as the author in almost all settings
- some results point to Solís influence in the last two thousand words
- the most reliable results: SVM and 100-500 MFWs range (from 54.8% to 81.2% accuracy, with the average of 72.75%)

Rolling Classify

- Problem that requires detecting multiple authorial voices
  - Salazar's voice and the anonymous author
  - Use of Rolling Classify (Eder 2016)

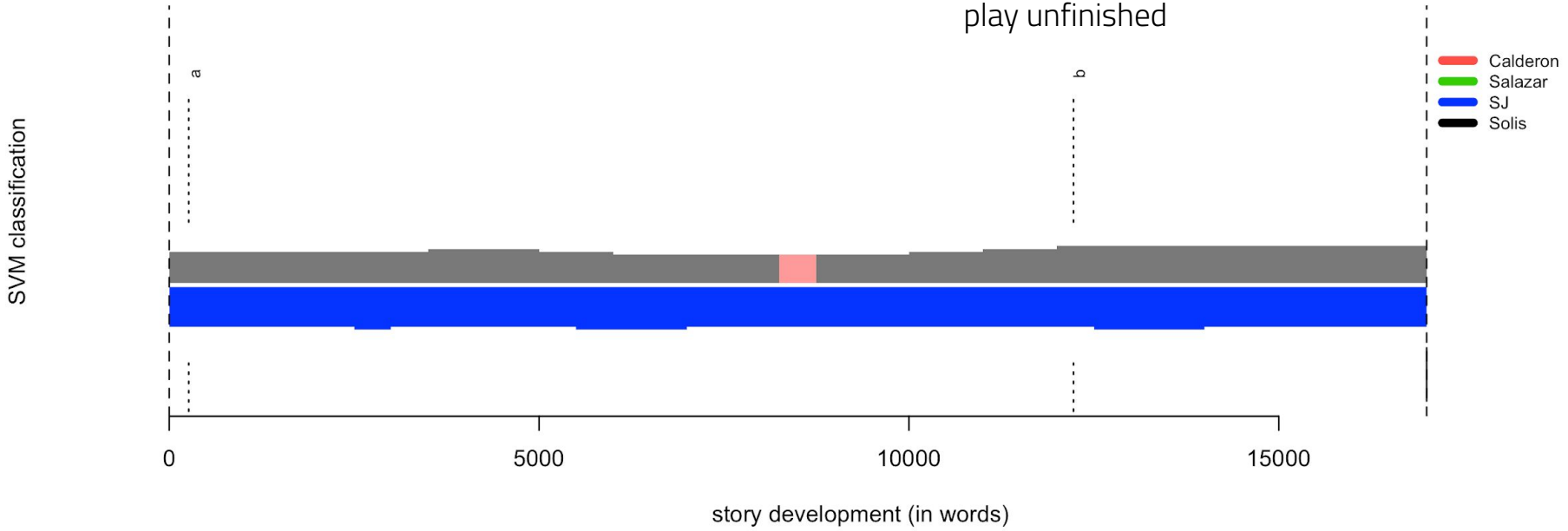
Rolling Classify

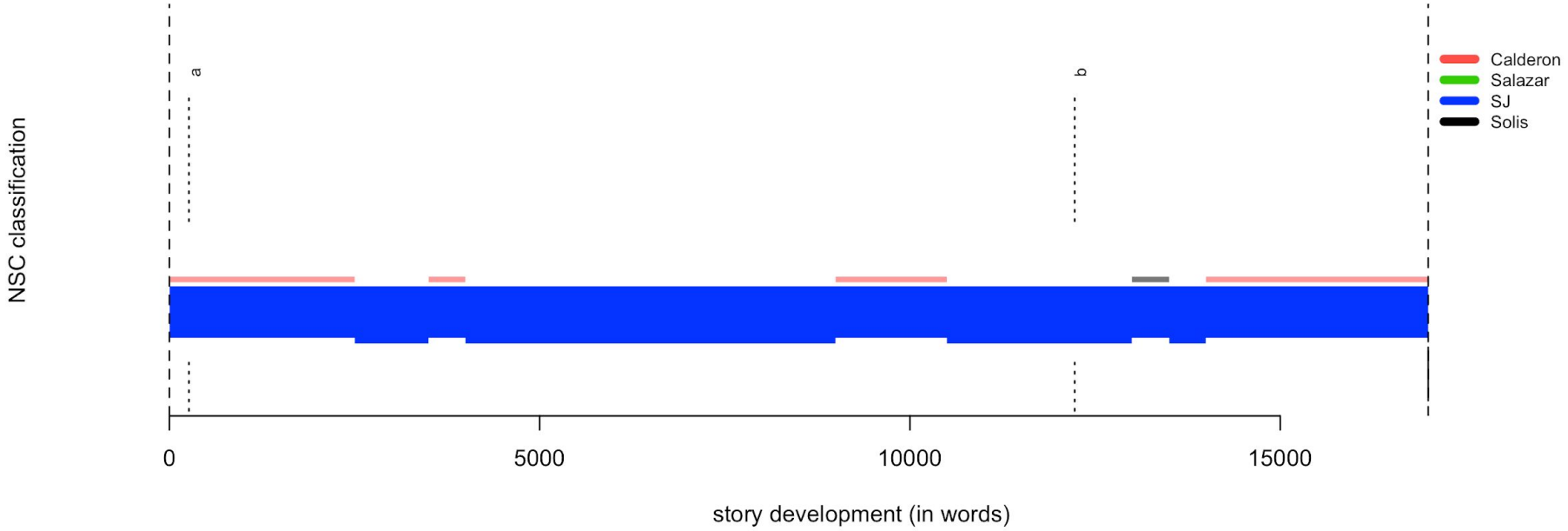
Experiment conditions:

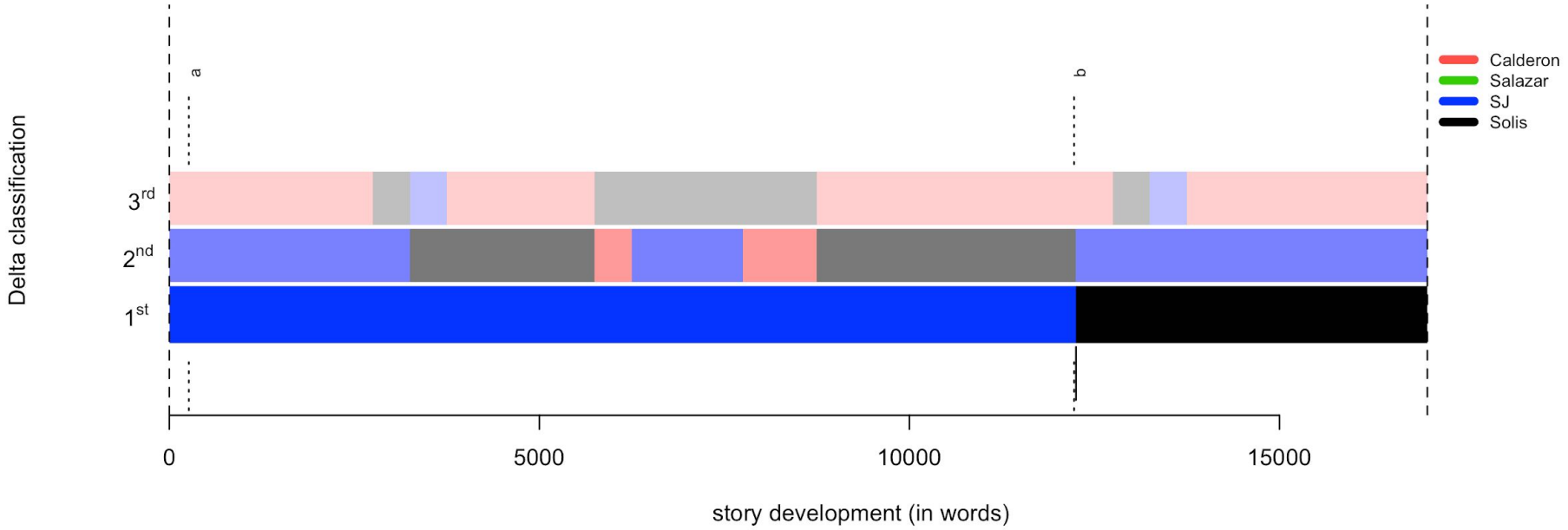
- SVM, NSC and Delta
- 500 MFW
- 5000 words-per-slice
- Authors: candidates (Salazar and SJ), control authors (Calderón and Solís)

beginning and the first encounter of protagonists, doña Beatriz and don Juan → very feminist confrontation

where Salazar left the play unfinished




NSC classification
a
b
Calderon
Salazar
SJ
Solis
0
5000
10000
15000
story development (in words)


Delta classification
a
b
3rd
2nd
1st
Calderon
Salazar
SJ
Solis
0
5000
10000
15000
story development (in words)

Who writes TV shows?

9 seasons of Doctor Who = 26 screenwriters

What about directors?

Script editors?

… actors?

Showrunner – the producer who actually [holds] ultimate management and creative authority for the program (...) the boss, the hard-nosed operational manager responsible for all creative aspects of a show and who runs interference with the studios and the networks.

Shane Bannon

https://www.theaustralian.com.au/arts/getting-the-run-of-the-place/news-story/78b66658bc660728f73a2def314a3788

Doctor Who (1963 –)

1963-89 ‘Classic’ series

- focus on the main character

1996 Film

2005- ? revival of the show: New/Nu series

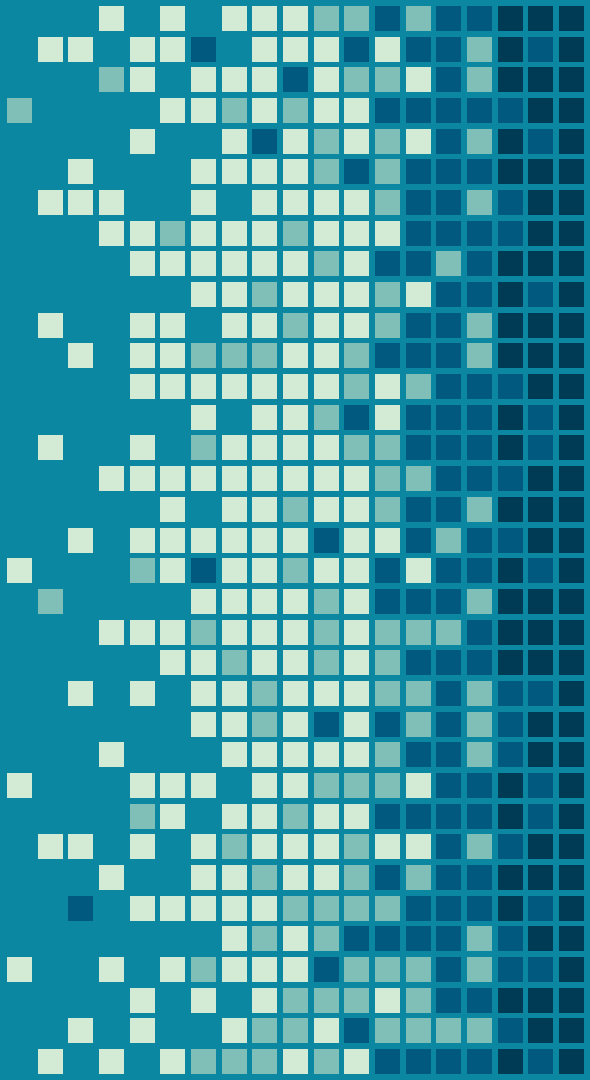
- transition to authorial – American model, increased role of a showrunner

I might change at least 30 per cent of the material, often 60 per cent, sometimes almost 100 per cent. I go over every line of dialogue, either adding new stuff or refining what's there; sometimes that means enhancing a line that the original writer hasn't realised is good. I'll bring out themes, punch up moments, add signature dialogue, clarify stage directions and make cuts. To every single scene, if need be. Usually, the basic shape remains intact, but sometimes I'll invent brand new characters and subplots... while at the same time remaining faithful to the original writer. I'll even impersonate them.

Russell T Davies

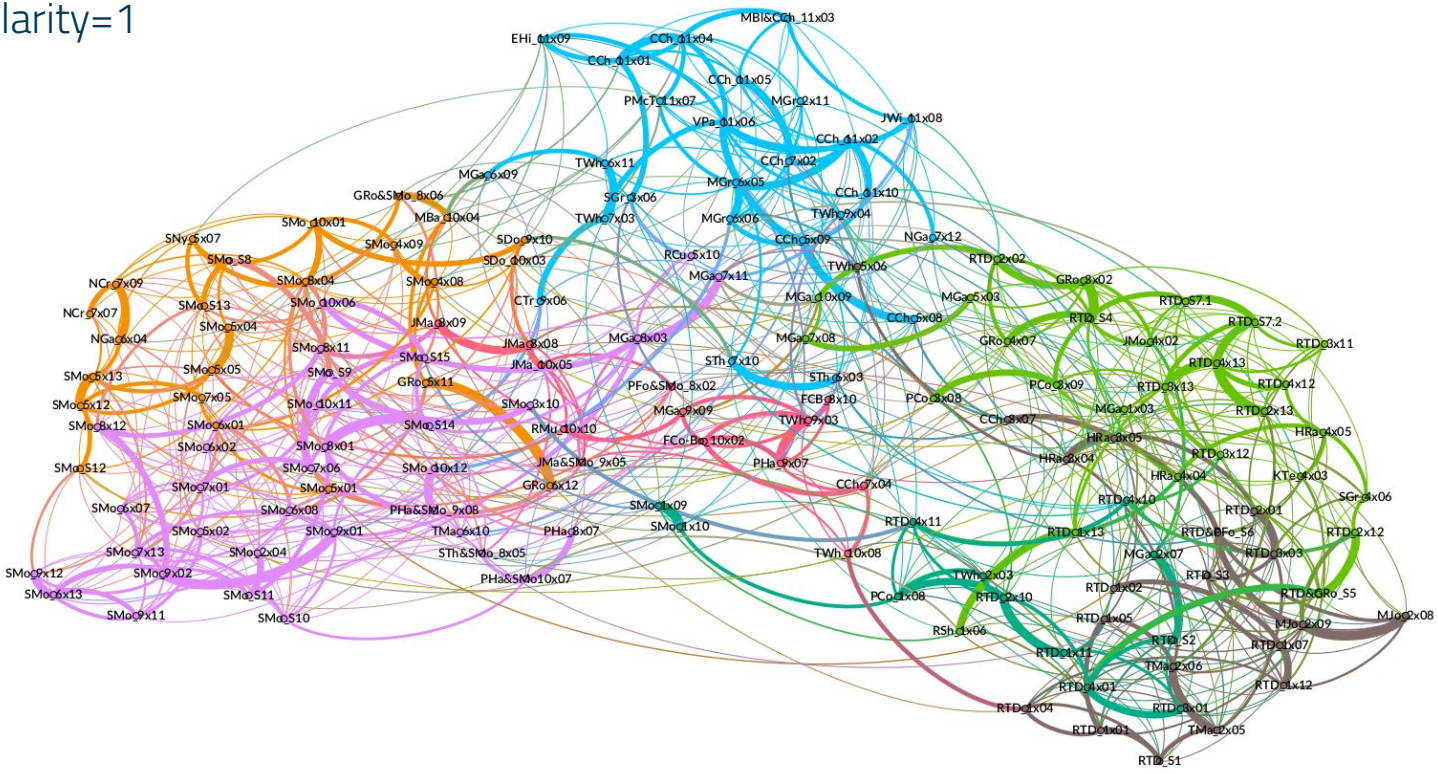
what you're there to do is to say in different ways every single day, to send out a consistent message: this is the show we're trying to make, it's like this.

Steven Moffat

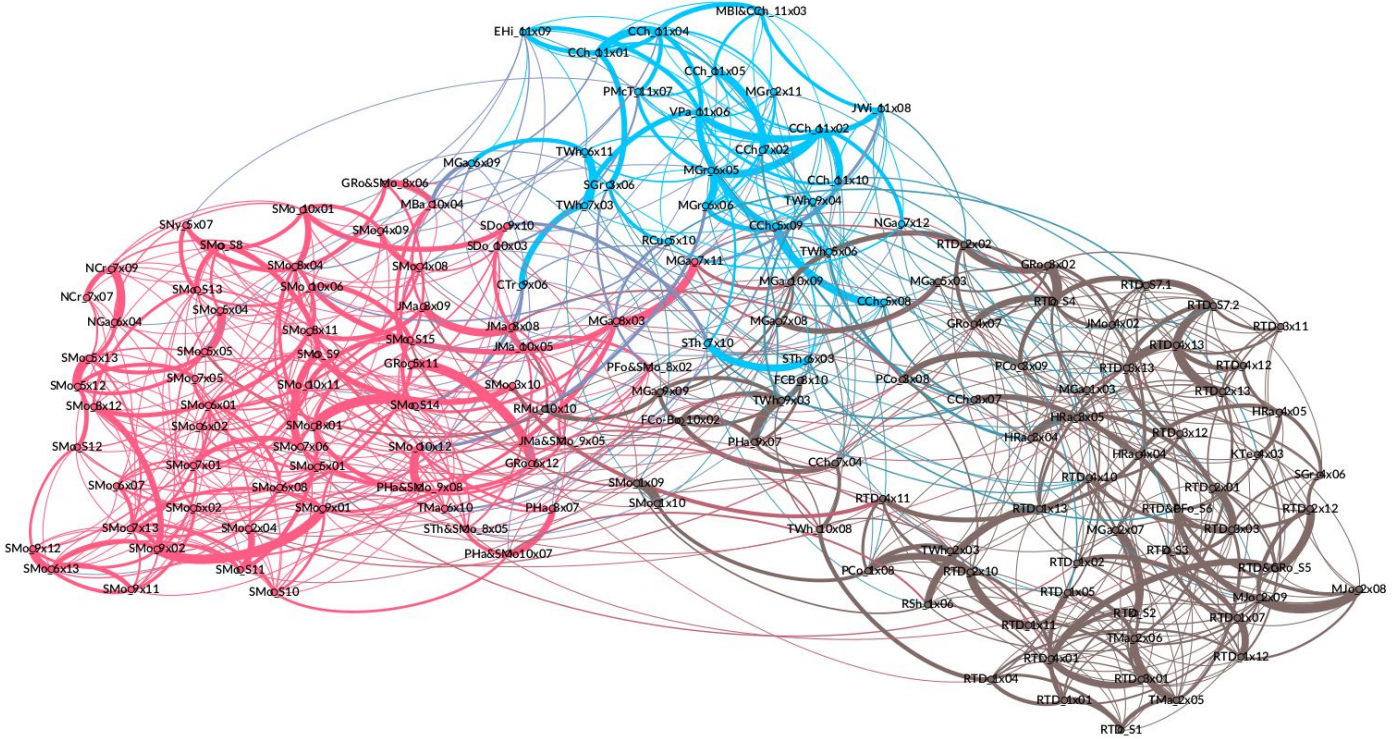


Authorial influence in Doctor Who

Writer clusters in Nu Who as detected by Modularity=1



Writer clusters in Nu Who as detected by Modularity=3



Observations

Russell T Davies – present in all (thematic) communities in his era

Steven Moffat – clustering with himself even when writing for others

Chris Chibnall – dubious, his work + other writers who became prominent on their own

Showrunners in detail

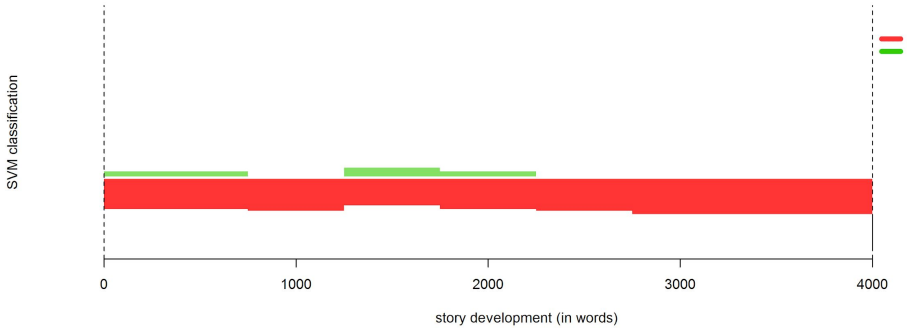
Rolling Delta

- Combined classification of smaller samples of the texts, added together as “moving window”
- RQ: Is the strength of showrunner’s impact the same?

Research questions

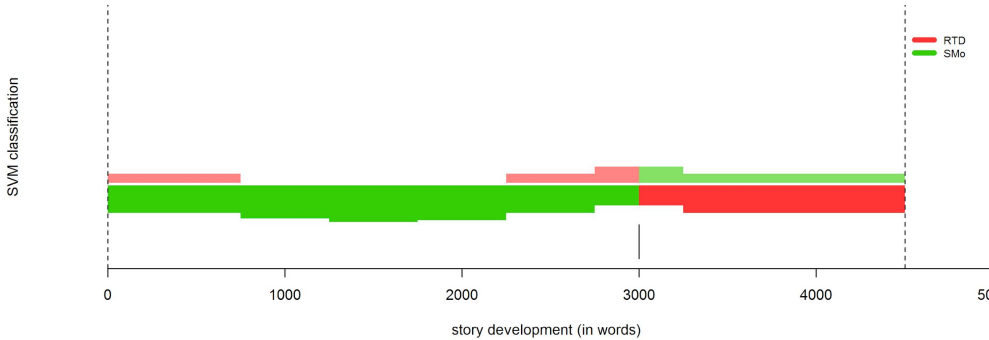
Showrunner’s signal usually stronger than writers’ and seasons’

Davies as a showrunner

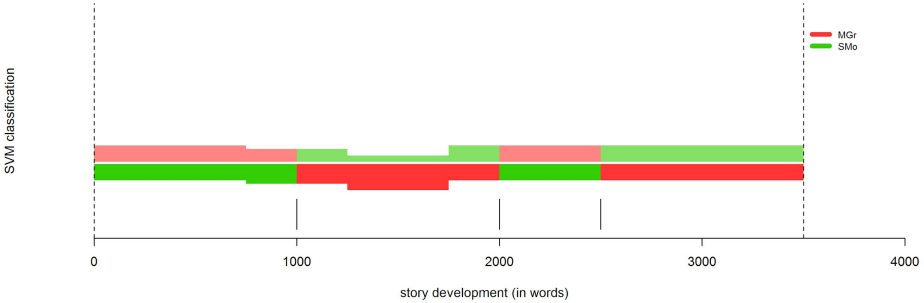


The Doctor's Daughter (4x06) contrasted sequentially against R.T Davies (red) and M. Gatiss (green); 100 most frequent words.

Silence in the Library (4x08) contrasted sequentially against R.T Davies (green) and S. Moffat red); 100 most frequent words.

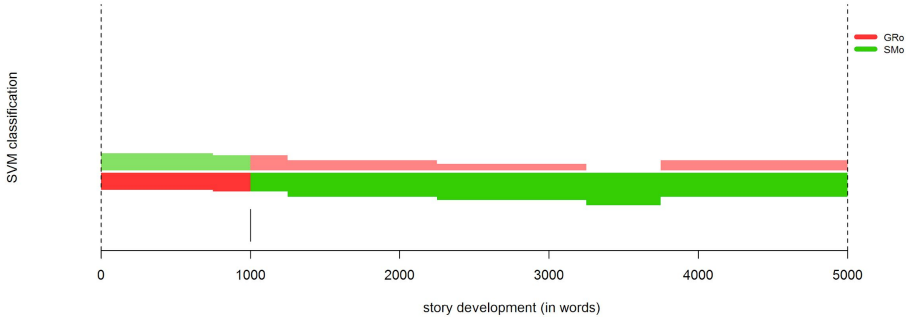


Moffat as a showrunner



The Rebel Flesh (6x05) contrasted sequentially against S. Moffat (green) and M. Graham (red); 100 most frequent words.

The Caretaker (8x06) contrasted sequentially against S Moffat (green) and G. Roberts (red); 100 most frequent words. This episode credits them both as authors.



Observations

Russell T Davies appears to be more engaging and editing / executing influence

Research questions

Observation from own 2017 study “Voices of Doctor Who”:

Showrunner’s signal usually stronger than writers’ and seasons’

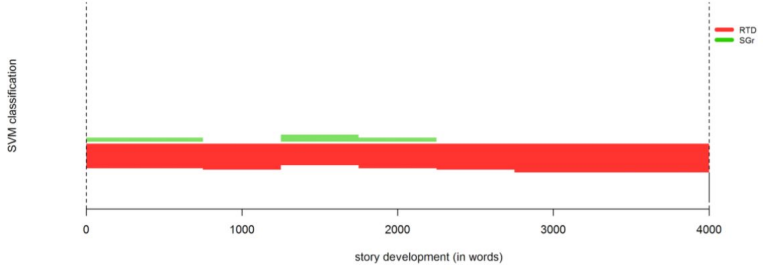


Fig. 35 “The Doctor’s Daughter” (4x06) contrasted sequentially against R.T. Davies (red) and S. Greenhorn (green); 100 most frequent words.

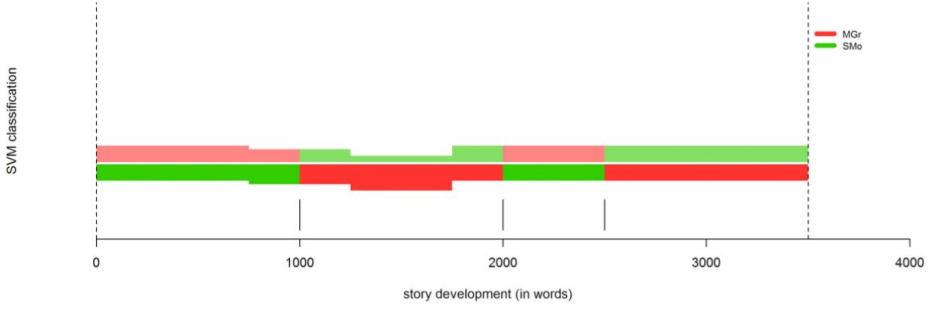


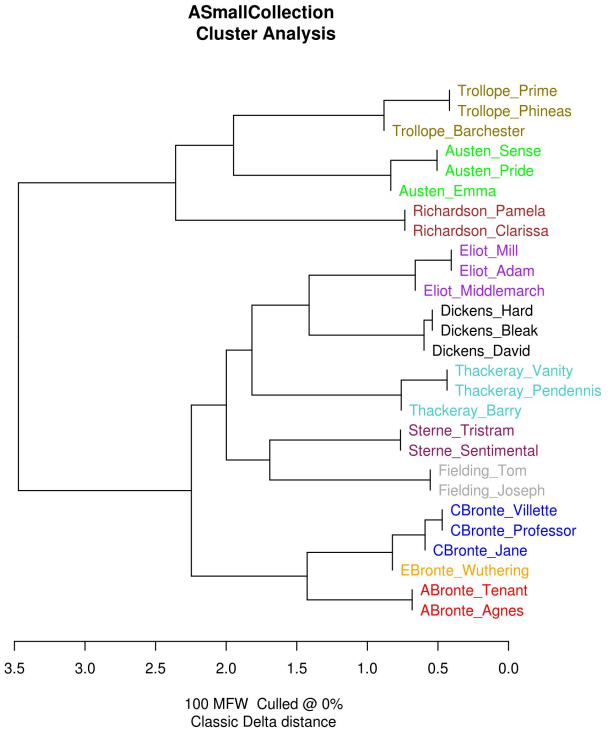
Fig. 41 “The Rebel Flesh” (6x05) contrasted sequentially against S. Moffat (green) and M. Graham (red); 100 most frequent words.

Classification in stylo

Classification algorithm

- k-nearest neighbors (v. good)
- Support Vector Machine (best)
- Nearest Shrunken Centroids
- Naive Bayes

What you've already done



Classify

- It trains a model for pre-defined groups of texts, e.g. authors.
- Then it computes distances (differences) between texts, …
- … represented as rows of frequencies of most frequent words.

Classify

- compares the trained models with test texts, using:
  - **Delta classifier** (lazy learner introduced by Burrows)
  - **k-NN classifier** (lazy learner relying on >1 neighbors)
  - **Support Vector Machines**, a high-performance non-probabilistic classifier
  - **Naive Bayes**, a classical yet slightly outdated classifier
  - **Nearest Shrunken Centroids**, a classifier for high-dimensional datasets
- A final report of the classifier's performance is outputted

Classify

Different structure:

- primary_set
- secondary_set

Running:

- library(stylo)
- classify()

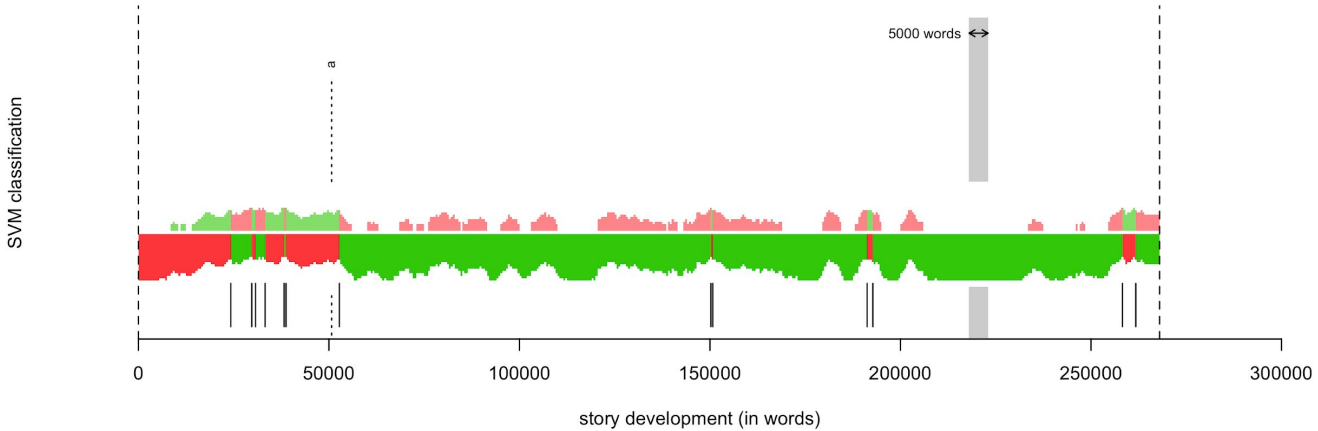
Good idea to save results in a new variable, e.g.

- results = classify()

Rolling classify

- Looks for traces of authors in a co-authored text…
- … by sliding through this text sequentially in order to detect peculiarities.
- Produces a graph of the respective strengths of these traces.

Rolling classify



Red = Guillaume de Lorris

Green = Jean de Meun

The **thickness** of the bottom stripe indicates **certainty of classification** and a *vertical dashed line* the *commonly-accepted division*.

Rolling classify

Different subfolder structure:

- reference_set (individual writings)
- test_set (collaborative text)

Rolling classify

Running the function:

- library(stylo)
- Example:

  rolling.classify(write.png.file = TRUE,
  classification.method = "svm", mfw = 100,
  training.set.sampling = "normal.sampling",
  slice.size = 5000, slice.overlap = 4500)