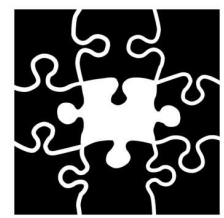
Literary Text Mining and Stylometry

DH Crash Course

Andreas van Cranenburgh



Huygens ING
Royal Netherlands Academy of Arts and Sciences

Institute for Logic, Language and Computation
University of Amsterdam

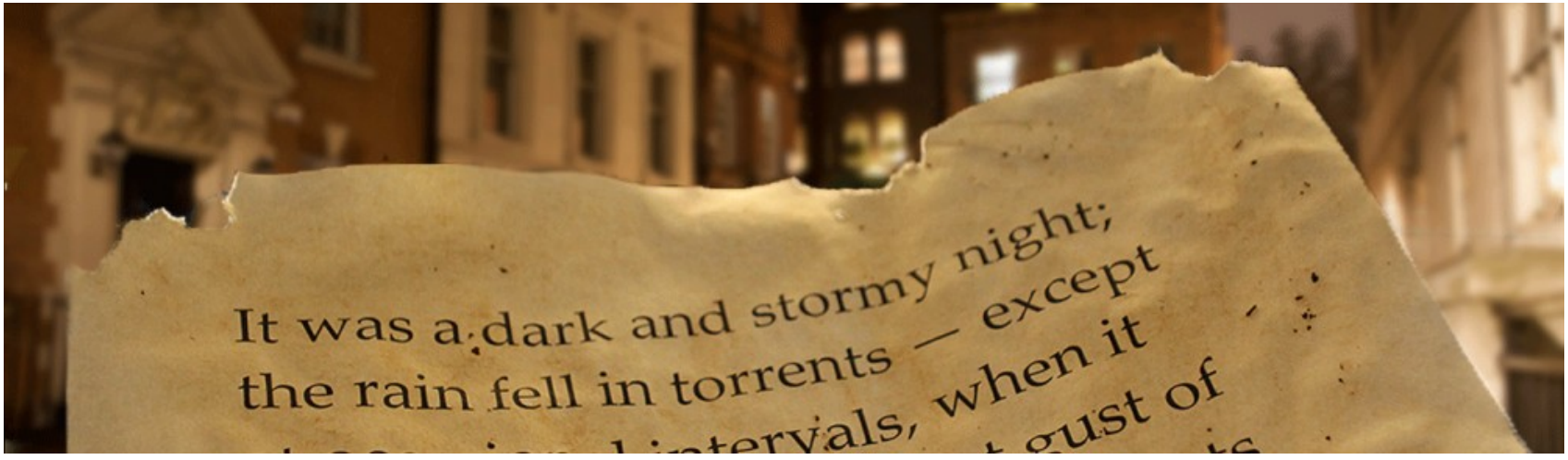
March 23, 2014

Amsterdam, 2014

Today's menu

1. "The Riddle of Literary Quality" project
2. Machine Learning
3. Your Mission

The project



The Riddle of Literary Quality*

*http://literaryquality.huygens.knaw.nl

Literary Quality: "low" versus "high" brow

Perceptions of literary quality due to:

- Social factors?
- Contextual factors?
- Individual factors?

Literary Quality: "low" versus "high" brow

Perceptions of literary quality due to:

- Social factors?
- Contextual factors?
- Individual factors?
- Textual characteristics?

Main research question

Survey: Two independent axes of quality:

1. good vs. bad
2. literary vs. non-literary

Main research question

Survey: Two independent axes of quality:

1. good vs. bad
2. literary vs. non-literary

Texts: Two kinds of text features:

1. low-level: directly extracted from text (e.g., sentence length)
2. high-level: analyze text with some model (e.g., deep syntactic structures)

Main research question

Survey: Two independent axes of quality:

1. good vs. bad
2. literary vs. non-literary

Texts: Two kinds of text features:

1. low-level: directly extracted from text (e.g., sentence length)
2. high-level: analyze text with some model (e.g., deep syntactic structures)

Question

Can we find correlations between quality judgments and text features?

Corpus

- 401 modern Dutch novels
- Published 2007–2012
- Selected by popularity

Survey

- Large reader survey
- Subjects select books they read from the corpus, and rate whether the book is good, literary
- about 14,000 readers completed the survey

Today's menu

1. "The Riddle of Literary Quality" project
2. Machine Learning
3. Your Mission

The Workflow

Definition

Text classification:

$$\text{Text} \Rightarrow \text{Features} \Rightarrow \text{Model} \Rightarrow \text{Predictions}$$

The Workflow

Definition

Text classification:

$$\text{Text} \Rightarrow \text{Features} \Rightarrow \text{Model} \Rightarrow \text{Predictions}$$

- Goal: generalization

Today's menu

1. "The Riddle of Literary Quality" project
2. Machine Learning
   - Features
   - Model
   - Predictions
   - Background
3. Your Mission

Feature vectors

Definition

Vector: a sequence of numbers

Feature vectors

Definition

Vector: a sequence of numbers

Each text will be represented by a vector of numbers.

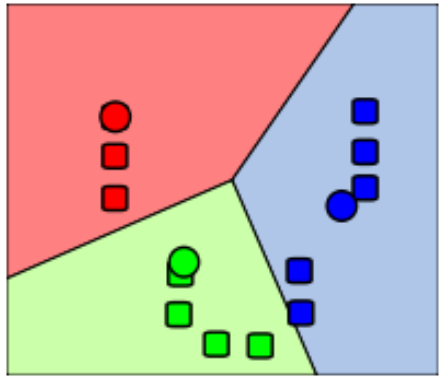
E.g.:

Author	*Shall*	*I*	*compare*	*thee*	...
Shakespeare	1	1	1	1	...
Me	0	9	0	0	...

The Vector Space Model

Definition

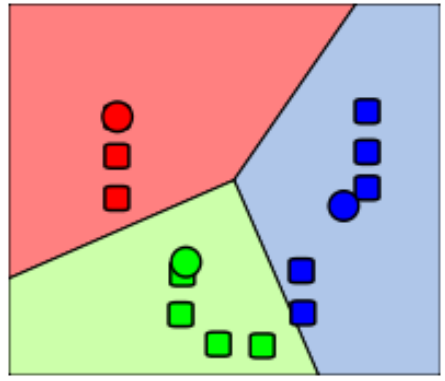
Space: place in which distances are defined



The Vector Space Model

Definition

Space: place in which distances are defined



- texts are more or less distant (dissimilar) in this space
- each vector element is a dimension
- the vector specifies a co-ordinate in the vector space.

Bag-of-Words model

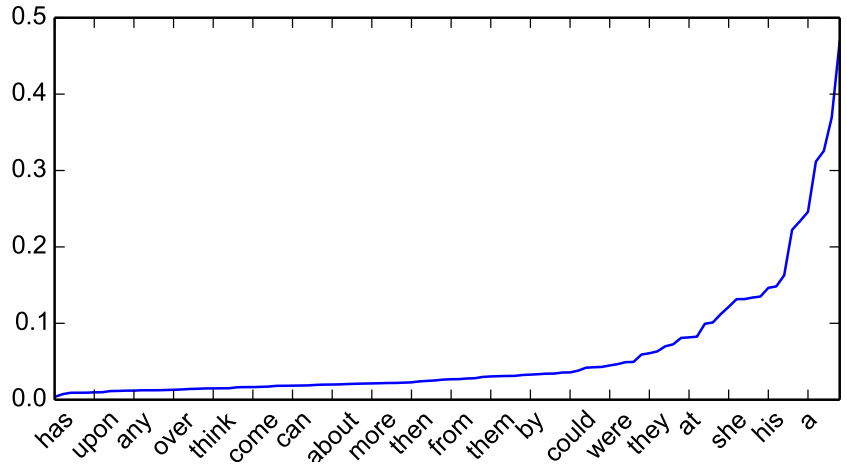
Definition

Bag-of-Words (BOW) model: use word counts as vectors

E.g.:

Author	*Shall*	*I*	*compare*	*thee*	...
Shakespeare	1	1	1	1	...
Me	0	9	0	0	...

Function words vs. Content words: I



Function words:

- Small words, highly frequent
- Unconsciously chosen
- Articles, pronouns, conjunctions
  E.g.: *the, I, and, of, in*

Content words:

- Low- to mid-frequency
- Chosen to match topic
- Nouns, verbs, adjectives
  E.g.: *walk, talk, ship, sun*

Function words vs. Content words: II

For text classification,

Function words:

- Useful for authorship attribution, gender detection
- Small set of words is sufficient
- Pennebaker (2011), The Secret Life of Pronouns

Content words:

- Good at detecting topics, related work
- Large vocabulary required

http://secretlifeofpronouns.com/

Model: making predictions

- Similar texts will have similar word counts
- Simplest model: for a new text,
  find its nearest neighbor
  and use that to make a prediction

Model: making predictions

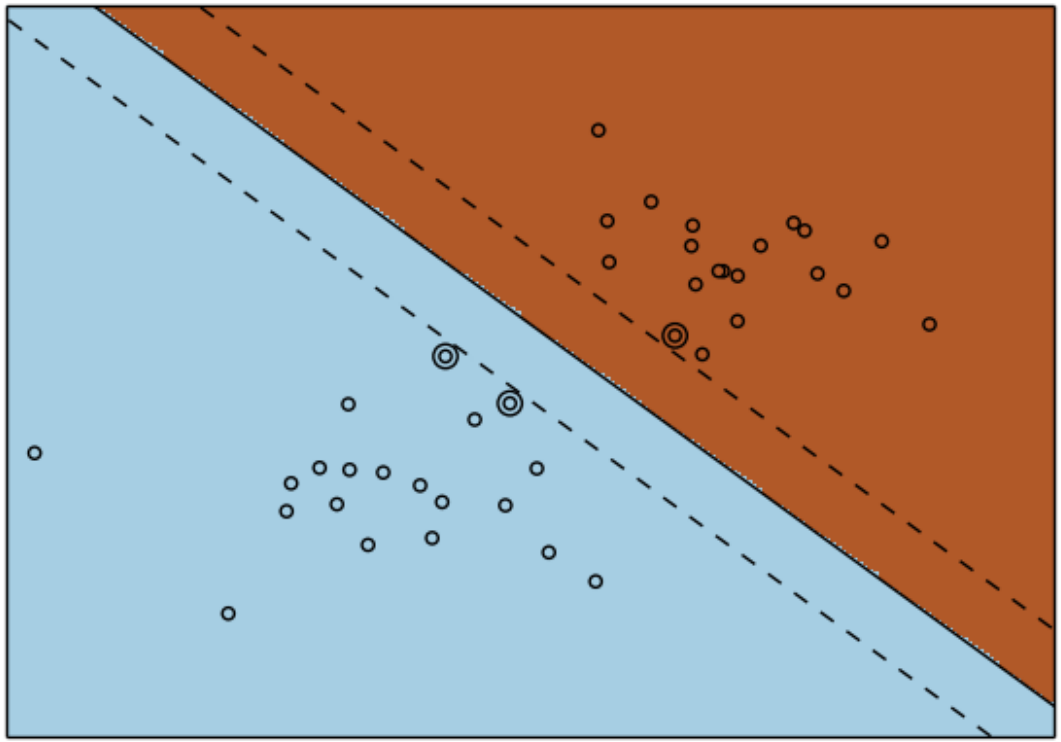
- Similar texts will have similar word counts
- Simplest model: for a new text,
  find its nearest neighbor
  and use that to make a prediction

This works, but ...

- Not all words are equally important
- Not all texts are as representative

Model: Support Vector Machines (SVM)

- Support Vectors are data points that maximally separate the classes to be learned;
- After training, each feature receives a weight that determines how much it will affect predictions
- The support vectors and weights define a line that separates the classes.



Predictions

- Authorship
- Topic
- Readability
- Prose genre (detective, thriller, sci-fi, &c.)
- &c.

Two fundamental problems: I

Problems in Machine Learning:

Definition

The Curse of Dimensionality:
Too many features.
Not enough data to learn interactions of features.

- Limit number of features.
- SVM handles large number of features well.

Two fundamental problems: II

Problems in Machine Learning:

Definition

Overfitting:
The training data has been learned so 'well'
that nothing else can be predicted.
$\Rightarrow$ undergeneralization

- Validate predictions on
  separate data set (train vs. test set)

Dimensionality Reduction

Issues with BOW model:

- Large vocabulary, high number of dimensions
- Would like to merge counts for similar words (e.g., *color/colour, problem/issue*)

Dimensionality Reduction

Issues with BOW model:

- Large vocabulary, high number of dimensions
- Would like to merge counts for similar words (e.g., *color/colour, problem/issue*)

Definition

Latent Semantic Analysis is a form of dimensionality reduction that attempts to summarize word counts as topics/concepts.

Limitations of Bag-of-Words models

Drawbacks:

- Word order information is lost
- Fixed granularity of individual words

Limitations of Bag-of-Words models

Drawbacks:

- Word order information is lost
- Fixed granularity of individual words

Alternatives:

- More complex features; e.g., grammatical.
  But: more complex features ...
  - are more often wrong
  - may have low counts,
    statistics will be less reliable/powerful

Limitations of Bag-of-Words models

Drawbacks:

- Word order information is lost
- Fixed granularity of individual words

Alternatives:

- More complex features; e.g., grammatical.
  But: more complex features ...
  - are more often wrong
  - may have low counts,
    statistics will be less reliable/powerful
- Incremental model; include context
  But: difficult to model influence of preceding text.

Aside: More advanced models

Topic Modeling Identify a number of topics (word distributions)

Deep Learning automatically learn good representations of data (features) using neural networks

Today's menu

1. "The Riddle of Literary Quality" project
2. Machine Learning
3. Your Mission

Today: Prose Genres

- Detective
- Thriller
- ...
- Literary fiction

Who, what defines genres?

- Publishers, critics
- Topics, style of texts

The Data

- 300+ novels from Project Gutenberg;
- Mostly 19th century;
- From following categories ("genres"):
  - Adventure
  - Detective
  - Fiction
  - Sci-Fi
  - Short
  - Historical
  - Poetry

Ashok et al. (EMNLP, 2013). Success with style.

Your Mission

. . . should you choose to accept it:

1. Install Python: `http://continuum.io/downloads`
2. Download corpus & code:
   `http://tinyurl.com/n9aaoht`
   - Unzip, open folder
   - Click on `start-windows.bat` or `start-osx.commmand`
   - A browser opens, open the notebook
     `DH-crash-course-riddle.ipynb`
3. Tweak parameters until score is acceptable
4. Interpret the results

THE END