# Voyages and Travels – an exercise in Naïve Bayes Classification for Travel Literature before the 1920s

Davide Brembilla – Digital Humanities and Digital Knowledge

## Abstract

This paper surveys the possibilities of analysing travel literature with computational methods. It employs a Naïve Bayes trained classifier, applying a Bag of Words model. Computational methods and state-of-the-art methods of analysis are discussed.

## Introduction

The present paper will apply computational methods to travel literature. I will apply distant reading methods to inquire whether data supports the idea of a unified genre of travel literature.

Distant reading can be a useful tool of analysis when inquiring large portions of literature. However, its validity can be questioned. A reason for this has been the limited scope of the studies, often aimed more at the tool rather than to literature itself (Hammond, 2017). In general, distant reading can be useful for explaining trends or claims or to inspect new possible paths to study rather than demonstrate an hypothesis.

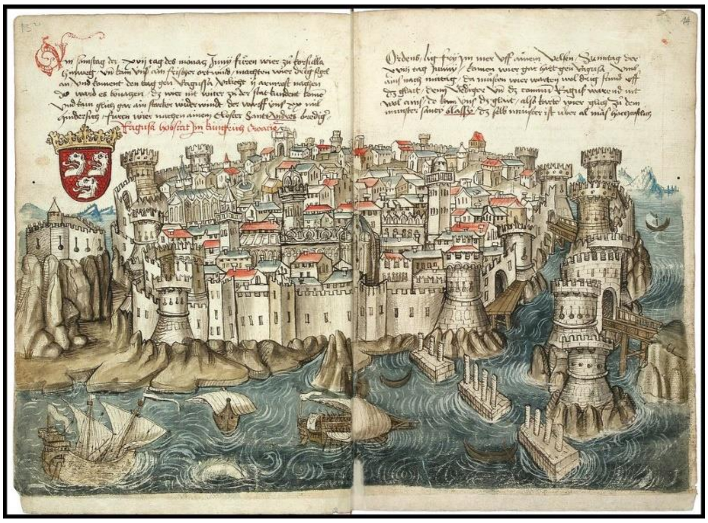

*Figure 1 Illustration of the book by Konrad von Grünenberg: Beschreibung der Reise von Konstanz nach Jerusalem 1487 Source: https://www.flickr.com/photos/bibliodyssey/3992494805*

## Travel literature

This project aims at studying the genre of travel writing. The notion of travel narrative (intended in a generic way as the narration of a travel) has existed for a long time, first as oral tradition, then as a trope of literature. Within the genre itself there is significant variety. We can distinguish between a mythological or documentary function of travel literature, or between the different voices that can be used for the account: one concentrated on the external world and its description, another aimed at the description of the self through the outside world (Das & Young, 2019). In fact, a fundamental element of travel literature is the presence of *alterity*, brought by a movement in space (Thompson, 2011). In summary, there are three objects of travel literature: the World, the Self and the Other. But, returning to the definition, what exactly qualifies as travel literature? Could we consider any book that contains some movement a book of travel literature? Or should we consider books with travel as a main element part of travel literature?

Fussell (1980, p.203) defined the travel book as

> "[…], at its purest, is addressed to those who do not plan to follow the traveller at all, but who require the exotic or comic anomalies, wonders, and scandals of the literary form romance which their own place or time cannot entirely supply".

While catching some important elements of travel writing, Fussell, who calls travel writing a subspecies of the memoir, leaves out a significant portion of travel writing especially common in travel writing prior to the XIX century, when the genre was rather called "voyages and travels" (Thompson, 2011). These writings were often not as much an autobiographical writing, rather they were a set of instructions for travellers who wished to take on a similar journey. In Fussell's eye, only a handful of people in XIX should be considered travellers, thus only a fraction of all writing about travels should be considered travel writing.

Overall, it is very difficult to exactly define what should be considered travel writing; in fact, following Thompson (2011), travel writing is a fuzzy concept with indefinite boundaries, in which a writer can communicate a fictional travel experience; it can span from texts with an emphasised literary vocation, like Calvino's *Collezioni di Sabbia*'s last chapters, to guidebooks, and it is probably better to consider travel literature in a wider scope than excluding a significant portion of travel narratives.

## Aims

The fuzziness of travel literature does not mean that there might not be any inherent textual features that could characterise the genre. An example of this can be the preference of parataxis over hypotaxis, especially for early travel literature, or the use of travel-related lexicon; at the same time, as more people started to travel, a more varied and ornated style became common, focused on the description of the extraordinary (Adams, 1983). In any case, what I will try to study here by using computational methods is whether pure textual evidence supports the idea of a unified genre or travel writing. To do that, I will train a classifier to distinguish whether a book is a of travel literature or not.

## Method

I will employ a Bag of Words (BoW) model and use Naïve Bayes method in a Python script[1]. This code could be reused for future analysis as a command line script.

*Preprocessing* – First, all text will be processed. I decided to try with two kinds of pre-processing, first a simple one present in the gensim library, that will not lemmatise words, and a more complex one that will lemmatise words with a lemmatiser (from the nltk library), converting them to their lemma (dictionary entry). What I expect by doing this is that, since travel books are often memoirs and thus narrated in the past, the simple pre-process will probably result in more precise predictions compared to the one with using just the lemma of words.

*Bag of Words* – The BoW model is a commonly used model for text representation (and in general for object representation), that considers the sub-elements of the text, in this case words, independently from the context in which they are used (Zhang, Jin, & Zhou, 2010). What this means is that I will convert each document in a n dimensional vector, n being the number of words or n-grams present in the text, that will be used by the classifier to predict its nature. In this case, I will also try to modify the scope of analysis including up to n-grams with 3 words, so that this model could detect simple syntactical features.

*Weighting* – For this test, I tried two different strategies. First I decided to weight the documents based on the tf-idf model, then I tried not weighting the terms; it is in fact possible that raw frequencies could also be efficient for this task. I tried in any case to apply the term frequency-inverse document frequency matrix (tf-idf), that weights a term based on how many documents it is present in; a term that is present in all documents will have a significantly lower weight compared to one rarely present.

*Classifier* – A classifier is a supervised machine learning algorithm that will, given a set of cases will learn to recognise in which category a text falls into. In this case, I used a Naïve Bayes Classifier. Naïve Bayes Classifiers are based on Bayes theorem (Zhang & Li, 2007). They study the probability

[1] Here is the link to the code: https://github.com/dbrembilla/travel_writing_model

of a hypothesis given the data available, in our case whether a book is a travel book or not, and returns the most probable hypothesis as the model's guess. This method is common in text classification tasks, as it is efficient in cases with just two categories and longer documents; while other methods can be more efficient, the heavier computational costs do not justify this slightly higher efficiency.

Prior analysis of this kind about this subject are not known to the author.

## Data

The texts used for the analysis were gathered from the Gutenberg project[2], which offers the possibility of search and browsing for subjects; in this way, I choose English texts or books translated to English from the Middle Ages, starting from Polo's *Milione*, until the 1920s. To train for non-travel books, I chose both fiction and non-fiction books that can be found in the most downloaded section of the site and that were not in the travel section. In total 100 texts were used, divided into "Travel" and "NotTravel" categories.

## Test results

These are the test results. They were limited by the machine I am currently using, that limited how deep the analysis could go. I ran 20 tests for each configuration.

| With tf-idf | 1-3 n-gram and full pre-process | 1-3 n-gram and simple pre-process | Unigram and full pre-process | Unigram and simple pre-process |
|---|---|---|---|---|
| Mean | 68% | 71% | 67% | 68% |
| Median | 67% | 72% | 66% | 69% |
| Standard Deviation | 0.07 | 0.13 | 0.09 | 0.17 |

| Without tf-idf | 1-3 n-gram and full pre-process | 1-3 n-gram and simple pre-process | Unigram and full pre-process | Unigram and simple pre-process |
|---|---|---|---|---|
| Mean | 84% | 84% | 75% | 73% |
| Median | 84% | 84% | 73% | 73% |
| Standard Deviation | 0.096 | 0.098 | 0.11 | 0.11 |

## Conclusion

My analyses return an overall efficiency of around 68% when using tf-idf, while without weighting the accuracy gets as high as 83% for the simple pre-process and 75% with the complete pre-process. In the tests, the accuracy does not vary much changing the pre-process method and the n-gram scope of analysis. The confidence of these results is however limited. Further studies would need to implement a more powerful machine and a significantly larger library, which may have influenced the result as well.

Travel literature is clearly a varied field, and the classifier is probably not very efficient because of that. In general, precision is higher than the recall in the classification and the model tends to predict accurately travel literature over negative samples, meaning that there are elements that

[2] https://www.gutenberg.org/

characterise travel literature over non-travel; more thorough studies in the semantics or syntax of travel literature could give interesting results.

## Credits

A significant part of the code was inspired by https://stackabuse.com/text-classification-with-python-and-scikit-learn

The code employs the following Python packages and libraries:

- sklearn
- nltk
- gensim
- re

## Author

Davide Brembilla is a student in the MA Digital Humanities and Digital Knowledge in Bologna after studying Lettere at the Università degli Studi di Bergamo. He is particulary interested in how technology can empower research and teaching in the Humanities field.

# Bibliography

Adams, P. G. (1983). *Travel Literature and the Evolution of the Novel.* Lexington: The University Press of Kentucky.

Das, N., & Young, T. (2019). *The Cambridge History of Travel Writing.* Cambridge: Cambridge University Press.

Fussell, P. (1980). *Abroad; British Literary Travel between the Wars.* New York: Oxford University Press.

Thompson, C. (2011). *Travel Writing.* New York, London: Routledge.

Zhang, H., & Li, D. (2007). Naïve Bayes Text Classifier. *2007 IEEE International Conference on Granular Computing (GRC 2007)* (p. 708). Fremont, CA, USA: IEEE. doi:10.1109/GrC.2007.40

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *Int. J. Mach. Learn. & Cyber*, 43-52. doi:10.1007/s13042-010-0001-0