A few questions concerning the exercises so far

Jan Jelinowski

08/04/2021

Contents

Dear reader,

I would have to share here a few questions I had while preparing for class. These questions span from almost petty nitpicking to questions about how to interpret results

1 stylo()

The work in *stylo()* has been carried out on a corpus of 65 Arabic novels from the *Nahda*.

1.1 MFW

We use the term *Most Frequent WORDS*. Why "words"? I always thought that, unless we lemmatise, what we are actually counting are forms. This is maybe less strong with English, were declination is absent, conjugation is very simple and there is no grammatical masculine and feminine, but still...

1.2 Different results with different settings

When testing *stylo()* I tried to play a bit with the settings, sometimes getting the same results, sometimes a different one. I was trying to understand how to interpret that.

1.2.1 Results for MFW=100 and MFW=1000

Identical results → as far as the algorithm is concerned, for the correct classification of the texts in a Tree, words 101 to 1000 do not modify in any significant manner the information extracted from words 1-100.

1.2.2 Results for MFW=1000 + Culling=50

Results with and without Culling are identical → as far as the algorithm is concerned, for the correct classification of the texts in a Tree, using MFW=1000 and MFW=1000 with a Culling of 50 brings the same results. → the information needed to classify the texts is found if we look at words that are present in at least 50% of the texts.

1.2.3 Results for MFW=1000 + Culling=100

This time the results differ.

Despite de setting, the algorithm works with MFW=182, which means that only 182 words are present in every text? → this seems awfully low. But if I remember well, *stylo()* only looks up a list of 5000 MFW in every text, so it probably ignores a lot of hapax… Still, 182 seems low, I always thought the common used vocabulary to be around 500 (taking without thinking about it the number usually given as "the amount of words everyone uses in daily life").

Proximities between authors are different, and a few texts of Ali Jarim and Jurji Zaydan are misplaced. → working on 182 MFW shared by all 65 novels doe not allow the algorithm to class them correctly. The use of the common vocabulary is only partially what distinguish the authors among themselves, and shows similarities (=close distances) unseen when taking author-specific vocabulary into consideration. This may or may not point to some other questions:

- what uses of said common vocabulary have been selected by *stylo()* to measure the distances?
- to what should those different uses be attributed? Are they marks of a personnal style, or effects of other sociolinguistic structures (gender, age, *regional dialect*)?

→ this in turn seems to suggest an analysis of the same corpus taking into account the parts of the Arabic world the authors are from.

1.2.3.1 Answering the question

I tested stylo() on the Nahda corpus, putting information on the authors' country of origin as the classifying factor. I have obtained two main results:

- MFW 100. Classic Delta does produce results where the geographical factor seems to have a strong influence, neatly organising the corpus (Syria, Lebanon, Egypt).

The only "mistake" in this classification is the literary production of Jurji Zaydan, which seems to be highly specific. → I would suggest the following interpretation of the results: the "personal style" of *Nahda* authors as selected in our corpus is highly influenced by the geographical factor. Local linguistic characteristics are not an unescapable determinism: Jurji Zaydan's style, by its originality, escapes it.

OR → what makes Jurji Zaydan's novels remarkable is another factor than style. We could suppose he writes in a more classical Arabic. Or a more modern one. What we can say for sure is that his language has highly significant measurable characteristics, and that those characteristics are not determined geographicaly.

- MFW182 with Culling = 100. Geography does not seem to be a valide interpretative key. Local linguistic characteristics can not be considered to be the factor at work behind the similar usage of the 182 MFW of the corpus common to all authors and all novels. To reformulate: the distances between the subsets of the corpus, when calculated on the basis of the 182 words common to all subsets of the corpus, does not show correlation with the countries of origin of the authors. But there still is something that allows *stylo()* to identify the distance as short between subsets written by the same author, with 4 exceptions.

  → Local linguistic characteristics are perceived by *stylo()* on the basis of region-specific vocabulary, not region-specific *use* (French *usage*) of the same vocabulary.

Conclusion: by comparing the *100 MFW* and *182 MFW Culling=100* lists of words for each subset, we could identify regional vocabulary. Moreover, words identified this way could be said to be not only localy specific, but carrying a local character strong enough to allow *stylo()*'s algorithm to correctly classify them by region of origin of their authors.

→ how to do that with R?

Of course these results may seem "obvious", not bringing anything into the scholarship, as I suppose a lot of what I wrote is known. Similarily, I suppose anyone with a decent level or Arabic will identify vocabulary that is characteristic for Egypt, Lebanon or Syria.

I would nevertheless stand by my results as:

- they are obtained statisticaly, thus having a different quality than "obvious" humanitic knowledge. Basically, those results are not intuitive and verifiable.
- they are obtained through *distant reading*: I have never read any of the novels, and have no prior knowledge about the authors.

1.2.4 Results for MFW=5000

The classification is not correct, two novels by Ali Jarim are separetad from the others by the novels of Khalil Jibran.

NEED FOR INTERPRETATION

1.2.5 Results for cosine distance with MFW=5000

Classification is inconclusive.

1.2.6 Results for cosine distance with MFW=100

Novels are grouped by authors, but the tree differs. It *does not* group the authors according to geography.

Conclusion: the characteristics of each subset that are prioritised by a Classic Delta correlate with local linguistic characters more so than the characteristics prioritised by a cosine analysis.

⟶ that means that knowing what both priotitise, we could deduce more information about which elements of the MFW lists correlate strongly with regional variation. It also points out to the characteristics prioritised by cosine analysis as author-distinctive characters with a weak relation to geographical determination.

⟶ there are several sets of countable characteristics of the vocabulary of a text that allow a classification of texts matching authorship. And we have not even tried to work with syntax.

1.2.7 Results for MFW100 and Canberra Distance

I am running out of time to comment, but I would like to add a short note: The classification is almost geographic, with Jurji Zaydan being isolated and two of his novels lost in the Egyptian space, between Ali Jarim and Muhammad Husayn Haykal.

Canberra distance is sensible to specific vocabulary and almost manages the geographical classification. Meaning that the earlier intuition about the weight of localy characteristic vocabulary is only partially right? Could it mean two of Jurji Zaydan's novels use specific vocabulary that the Canberra distance has identified as characteristic for the "Egyptian space"?

1.3 Consensus tree

It works. What does it bring?

1.4 How does *stylo()* work?

The tree is build by pairs. So *stylo()* makes lists of MFW (or n-grams) according to the settings for each subset of the corpus. Then, it compares all those lists and pairs them. Then, it compares the pairs and pairs them ?

⟶ that would mean the second level of the tree compares lists of 200 items, where many items appear up to two times? Every third leve would then compare lists of 400 items, where many items appear up to 4 times…

⟶ what does that imply for Culled trees? Thatwhat is compared are always lists of identical items, that are then paired according to either or both the order of those items of their number of occurrences.

Question: how does the algorithm decide about the level at which to link to subsets or sets of subsets. How does the horizontal distance work?

1.5 Multidimensional Scaling and PCA

How to those work?

How to interpret the PCA graphs?

→ they are showing relative distances of the subsets of the corpus, so what is key to the interpretation is to identify what are the axis, what are the oppositions? Will we try to do that in class?