Beyond Verbal Identity
Contextual Features for Intertext Discovery

C. W. Forstall1, L. Galli Milić1, N. Coffee2 and D. Nelis1
1. Université de Genève 2. University at Buffalo, the State University of New York

Motivation

We use the text-reuse detection tool Tesserae to locate potentially interesting allusions in Flavian epic poetry. While thematic resemblance at the scene level is often important to establishing the connection between two passages and thus the significance of an allusion, Tesserae presently focuses on localised re-use of specific phrases and may miss higher-level contextual cues.

We are testing the viability of larger-scale, “thematic” features targeted at the scene or paragraph level. Our goal is to modify the rankings of verbal correspondences identified by Tesserae according to the similarity of the respective contents of the phrases.

For example, the pair of phrases below was ranked 379th of 912 results by Tesserae, but in the context of systematic structural similarity (see right), otherwise lower-ranking text-reuse becomes more interesting.

... et enim dat candida certam
non Helicon.

[Val. Flac. 5.70]
adspiram auras in noctem nec candida circa
luna neagit, splendit tremula sub lumine pontus.

[Verg. Aen. 7.8]

The effects of authorship

The box plot at right shows correlation between k-means clustering and true authorship, over 10 repetitions of the clustering for each treatment: tf-idf scores on the left, and LDA topic scores on the right. We chose k = 11, the number of authors in the corpus.

LDA was effective at reducing the otherwise significant impact of authorship on the classification.

Cluster stability

We varied k, the number of classes, from 2 to 12, and for each value of k we generated 15 clusterings. Adjusted rand indices were calculated for each of 105 possible pairs of clusters for a given value of k.

The distributions of these (right) provide an indication of the stability of such configuration of classes: small numbers of classes are highly stable; among larger values of k, divisions into 6 and 7 classes are most stable.

Sample results

Below we show one example of k-means clustering into 7 classes, taken from the topic-stability experiments described above. Point size shows how often, in 100 different tests, each sample fell into the class shown here.

Below: a close-up showing only Vergil’s Aeneid and Valerius Flaccus’ Argonautica. This is the type of result that we are looking for: samples fall into multiple classes and are not segregated by author.

Methods

Corpus and text preparation

Our corpus was primarily epic, enlarged to include Ovid’s Heroides and Seneca’s Medea, which we felt might show affinities of style and content to our test of interest, Valerius Flaccus’ Argonautica.

Each sample was 10 lines of text. Inflected forms were reduced to lemmata, using methods comparable to those in Tesserae.

All preprocessing and subsequent analysis was done using R, with the help of the cluster, mclust, lm and topmosaic packages.

Unsupervised classification

We used k-means clustering to search for stable clusters of passages that shared similar language across works. Clustering was performed on two different feature sets:

1) TF-IDF weighted scores for all the words in the corpus common to two or more 10-line samples. Each sample was represented by a vector of approximately 8,000 frequencies.

2) A set of 50 topics generated using Latent Derchti allocation (LDA). Each sample was represented by 50 values, representing its scores for each of the topics.

Correlation between clusterings

We tested correlation between the clusters generated by k-means using the adjusted rand index. This gives, for two classifications, a measure of their correlation above what is expected by chance.

The figures above show the author effect graphically: for the TF-IDF features the distinctness of authors such as Ovid, Seneca, Lucan, Silvus Italicus and Corippus from the central cloud is apparent. Under the LDA treatment, only Ovid maintained the same degree of separation.