1. Title of the Project
Using human in the loop to bridge the gap between close and distant reading

2. Coordinators
Inge van de Ven (Culture Studies)
Menno van Zaanen (Communication and Information Sciences)

3. Project Summary
A key strategy for scholars in the humanities is the *close reading* of cultural objects: reading to uncover layers of meaning that lead to deep comprehension. Close reading involves carefully reading and reflecting on a cultural object, in which we pay special attention to things like characterization, the pace of the plot, the symbolism and imagery found throughout a work. It can also be put to use to uncover and engage critically with power relations that are inherent to each cultural object.

The current digitization of old texts and imagery, and the birth of a whole range of new, digitally native genres, however, poses new questions for the practice of close reading. This is exacerbated in the context of big data and superdiversity, in which it seems increasingly problematic to make inductive statements based on the contingent cultural objects that we study. It simply is not possible to read everything. *Distant reading*, in this context, refers to a reading method that relies on computer programs. The strategy represents an attempt at utilizing big data analytics for the purposes of literary scholarship. Since 2000, many have followed Franco Moretti's provocative call for distant reading. Moretti deemed close reading “a theological exercise” and urged humanists to “read less”. Others, like Michael Manderino (2015) and Antoine Compagnon (2014), attempt to rehabilitate close reading, arguing that we need its associated skills and strategies more than ever in our media-saturated age.

The issue is to a considerable extent a crisis of attention. Moretti has noted (2000, p. 57) that close reading necessarily implies a select canon, or a small slice out of the available data. However, acts of selection are losing currency as big data theorists today deem sampling "an artifact of a period of information scarcity, a product of the natural constraints on interacting with information in an analog era" (Mayer-Schönberger and Cukier 2013: 16-7), and as companies like Google strive to collect and organize the world's information (Vaidhyanathan 2011: 2). A close-up perspective pertains to the small; distance allows us to see the bigger picture, and the latter is currently privileged.

Both close and distant reading seem to have their own function. Close reading allows for the identification of minute detail, whereas distant reading shows large patterns. If both approaches have their own benefits, how can we bridge the gap and unite the most valuable properties of both approaches to textuality? This study seeks to reflect on and develop analytical instruments that combine classical-humanist attention to the singular object with methods applicable to variable scales of textuality.

In this project, we will analyze corpora of literary and non-literary texts from the minimalist to the maximalist that solicit readings which zoom in and out between part and whole, micro and macro, surface and depth. On these basis of these, we propose reading strategies that move beyond the dichotomy and allow us to oscillate between the close and the distant, small and large-scale, minimalist and the maximalist, deep and hyper attention. The aim of

the project is ultimately to add grey scales to the originally black/white distinction of close and distant reading. Can we develop tools that allow us to identify patterns to indicate larger trends within or between documents, while at the same time identify outliers or documents that may propose alternative views on the topics? This investigation requires both the development of computational tools (e.g., topic identification, summarization) that can deal with large amounts of documents, but in order to evaluate the computational analysis, in depth, qualitative analyses of the performance of the computational analyses is essential.

Last year, we have investigated the use of topic modeling techniques (distant reading) to automatically cluster texts in ways similar to how people would do this for close reading purposes. It turns out that this approach leads to interesting and useful results, but that the unsupervised computational techniques are not sensitive enough to the human demands for the task. To properly develop methods that go between distant and close reading, human-in-the-loop approaches are required, such as labeled LDA.

4. Project timeline

Tasks

- *Collection of suitable corpora (text collections) on a range of topics and a variety of scale.* Last year, we used a Reddit thread, which has been manually annotated. Another dataset is now available (https://research.googleblog.com/2017/05/coarse-discourse-dataset-for.html), but this still requires downloading the actual textual contents. Additional corpora dealing with topics in which most texts deal with one view and some other texts present a different view on the topic may also be used. These include online forum posts (large scale, short texts) or academic articles on topics that have attracted much discussion (small scale, long texts).
- *Development of distant reading methods in the context of close reading.* Last year, we investigated topic analysis of texts using Latent Dirichlet Allocation (LDA). This provided useful (for close reading purposes) clusters of texts, but also indicated that additional direction of the computational tools is essential. Alternative techniques, such as Labeled LDA might be useful here and will need to be evaluated.
- *Evaluation (through in-depth analysis, close reading) of the identification of the computationally assigned topics.* The evaluation of computational distant reading techniques for close reading purposes shows how reliable distant reading is with respect to several tasks, such as getting the gist of the texts in the corpora, or understanding the subtle discussions and arguments made between the texts on a shared topic.

Milestones

- Collection of (additional) corpora on a range of topics that allows for the evaluation of close reading versus distant reading. These corpora will be identified by the student assistant with a cultural background (making sure a difference in close and distant reading can be found) as well as the student assistant with a computational background to make sure the collection of the texts is practically feasible.
- Annotated version of the corpora where the annotations indicate different viewpoints on a particular issue. This annotation is based on close reading to serve as a gold standard for alternative, computational approaches.

- Application and evaluation of human in the loop methods for the identification of clusters in the corpora will be performed. This includes automatic, quantitative analyses of the resulting clusters, but also manual, qualitative analyses.

Article on the use of human in the loop clustering methods to identify clusters (through computational distant reading) that are useful for close reading.

5. Research Trainee Profile

To perform the research proposed in this project, both quantitative and computational analyses are required. Even though the coordinators are keen on working together on this project, they will need to cross the boundaries of their own expertise area to properly combine the techniques that are required as this project relies on the combination and interaction of computational and qualitative research. The student assistants will help in bridging these boundaries. At the same time, the students get experience in performing research (in their own field), while at the same time recognizing that alternative research methodologies (performed by the other student) exist and lead to useful results.

Because the proposed research deals with both cultural, communication sciences, and computational topics, the positions could be interesting for students from a wide range of backgrounds. In particular, students with a background in the following tracks should consider applying:

- Culture studies: Online culture
- CIW: Cognitive Science and Artificial Intelligence (or Human Aspects of Information Technology)
- CIW: Data Journalism
- CIW: Text and Communication/Communication Design
- CIW: Data Science: Business and Governance

However, students with other backgrounds might be suitable candidates for the positions as long as they have a background and interest in either cultural/qualitative or computational/quantitative research.

The educational level of the student assistants (Ba, Ma, ReMa) is not particularly important. They should, however, be motivated to learn about a range of methodologies, which may include methodologies that are new to them. Since the project relies on the interaction between the in-depth analysis of texts as well as automatic (computational) analysis of texts, our preference goes out to one trainee with a background in culture or text studies and another who is knowledgeable in computational, big data techniques.

How to apply

Send a curriculum vitae as well as a brief motivation letter to
Inge van de Ven (I.G.M.vdVen@uvt.nl)
Menno van Zaanen (mvzaanen@uvt.nl)