## Exercise 4: Thematic Latency and Topic Modeling

---

**Timing.** Students can begin the exercise in class, complete it outside of class, and report their conclusions in a subsequent class.

**Tools required.** MALLET Topic Modeling Tool, Voyant Tools, a digital archive such as Project Gutenberg, and a scholarly database such as JSTOR.

**Purpose.** In the fourth exercise, based on Jockers' discussion of the computational analysis of literary themes, we will use topic modeling, specifically latent Dirichlet allocation (LDA), to mine major canonical novels for new critical insights. How can computers aid literary critics in discovering latent themes in these works that previous scholars have somehow missed? Why can machines sometimes find themes in literary works that scholars cannot? What can topic modeling reveal about critical presuppositions or bias? Students will choose as case study a major literary work or a collection of literary works by a single author and use LDA, in conjunction with standard databases like JSTOR, to write a prospectus for an original critical approach to the work or the author.

**Step 1.** As in the previous exercise on stylometry, students will learn topic modeling first by learning to interpret results, then by learning to produce these results, and finally by using the technique to frame and answer original research questions. Begin by distributing the LDA output for four groups of novels, contained in "Novels LDA output 2" in the module resources, explaining that the results were generated using the MALLET Topic Modeling Tool. You can also generate your own examples. Ask students, individually or in teams, to examine the twenty groups of collocates, or "topics," for each group, asking them to identify and assign a title to the four or five that seem most "coherent," according the Jockers' definitions of coherent and incoherent topics. Using the website, "Basic Strategies for Interpreting Topic Models," discuss with students their reasons for identifying coherent topics and the potential research questions that they might derive from these topics.

**Step 2.** Ask students to generate their own topic models using the Topic Modeling Tool. Although detailed instructions for using the tool appear on the site where you can download it (included in module resources), you will probably need to walk through the process with students before they can obtain and interpret topic models on their own.

As with the literary fingerprinting technique, this technical process is crucial to the completion of the exercise. Create the text files for submission to the Topic Modeling Tool using Project Gutenberg or another literary database, or use the text files containing Alcott's, Barnard's, Cooper's, and Radcliffe's novels included in the module resources. Encourage students to be persistent in working through glitches. As the website explains, the topic modeling tool will generate results in several different formats. For the purposes of this exercise, students will need only the "Topic Index" or "List of Topics" contained in the "all_topics.html" file. Ask them again to identify the most coherent topics from the list that they generate.

**Step 3.** As an extra though not essential step, students can take the list of words in each coherent topic and apply them to the same corpus in Voyant, checking frequency and colocation. Visualizing the results of the Topic Modeling Tool in Voyant helps the students not only in understanding "topics" as a function of colocation but also in gathering, interpreting, and displaying data for their report at the end of the exercise.

**Step 4.** As in the stylometry exercise, help students to use topic modeling to frame an original research question. Ask them to choose a group of works by the same author or a group of works affiliated by some other category and to generate a topic model for this group of texts. Use the most coherent topics as a guide for searching existing scholarship on the selected author or texts in JSTOR or another academic database. When students find a coherent topic that does not seem to be addressed in existing scholarship, ask them to write a prospectus for an article using the topic model as aid to developing a new scholarly approach to their author or texts.

**Conclusion.** This exercise invites the interesting question, implied in Jockers' chapter on topic modeling, of whether computers can recognize themes in texts that human readers cannot, if we define "theme" as a set of conceptually related words that appear in proximity to each other within the text. In more practical terms, the exercise will demonstrate that students can use topic modeling as an aid to their own critical analysis, a way of making an "x-ray" of texts to find significant internal structures invisible on the surface. In my own experience, students formed a series of surprisingly detailed, coherent, and original critical approaches to the novels of Cooper, Alcott, Twain, and several other authors by cross-referencing their topic models of these authors' novels with existing scholarship on JSTOR.