The Topic Modeling Game  
LDA @ the Farmers’ Market

**Materials for the Instructor**

**Background**:

The Topic Modeling Game is designed to help humanities students get a preliminary grasp on how Latent Dirichlet Allocation (LDA) in MALLET works. It’s not designed to be a perfect replication of topic modeling, rather as a means of simplifying it enough to make the logic behind the method clearer and to allow students to read topic model data critically, to ask relevant questions about how the model was created, and to recognize the difference between reliable and unreliable practices. The game is not an example of problem-based learning; however, it is an example of student engagement and pedagogies that lead into successful problem-based learning[[1]](#footnote-1) activities.

**Objectives**: (can vary based on how much detail is put into the game, more could be added)

By the end of the Topic Modeling Game, students should learn:

* The assumptions LDA makes in order to create the constraints that make the algorithm work.
  + There are a “certain” number of topics in a collection of texts.
  + LDA adjusts to however “topics” are entered by the user.
  + All texts include some proportion of all topics, even if it is extremely small
  + Texts are created through the selection of words across a distributed vocabulary.
* The definition of a “topic” as a “distribution over the language” by creating simple topics and keyword distributions themselves.
* That there is a distinction between LDA and its “wrapper” (in this case, Mallet).
* To differentiate between human methods of “reading” and organizing documents and LDA’s use of co-occurrence to make distinctions between words rather than sematic meaning, and how this relates to LDA’s treatment of polysemy (multiple meanings of the same word)
* Through trial and error the basics of iterating LDA, how topic proportions, document distributions, and word distributions are derived (very rudimentary)
* The influence of “chunk size” on topic keyword results.

**Materials needed:**

* Something to topic model. Ideally, plastic fruit with a few colorful blocks (think children’s toys) might work, since sorting them depends less on verbal sense-making. The first time this activity was attempted, we used small pieces of paper with types of produce: (20 slips with a type of red apple, 20 slips with a type of green apple, 10 pear of assorted variety, 10 types of squash of assorted variety, 10 melons of assorted types, 5 beans of assorted types, 5 types of leafy lettuce, 5 broccoli, 5 tomatoes, and 5 miscellaneous other types of produce)
* Something to collect the produce in. If you are playing with sheets of paper with produce word types, they can be taped to paper as if the paper were the “basket/document.” If it’s plastic fruit and/or blocks, you may want to use an actual basket for collecting purposes and then an additional sheet of paper to keep tallies.
* You’ll need a chalk or white board for whomever decides to be the record keeper for the “MALLETers”

**Time:**

It will take a considerable amount of time to work all the way through the topic modeling game, most likely 1 ½ to 2 hours or so if uninterrupted. Consider carefully how much time you have to dedicate to the TM game. For 50-60 minute classes, consider where a natural stopping point might be. Students might continue the activity in groups outside of class and report back on their findings. Another alternative is to record where you are (using photography, collecting and saving worksheets, etc). Between class writing activities such as asking students to propose questions and to answer them among themselves might be useful. Then pick up the activity during the next class after responding to and discussing questions raised asynchronously outside of class.

Math:

The game does not require the same advanced math that actual LDA performs; however, an understanding of some core math concepts will be important for students to review in order to understand how to create predictions about a corpus of texts:

* Proportion
* Average
* Probability
* More may be added…

Reductiveness:

It is important to present the caveat that this activity is a \*simplified\* version of LDA. Once students grasp hold of this understanding, future use of topic modeling would require improving their nuanced understanding of how LDA works. The Topic Modeling Game is a simplified and less accurate demonstration designed for the purpose of introducing students to the core concepts of LDA.

Possible additional activities:

1. Distribute the output of an actual topic model of many texts. Include the topic proportions, keyword distributions, hyperparameters, and document distributions. Then, ask students before they play the game to explain what they think each of the parts of the output means. Collect those predictions (or have them posted on a collective digital space). Next, do the Topic Modeling Game. After finishing the game, ask students to use “track changes” or some other versioning tool to change their original predictions about the output files from topic models. Explain how each of the parts of the actual model relates to the model created in the “game.” This will allow both you and the students to see what is still foggy.
2. As students to give one another a “tour” of the model’s output files once the game is over.
3. For other kinds of topic models that include chronological change, semi-supervision, etc. More could be added here.

**Things to remember:**

Students, particularly graduate students, in the humanities have been trained not to fail and more importantly not to make mistakes in the classroom. Students are also inclined toward not wanting to feel as though they are “playing games” in as much as “play” in this sense is contingent upon a sense of frivolity and a lack of seriousness. The Topic Modeling Game will expose both these resistances. For the game to work, however, students need to make mistakes and they need that opportunity to revise their predictions publicly so that their understanding of how LDA works becomes deeper and more lasting. Therefore, spending time leading up to this in-class activity on creating a supportive environment in which students can make mistakes and revise their predictions publicly will go a long way to improving the outcome of the activity.

**Keep in mind that this will be an uncomfortable process for students. That’s ok. It some encouragement and prompts, such as “You’re on the right track” or “Remember takes time and some frustration. As students become “stuck” it is a good idea to offer that you may have to try this process several times before you can “name” the topics. The point is for students to make mistakes, to revise their predictions.**

1. For more information on problem-based learning, see Samford University Center for Teaching, Learning, and Scholarship: <http://www.samford.edu/ctls/archives.aspx?id=2147484113>. [↑](#footnote-ref-1)