**Notes by Alan for** [**WE1S Meeting (2016-11-11)**](http://4humwhatevery1says.pbworks.com/w/page/112721815/Meeting%20(2016-11-11)) **on Interpreting Topic Models**

(last rev. 11 Nov. 2016)

**In general,**

* the sample corpus we are working with is too small at present (increases arbitrariness in what articles are associated with which topics)
* there is too much noise in the corpus from a variety of sources (article types, scrubbing problems, etc.)
* there is too much friction in using dfr-browser and other tools in conjunction with the plain-text articles to facilitate good understanding.

**Interpretive Method**:

(Using Dfr-browser and reading articles in the CSV files; these notes adapted from Jeremy and Alan's notes)

1. **List view** > sort by rank (greatest to least)
   1. observe the top topics (which can often be overly general topics)
   2. observe the next tier of topics
2. **Scale view** 
   1. See if there are clusters of topics (central clusters and outlier clusters). The thickest rings are the top topics. Mouseover to confirm the topic numbers of interest. Are they near each other? Are they central or outliers? Are there other topic numbers of interest near or far from them?
   2. Also: if available, use dendogram analyses to observe clusters of topics.
3. **Topic view** -- observe individual topics. When looking into a topic:
   1. Make a guess about a label-able "idea" ("theme") of the topic by looking at the top frequency words in it. (Also, if available, use word cloud visualizations of topics)
   2. Then observe the top two or so tiers of articles in the topic. Articles are ranked by the percentage % of tokens in them (loosely: words) that match the set of topic words (the absolute number of such matching tokens is also shown as in the column labeled "tokens"):
      1. Read sample articles of interest (and search for some of the topic words in the articles). Try to understand the relation of the article to your idea of the topic.  
          (Note: currently, we're limited to reading the articles in the CSV metadata files. For example, open the six files for 2007-2009 LA Times and Wash Post "humanities", and search the relevant one for the article you are interested in. You can copy the article into a Word doc for ease of reading and word searching if you wish.)
   3. Iterate the above, seeking to adjust your idea of what a topic is about,
   4. Assess whether there is a valid or approximate match that is possible between a human reading of the articles in a topic and your idea of what the topic is about.
   5. Assess if you think the topic model is at the right level of granularity (number of topics). Would interpretability be improved if themes and articles were divided between a larger number of topics?
4. **Word view** > List of All Words > search on "humanities" (or other words of interest)-
   1. Observe the topics and articles (as per above)
   2. If you find yourself moving associatively between separate words that seem to be related to the set of articles you are interested in, you might make a list of these clustered words.

**Noise Problems**:

* **Short articles** 
  + We should do a study of genres of short articles.
  + Should we filter them out, or classify them and set them aside?
* **Events** 
  + Can we think of a way to identify the signature of event listings automatically and set them to one side?
* **Scrubbing**:
  + We should experiment with more Antconc work to identify high-frequency words and phrases that are a problem. E.g.,
    - days of week
    - "Washington Post"

**Tool Problems:**

**DFR-browser Wish List**:

* Show title of topic model
* Show list of metadata files included in model
* Show total number of words (and tokens) in an article
* Better way to read articles (linked from DFR-browser)
* Better implementation of the "scaled view"
  + Clearer indicators of weights of topics in the corpus
* Filter by size of article (or number of tokens)
* Grab a cluster in the "scaled view" and open just those topics
* Consistent, normal browser behaviors in the various topic and article views:
  + Ability to open links in new tab
  + Ability to copy text of linked article titles
* Start topic number at "0" instead of "1" (to match with the way Mallet numbers topics)

**Other Tool Wish List:**

* Automated method of generating: topic clouds, dendograms (with links to topics in dfr-browser)
* Method to annotate or record observations about topic models
  + E.g., ability to take a snapshot of the three or four views of a topic model (and sample articles) that lead one to frame a hypothesis about what the model is telling us.