Topics for Discussion:

1. Stop words and Consolidation

Should we convert contractions to uncontracted bigrams? Should we convert them to underscored bigrams? Or will most contractions be made up of stop words anyway?

2. Topic Numbers

The more topics you have in your model, the greater the likelihood that individual topics will derive from a single or a small group of articles focused on a specific name, place, or event. If a word like "Wellesley" is prominent in a Women and Gender topic, is this topic really about Women and Gender or about a specific event or issue related to specific entities.

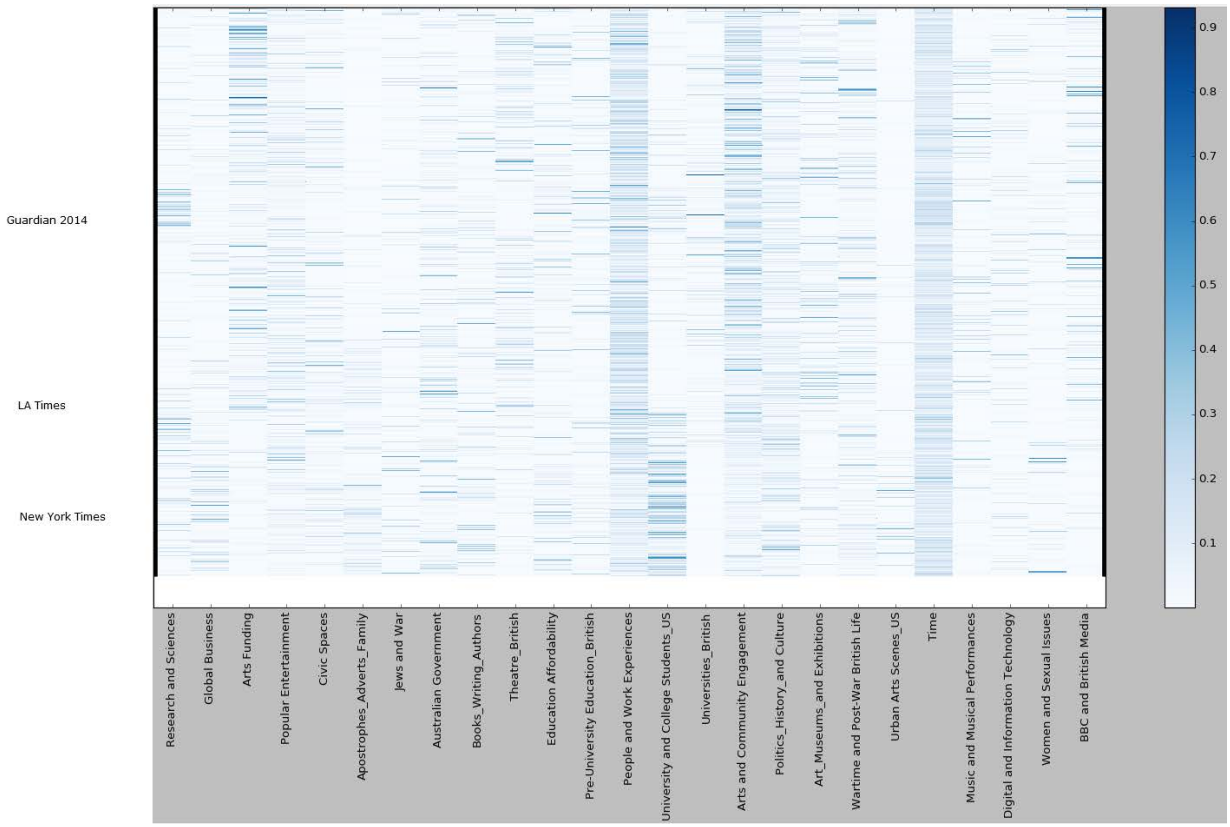
If we measure the similarity between apparently similar topics from different sized models, is this a good way to determine whether we are really looking at the same topic?

3. Effect of Corpus

The query terms "humanities", "arts", and "liberal arts" produced the following numbers of de-duplicated documents:

	2013	**2014**
The Guardian	1326	1257
Los Angeles Times	76	62
New York Times	291	302

This means the most prominent topics in the corpus likely gain their prominence because of one publication (in this case, *The Guardian*) or potentially a given time frame (for a model containing documents from a range of years). The distribution of topic prominence across the corpus looks something like the graph below. Labels along the y-axis overlap so tightly that they are useless, so a few labels are used to mark approximate transitions between years and publications. In many ways, it is the least specific topics that are evenly distributed in an uneven corpus.


Guardian 2014
LA Times
New York Times
Research and Sciences
Global Business
Arts Funding
Popular Entertainment
Civic Spaces
Apostrophes_Adverts_Family
Jews and War
Australian Government
Books_Writing_Authors
Theatre_British
Education Affordability
Pre-University Education_British
People and Work Experiences
University and College Students_US
Universities_British
Arts and Community Engagement
Politics_History_and Culture
Art_Museums_and Exhibitions
Wartime and Post-War British Life
Urban Arts Scenes_US
Time
Music and Musical Performances
Digital and Information Technology
Women and Sexual Issues
BBC and British Media
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1