Visualization of topic models for varying document size

In the following visualizations, LDA models have been trained with the same content, but split in documents of various sizes. In this case, the number of topics is always 80, the implementation use from the gensim module. The documents in the first case have been split according to the paragraphs found in the original source.

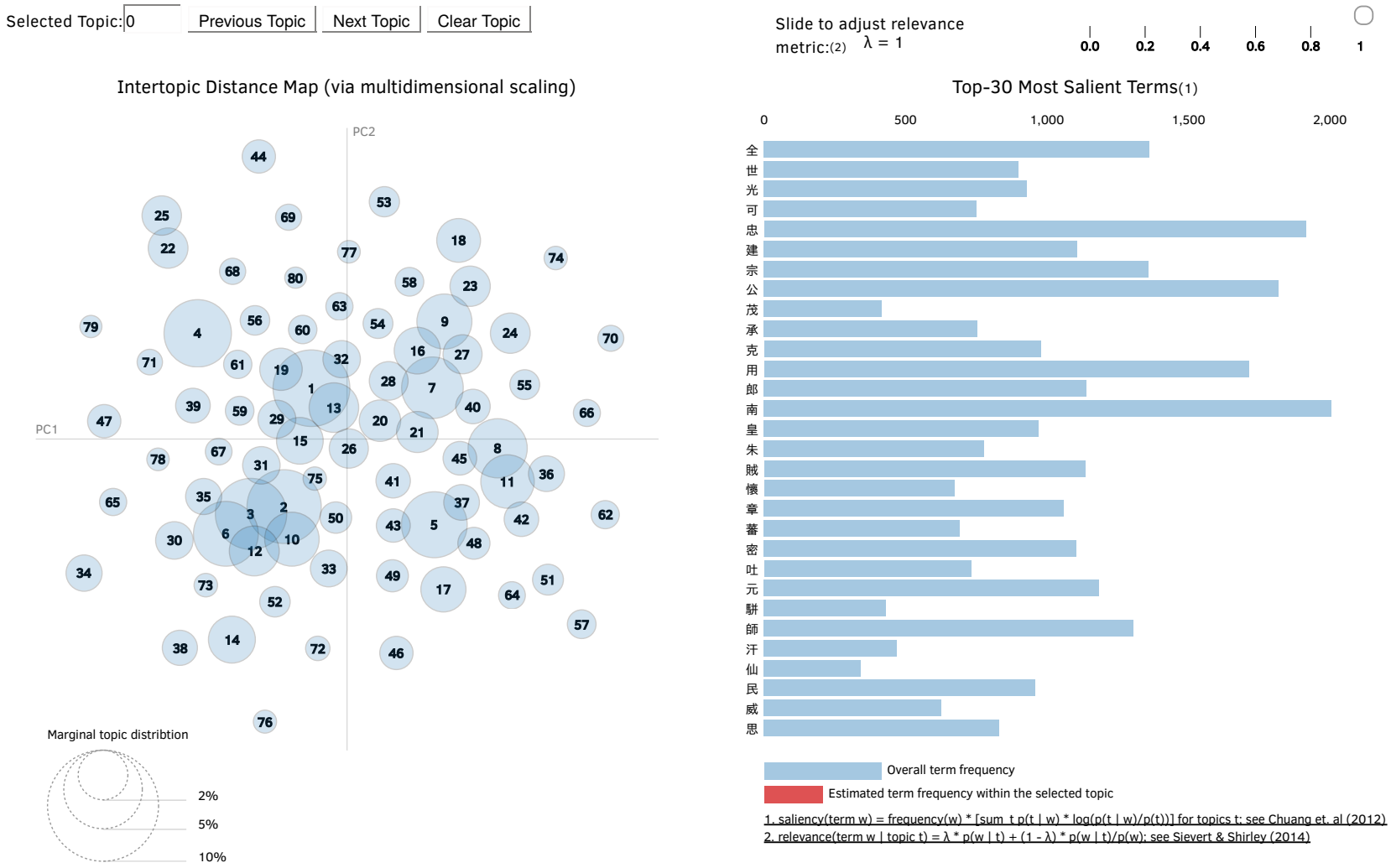
In [1]:
```
import ktm_prepviz, pyLDAvis
#default for mds is mmds, tsne is also possible.  pcoa is not recommended since it gives sometimes errors in the returned JSON
vis = ktm_prepviz.prepviz("doclength", mds="tsne")
```

Distance measures

There are various distance measures, in this case dimension reduction via Jensen-Shannon Divergence & Metric Multidimensional Scaling has been used

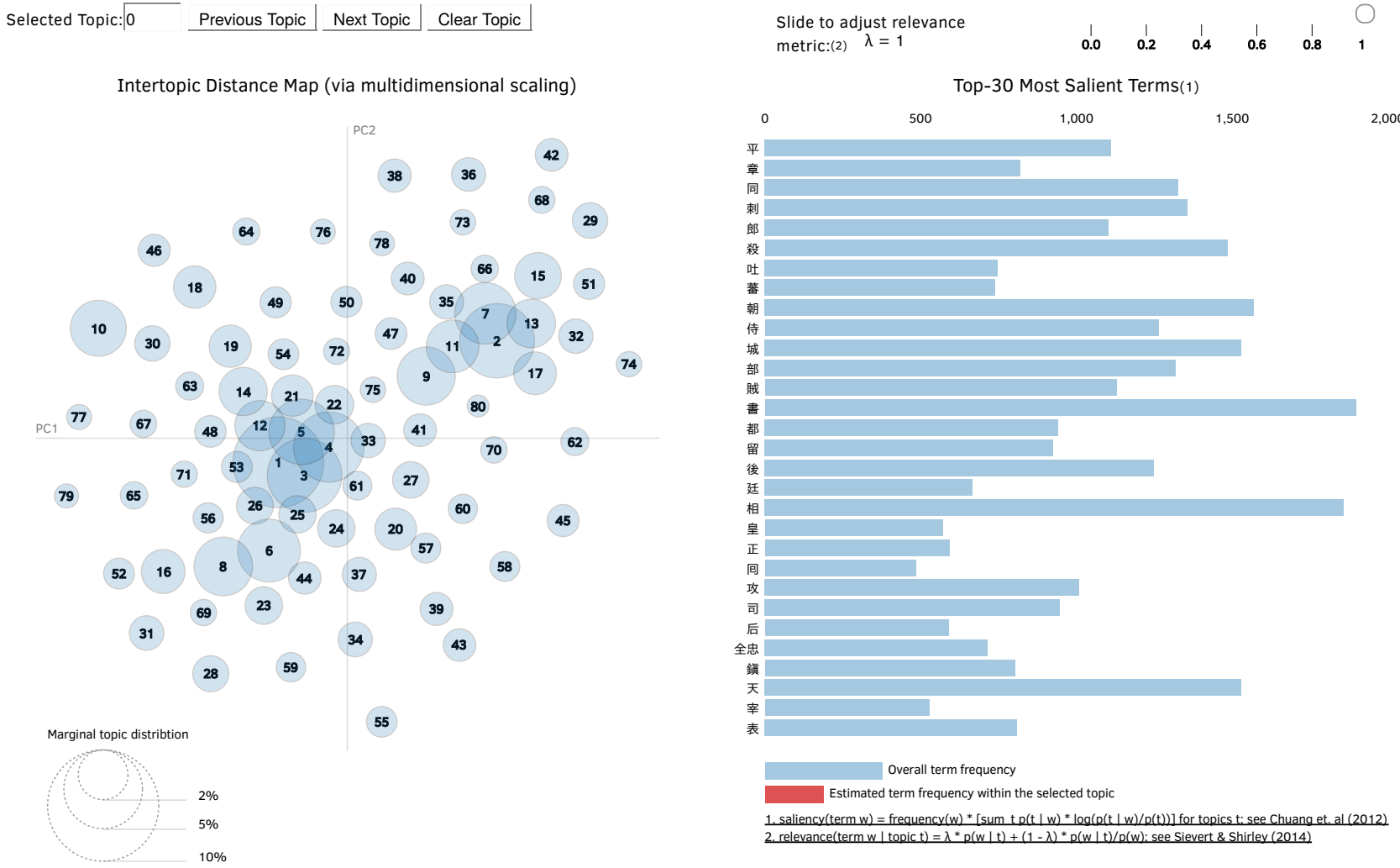
In [3]:
```
pyLDAvis.display(vis[0])
# Variable length of documents
```

Out[3]:



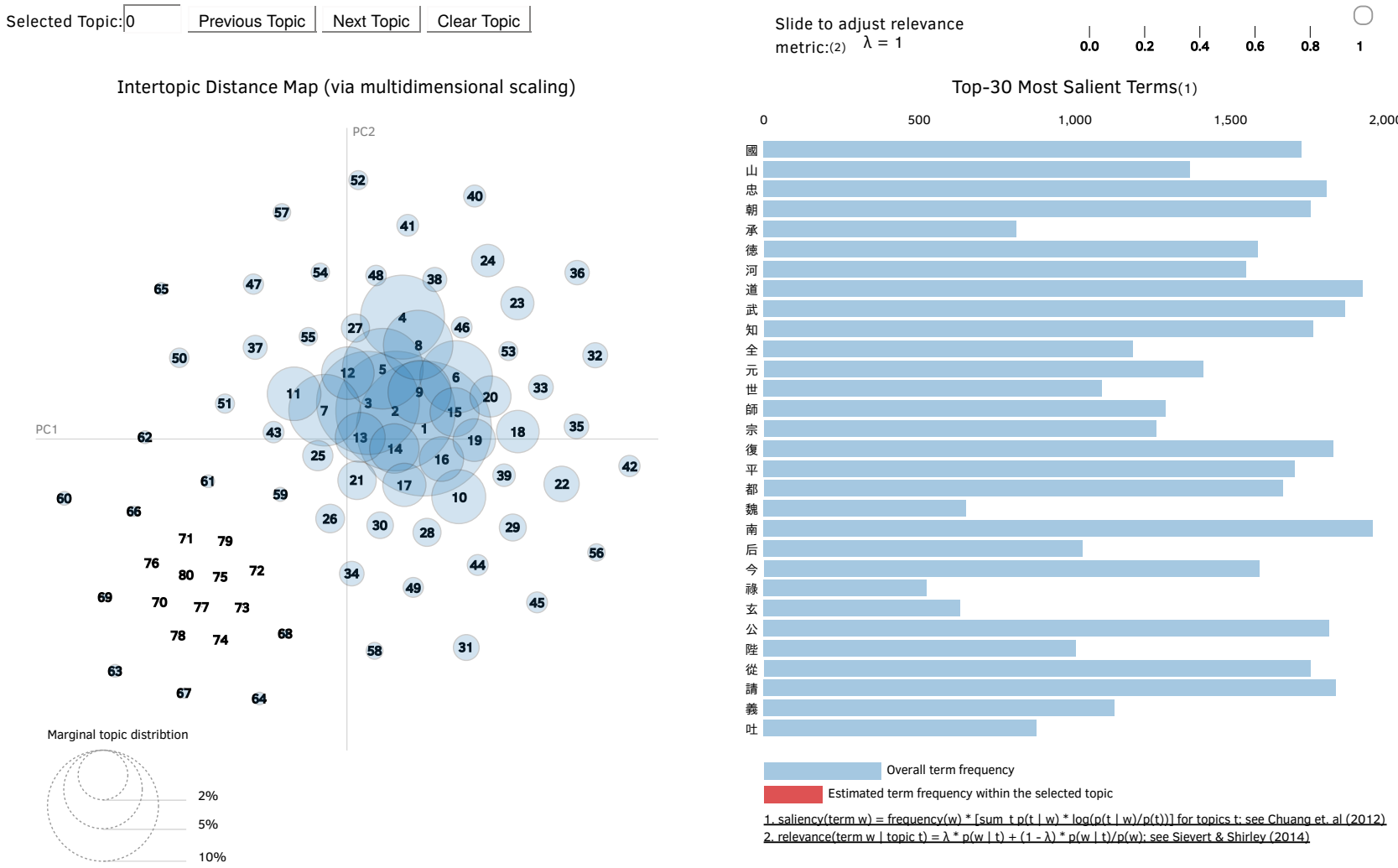
In [4]:
```
pyLDAvis.display(vis[1])
# Documents with average length of around 25 tokens
```

Out[4]:



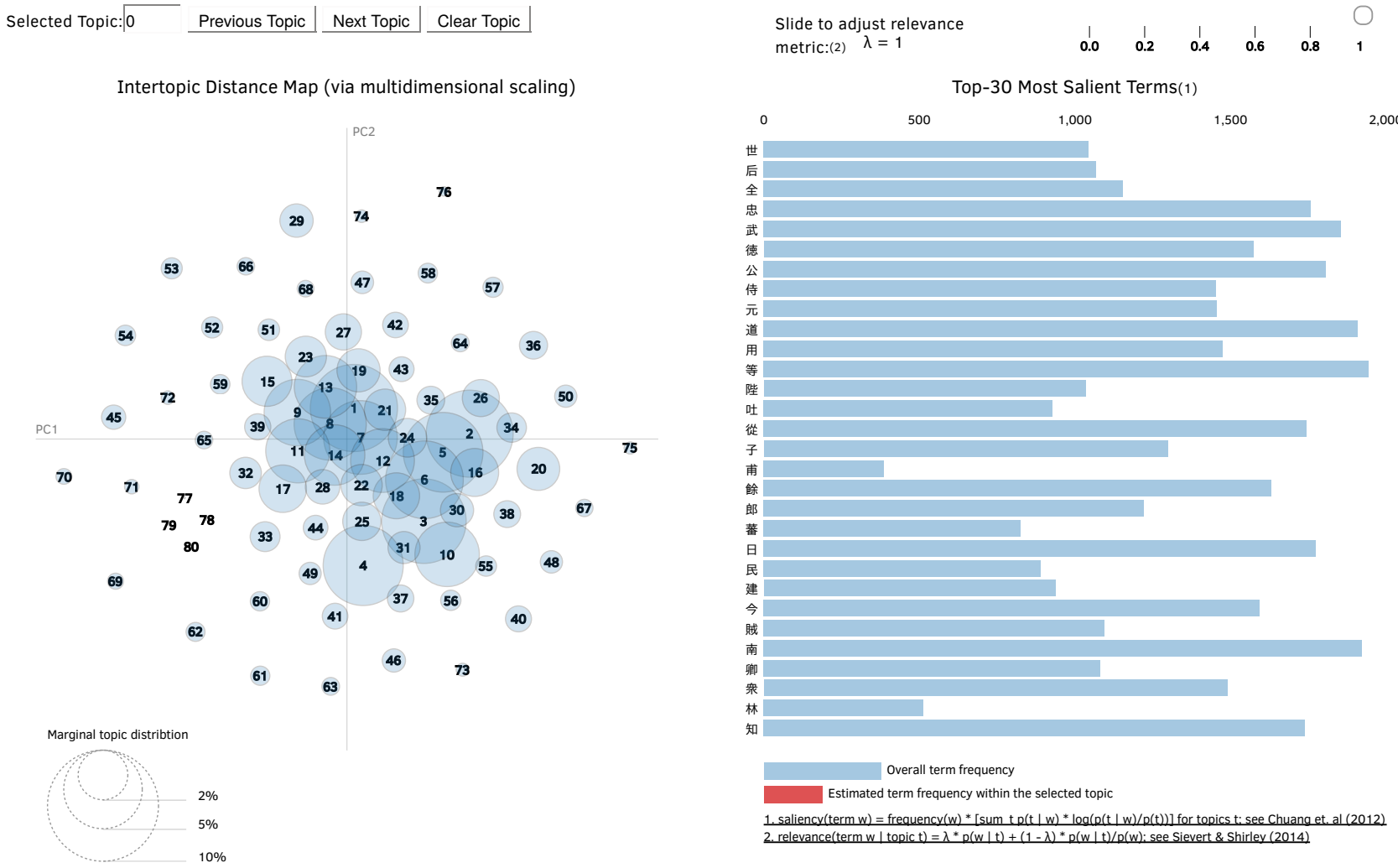
In [5]:
```
pyLDAvis.display(vis[2])
# Documents with average length of around 100 tokens
```

Out[5]:



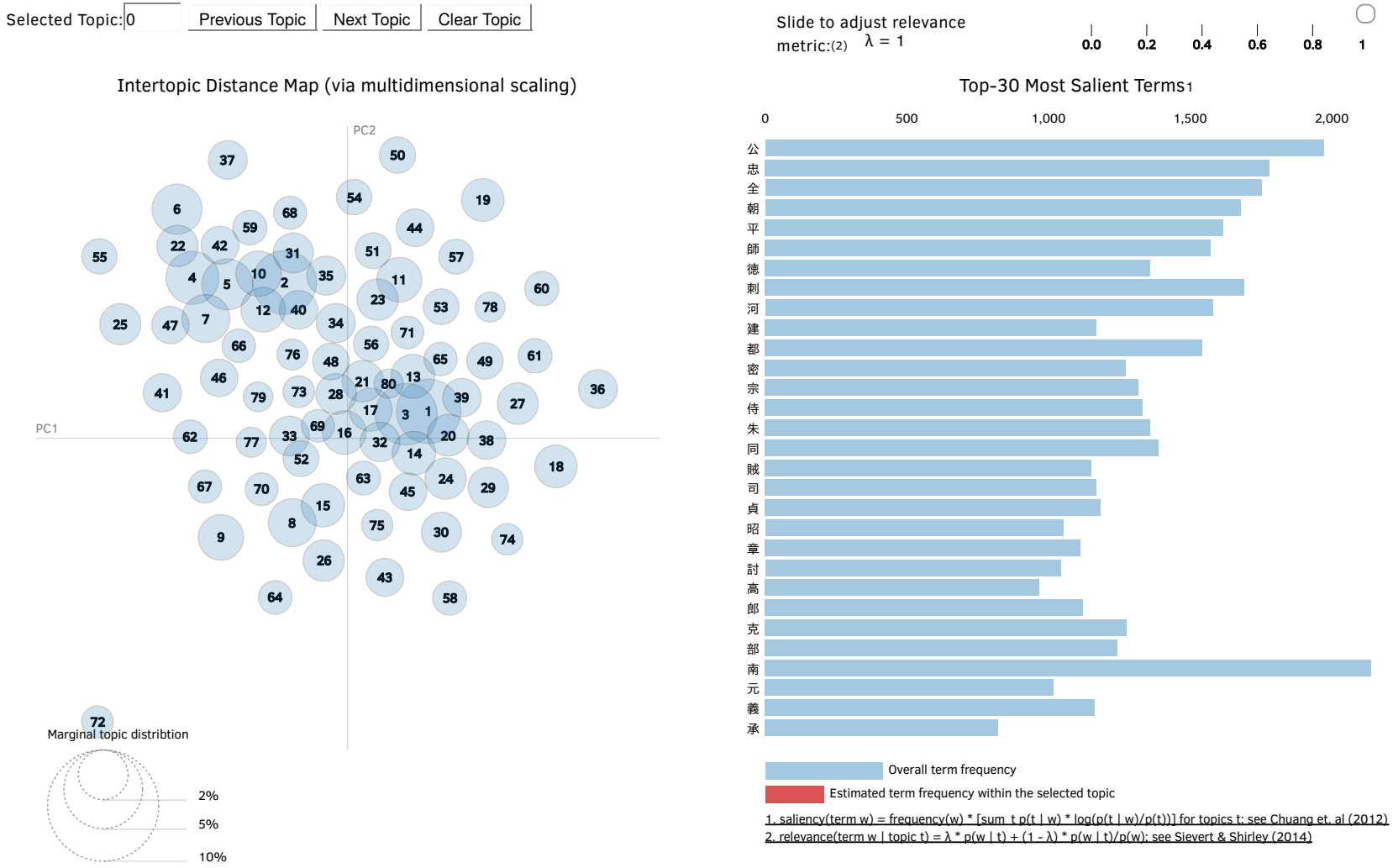
In [6]:
```
pyLDAvis.display(vis[3])
# Documents with average length of around 500 tokens
```

Out[6]:



In [7]:
```
pyLDAvis.display(vis[4])
# Documents with average length of around 1000 tokens
```

Out[7]:



In [ ]: