Тематическая модель новостей6

1. Кластеризация документов – возможно разбить корпус на кластеры. В этом случае документы, которые ближе всего к центроиду кластера, могут иметь наибольшую похожесть между собой.

Также они могут быть наиболее важными («представительными») для данной темы.

1. Попробовать отнести документ к нескольким темам. В случае, если документ относится к большому числу тем, он важен, так как затрагивает сразу несколько областей.
2. Можно отслеживать долю статей по конкретной теме в выборке. Если статей по теме много, то тема является «хайповой». Но в данном случае возникает проблема с обнаружением малозаметных статей – они могут содержать важные темы,