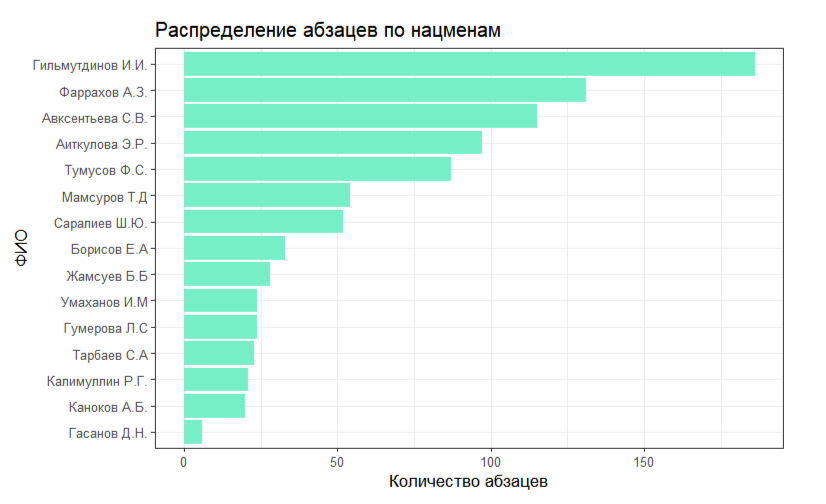
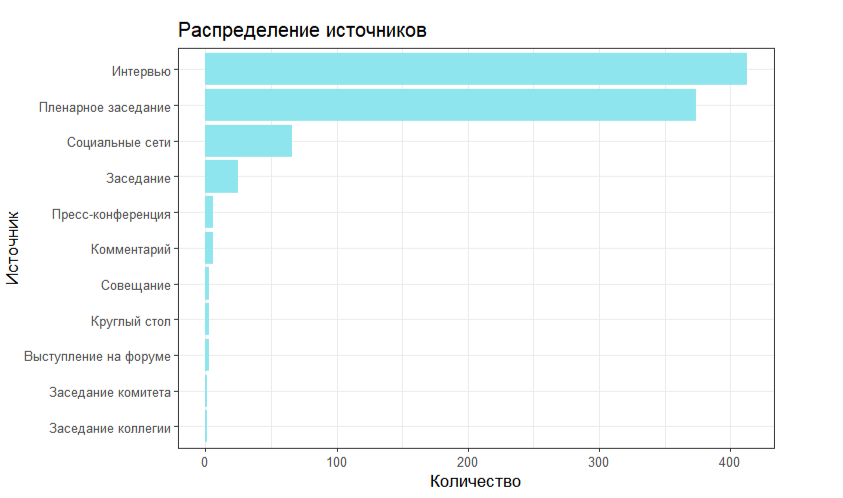
**Тематическое моделирование**

**Исходные данные:**

* 138 текстов (901 абзац) на 15 человек
* цифровой контур: интернет упоминался 6 раз (всего 27 243 различных упоминаний), цифровой упоминается 5 раз, реклама и стриминг не упоминаются
* языковая повестка: есть, язык встречается 119 раз, родной 44 раза, традиционный 18 раз



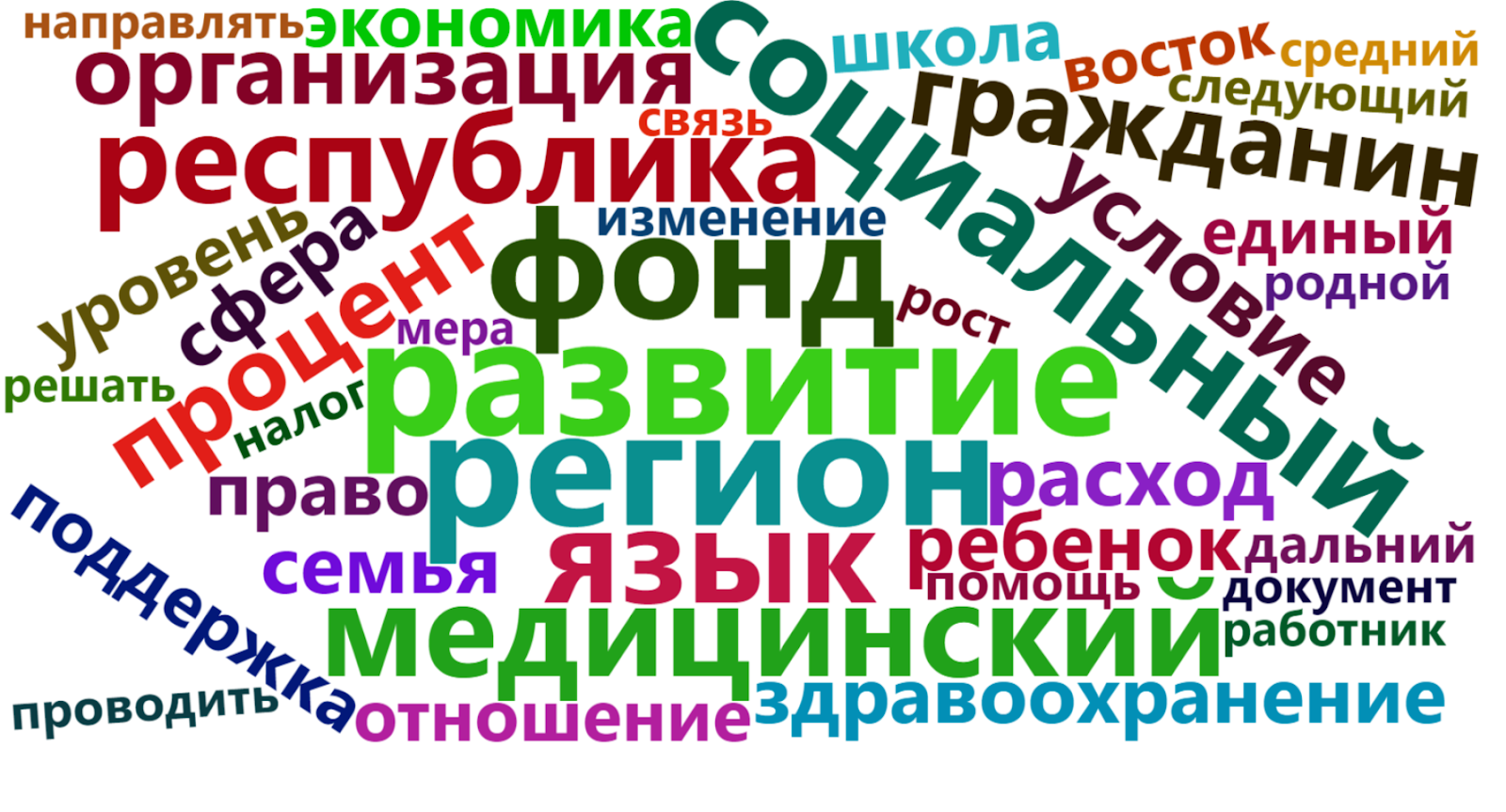


**Предобработка:**

* стопслова: в основном глаголы, местоимения, слова касающиеся внесения законопроектов



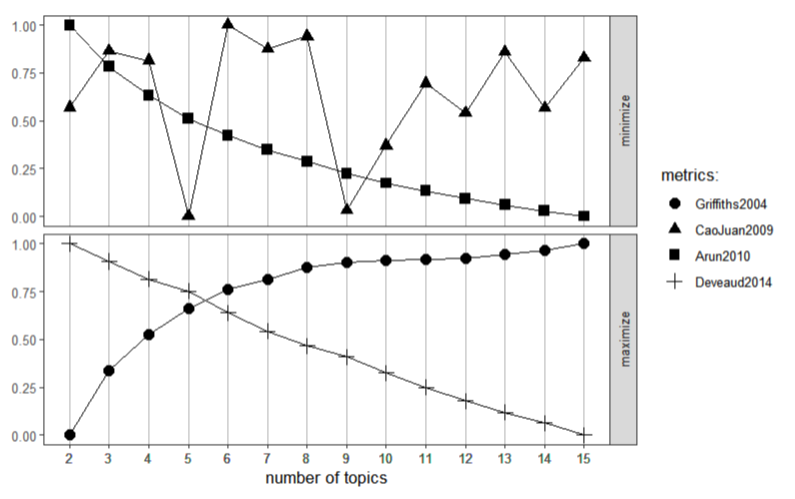
* удаление пунктуации, цифр и лемматизация (mystem) есть в rmd
* наиболее частые слова:



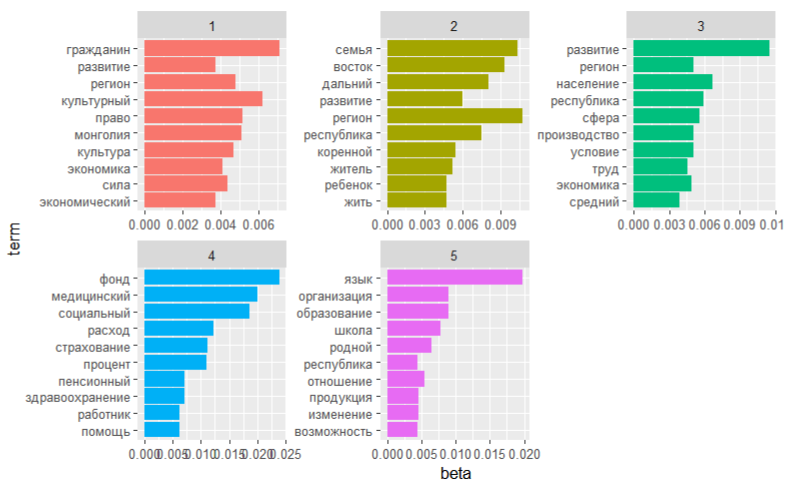
**Построение моделей:**

LDA по абзацам (слова)

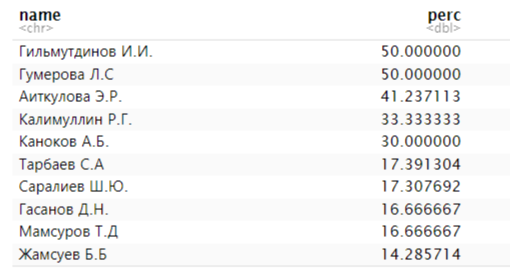
* Оптимальное количество тем: 5



* слова в каждой теме: наиболее подходящая тема под номером 5 (такие слова как родной, язык)



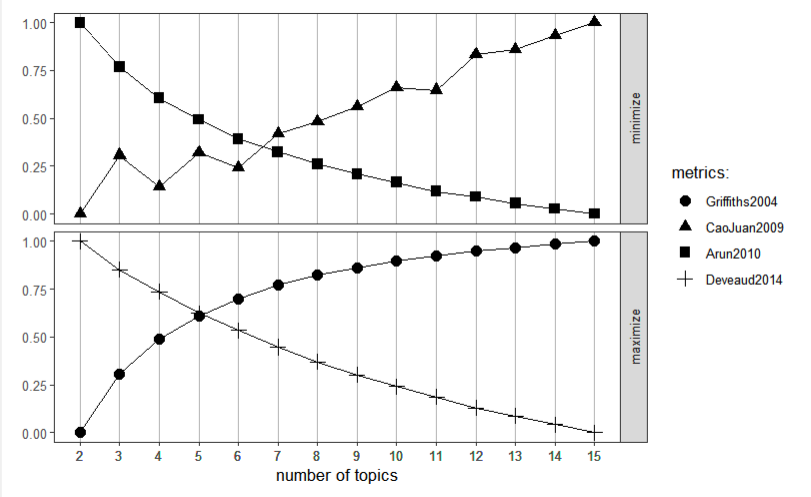
* выбор приоритетной темы (5) и подсчет процента документов, относящихся к этой теме



Гильмутдинов высказывался на тему языка. Гумерова высказывалась на тему языка (но основная повестка - образование - тоже затрагивается в выбранной третьей теме). Аиткулова высказывалась на тему языка. Калимуллин не высказывался на тему языка - при этом высказывался на тему образования - и скорее всего поэтому попал в данную тему.

LDA по абзацам (биграммы)

* Оптимальное количество тем: 5

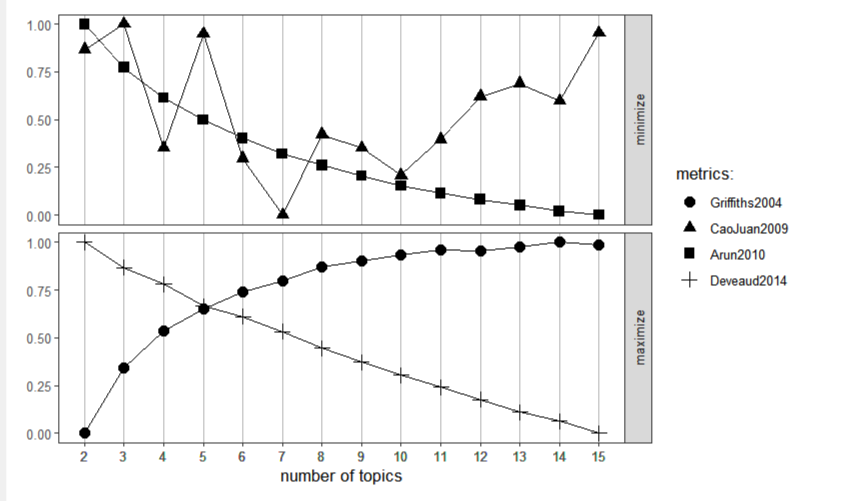


* темы выделяются плохо в данной модели

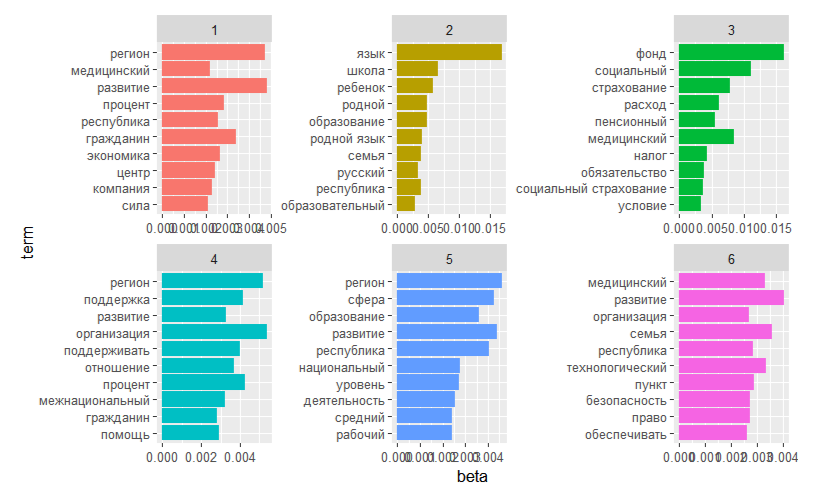


LDA по абзацам (биграммы+слова)

* оптимальное количество тем: 6

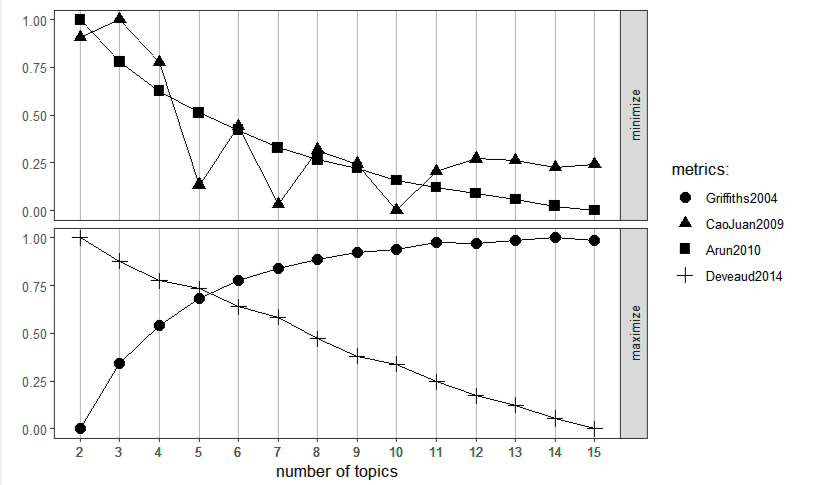


* модель лучше, чем просто биграммы, но примерно такая же, как и модель со словами - тема связанная с языком под номером 2 (язык, родной, родной язык)

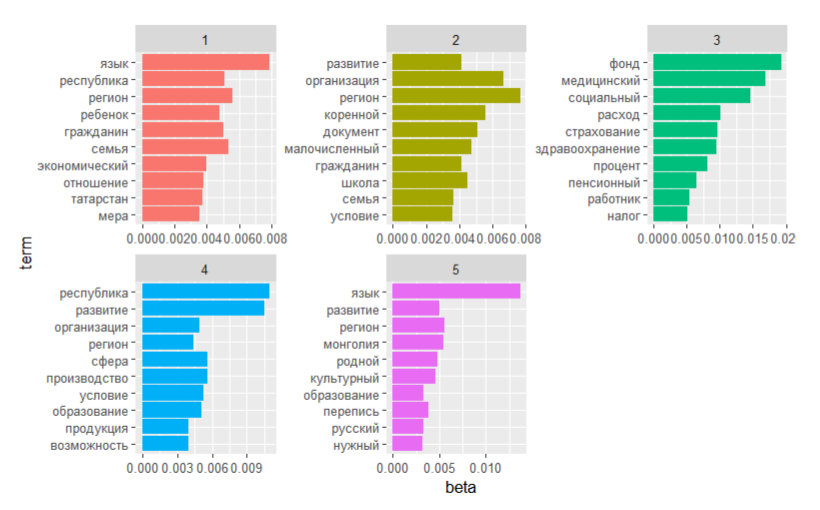


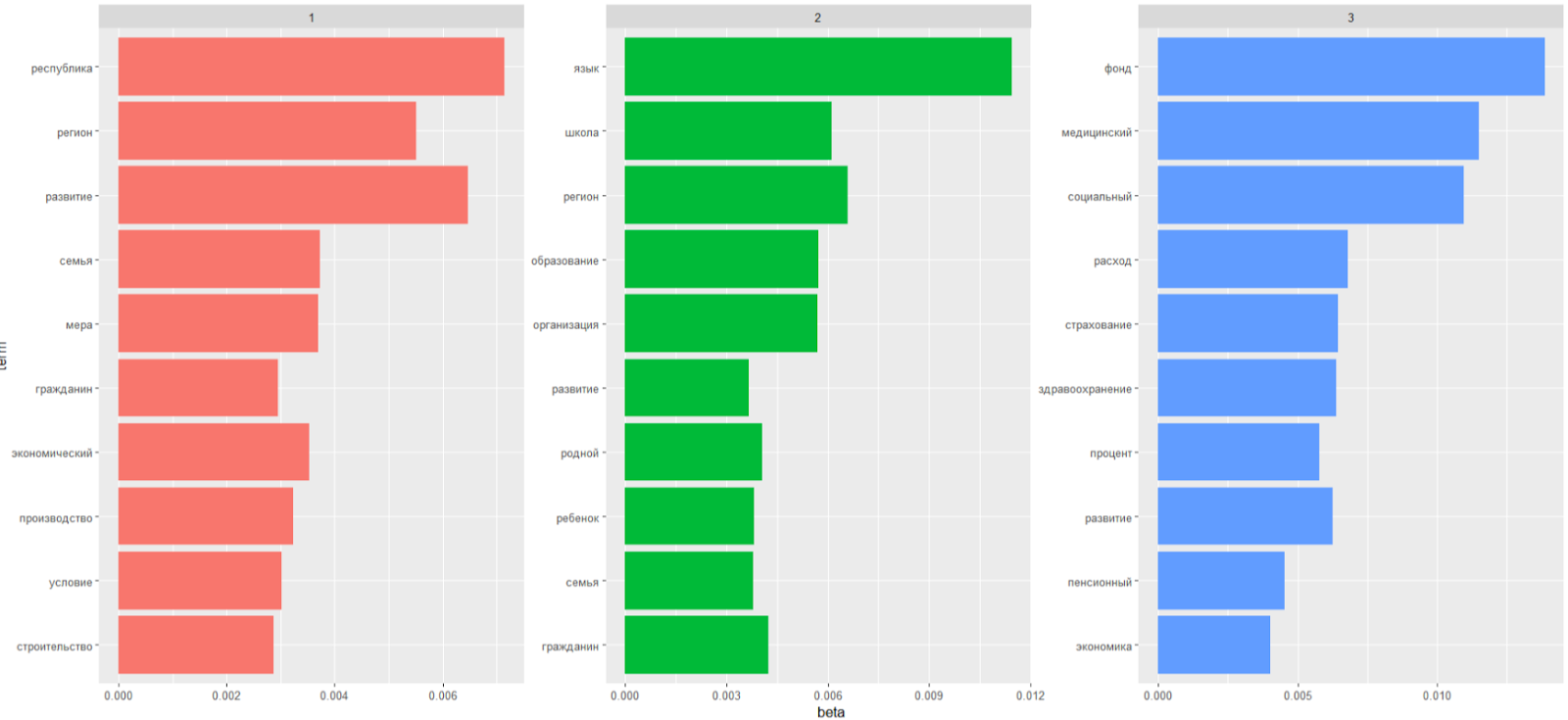
LDA по целым текстам (слова)

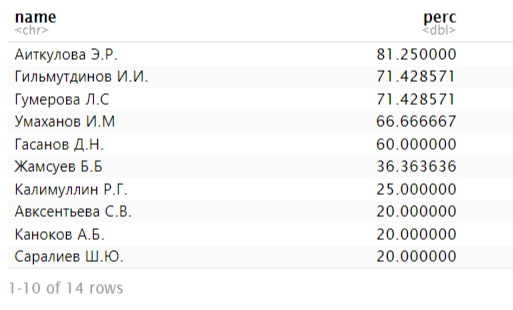
* оптимальное количество тем: 5



* если построить модель по 5 темам, то несколько тем подходят под языковую повестку (1 и 5)



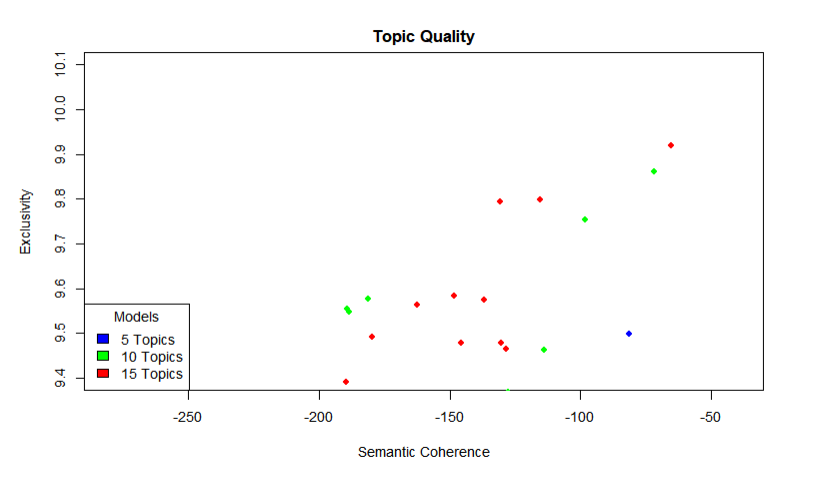
* если построить модель по 3 темам (что не рекомендуется согласно подсчетам оптимального количества тем) - темы выделяются более широко - и вторая тема нам подходит
* 
* так как это тексты, а не абзацы, выберем документы, где покрытие темы не менее 60%



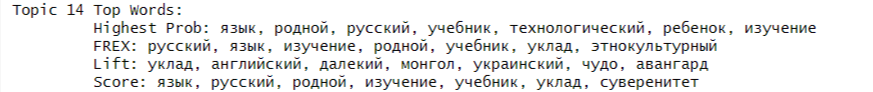
Гильмутдинов высказывался на тему языка. Умаханов также высказывался на тему языка (в основном о мультикультурализме). Аиткулова высказывалась на тему языка. Гумерова высказывалась на тему языка. У Гасанова всего 5 текстов. Таким образом, можно сказать, что наиболее точные результаты по Гильмутдинову и Аиткуловой - в абсолютных значениях и в процентном соотношении. И он, и она высказывались на тему языка.

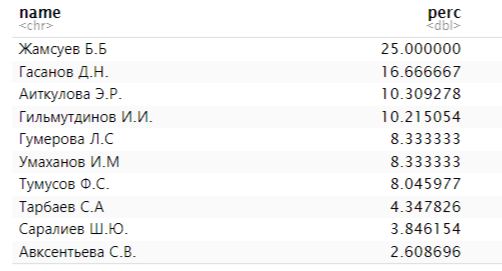
STM по абзацам (слова)

* построено три модели: 5, 10 и 15 тем
* оптимальное количество тем: 15



показатель dispersion самый низкий у модели с 15 темами. Подходит тема под номером 14.





У Гильмутдинова и Аиткуловой была языковая повестка. Жамсуев высказывался про язык (но про английский и русский). У Гасанова слишком мало текстов

Summary:

* Деление на абзацы или на тексты не слишком принципиально
* Отдельные слова лучше чем биграммы
* Если берем абзацы, то выбираем только доминирующую тему и смотрим у спикера процент текстов с этой темой
* Если берем целые тексты выбираем те, где тема покрывает 60% документа, и смотрим процент такой темы у спикера
* STM лучше чем LDA по делению тем, но по спикерам LDA выделяет лучше
* В выборке не хватает текстов про цифровизацию и мало текстов про язык в отрыве от образования
* Список стоп-слов можно увеличить ещё
* Кого выбирать в качестве спикера по повестке можно посмотреть по пересечению моделей: здесь это Гильмутдинов и Аиткулова