Моделирование речевого поведения носителя миноритарного языка РФ в социальной сети

Людмила Зайдельман

научный руководитель: к.ф.н. Б. В. Орехов

НИУ ВШЭ

10 июня 2016

Постановка задачи

- Описать поведение отдельно взятых носителей миноритарных языков в социальной сети Вконтакте;
- Описать поведение групп носителей – сообществ;
- Проследить связи и зависимости между разными признаками, характеризующими поведение носителей;
- Разделить языки на группы по их представленности в сети Вконтакте;
- Построить портрет (один или несколько) типичного носителя миноритарного языка РФ.

Данные

- 43 языка России, для 31 языка нашлось по крайней мере одно сообщество Вконтакте;
- 1595 сообществ, в которых есть тексты на миноритарных языках.

О каждом сообществе:

- название;
- число участников;
- список всех постов и комментариев.

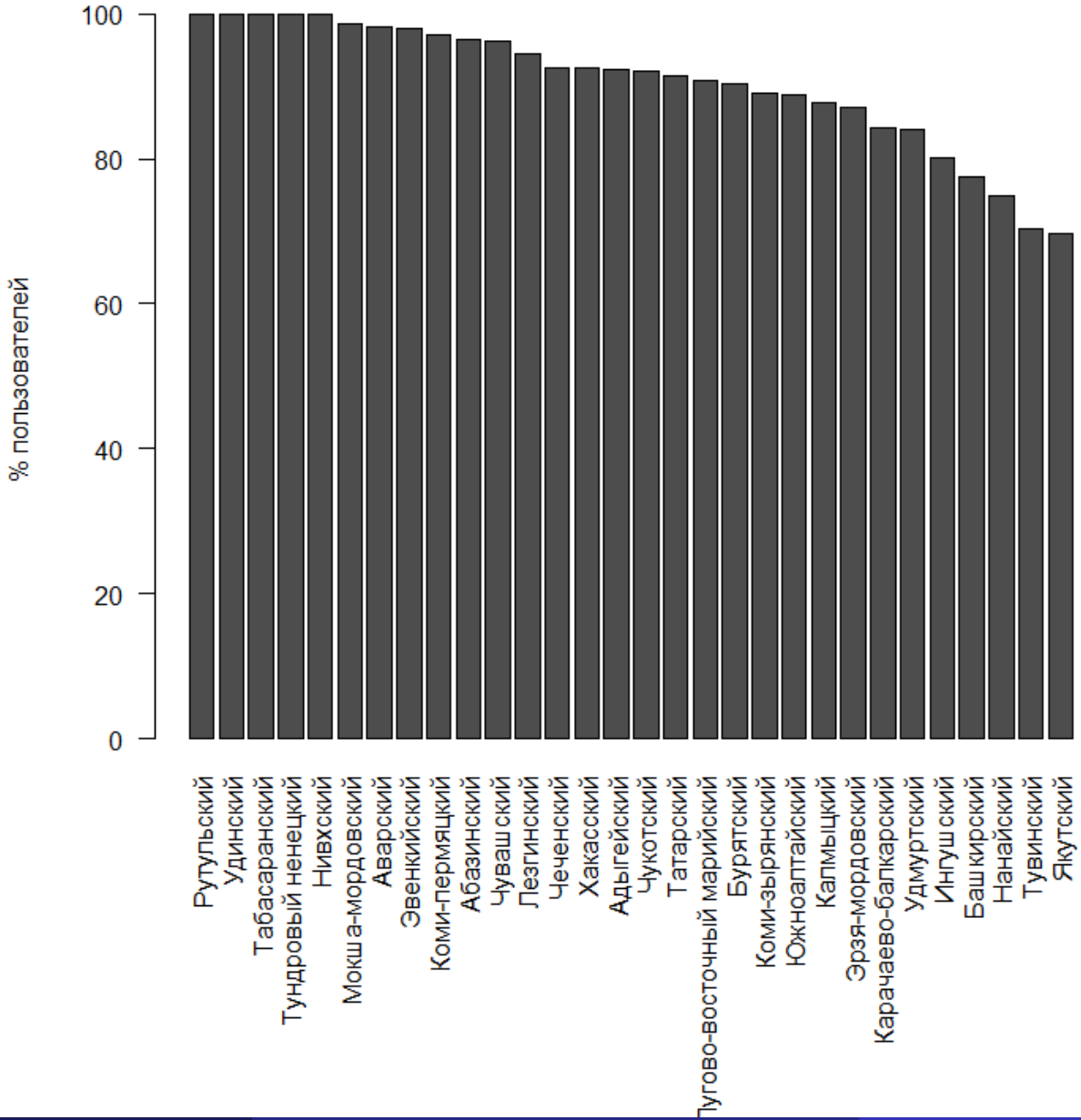
О каждом посте/комментарии:

- текст;
- дата создания;
- число комментариев (для постов);
- id поста-родителя (для комментариев);
- информация об авторе: id, дата рождения; город проживания, пол.

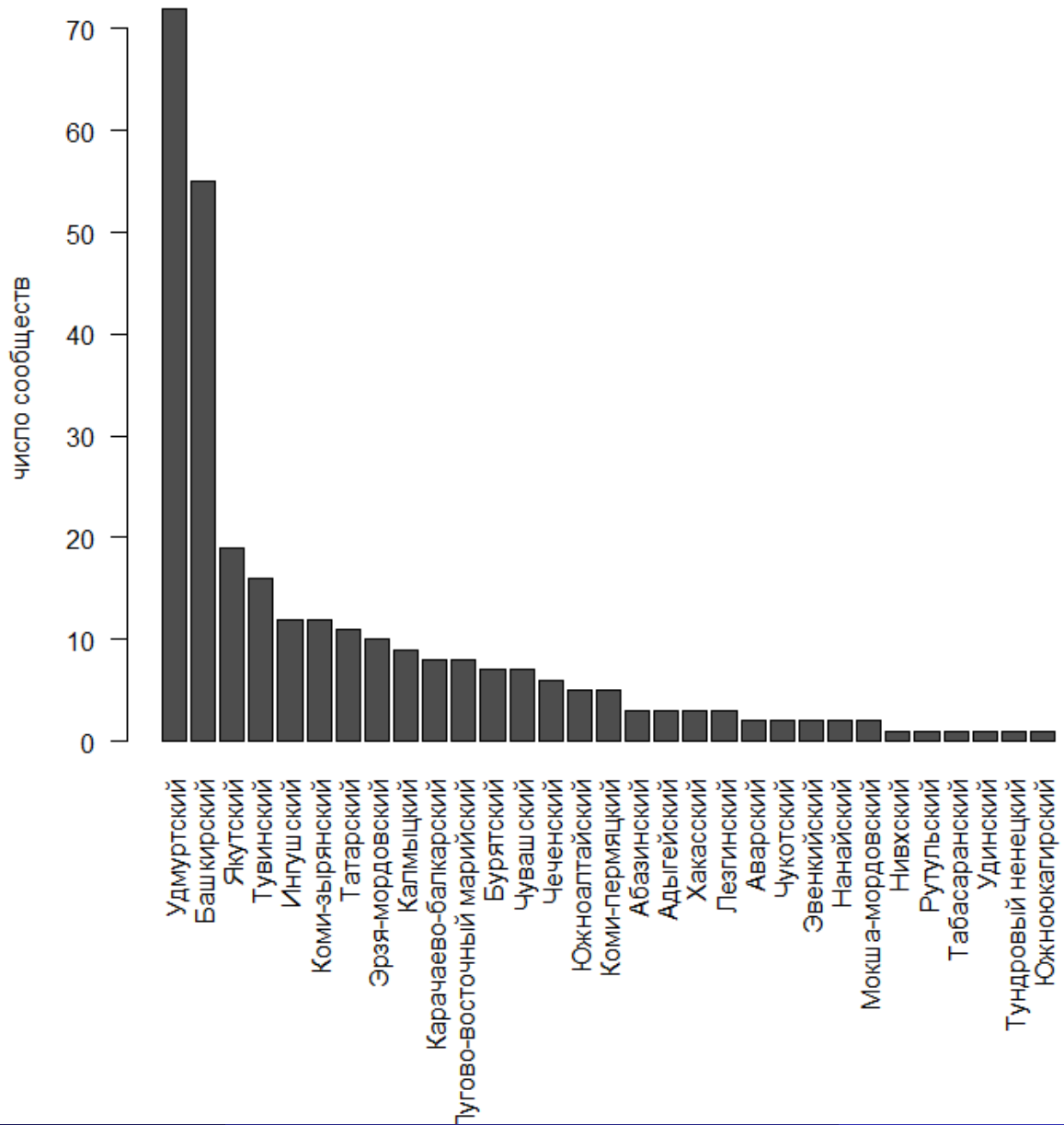
Пользователи

- Пользователь является носителем миноритарного языка, если он написал по крайней мере один пост на этом языке;
- Количество пользователей-носителей не зависит линейно от количества носителей, известного из переписи населения;
- Среди носителей 23 языков регионом-лидером является титульный регион;
- Средний возраст носителей всех языков находится в промежутке 19,6-31,3 лет; самым «юным» языком является чукотский, а самым «взрослым» – коми-зырянский;
- Абсолютное большинство всех пользователей участвует в жизни только одного сообщества, «говорящего» на миноритарном языке: наименьший показатель у пользователей якутского языка – 69,6%;
- 57,4% пользователей-носителей написали только один пост на миноритарном языке; средний показатель равен четырем.

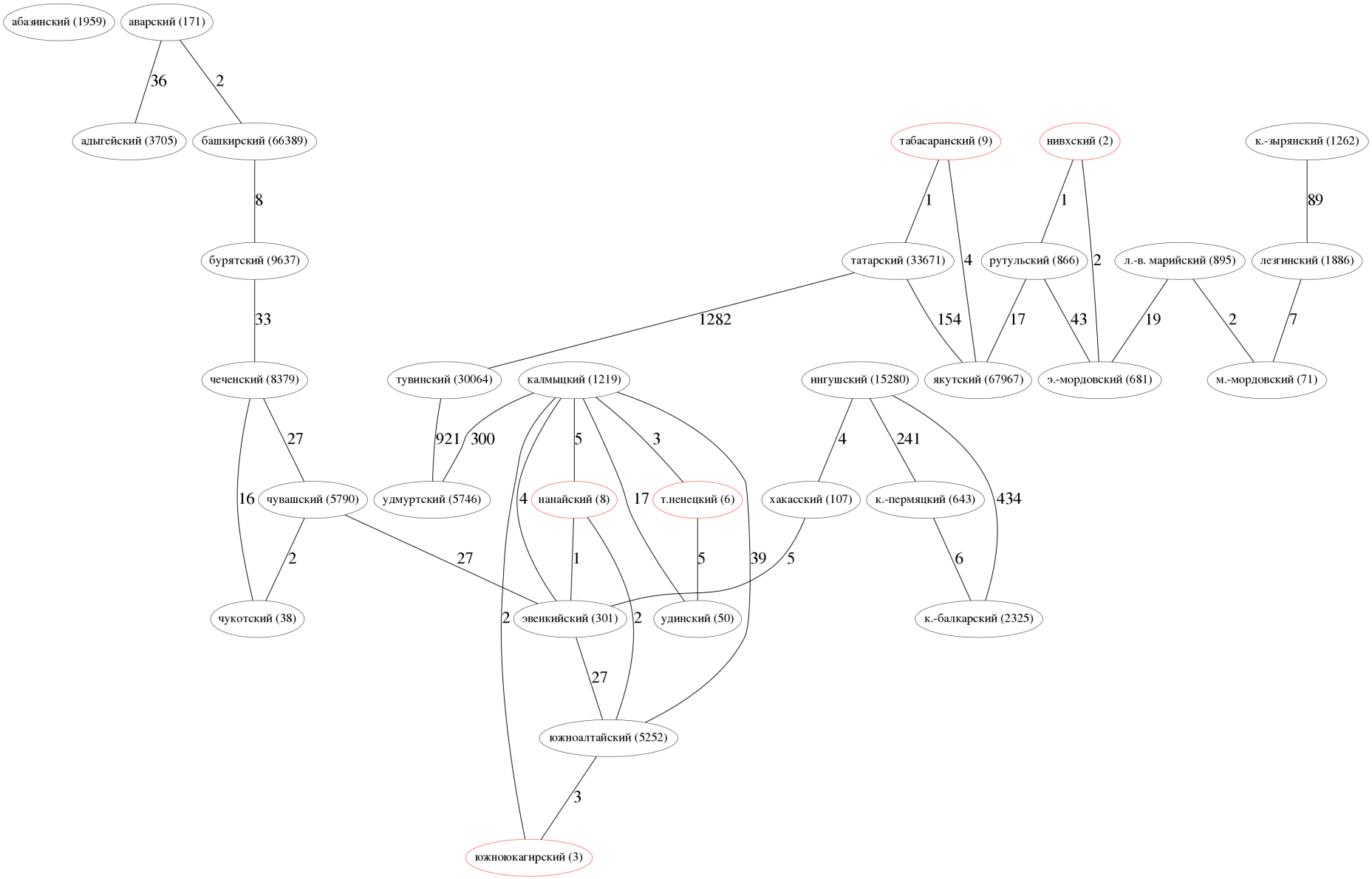
Доля написавших только в одно сообщество



Максимальное число сообществ у одного пользователя



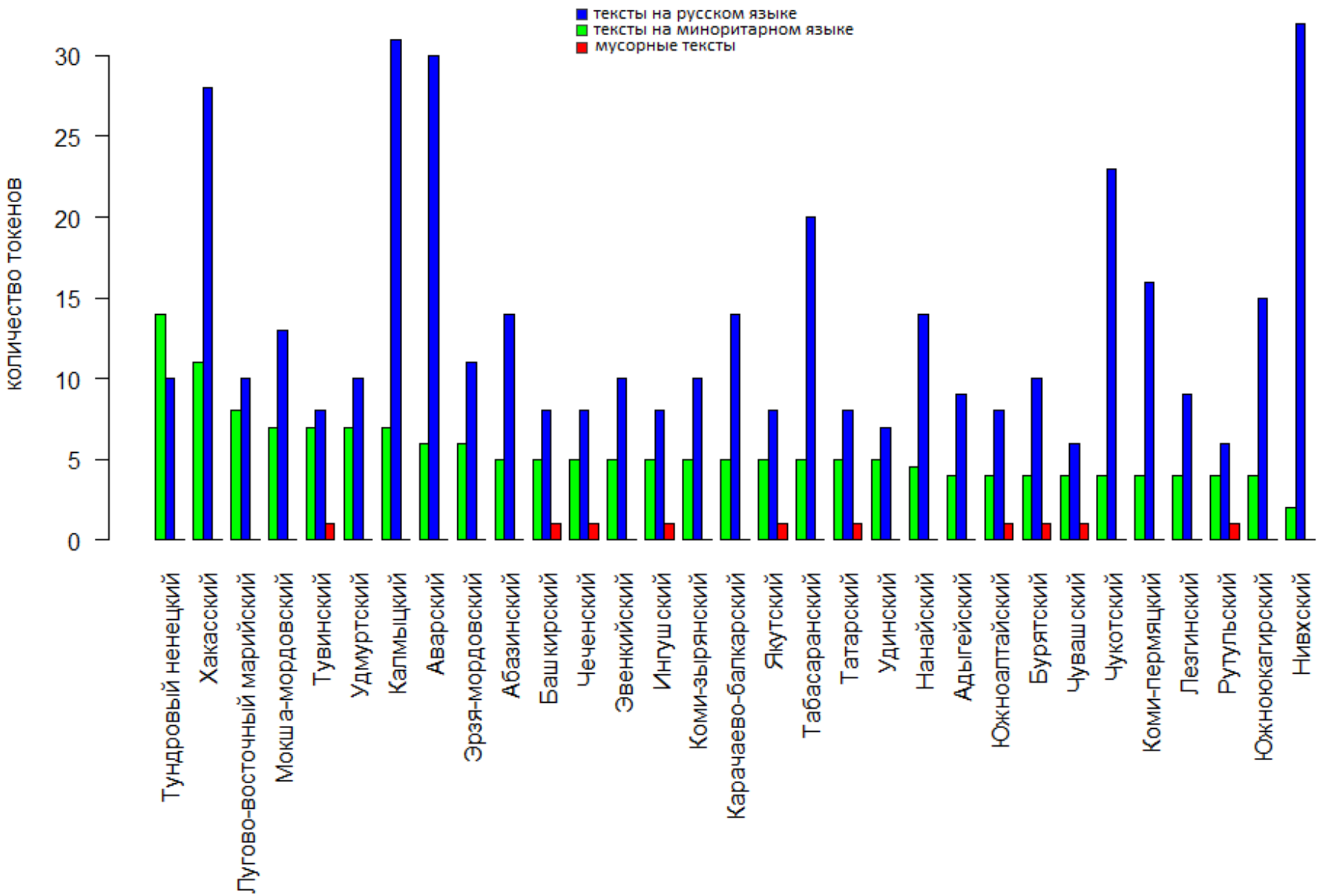
Языковые способности пользователей



Сообщества

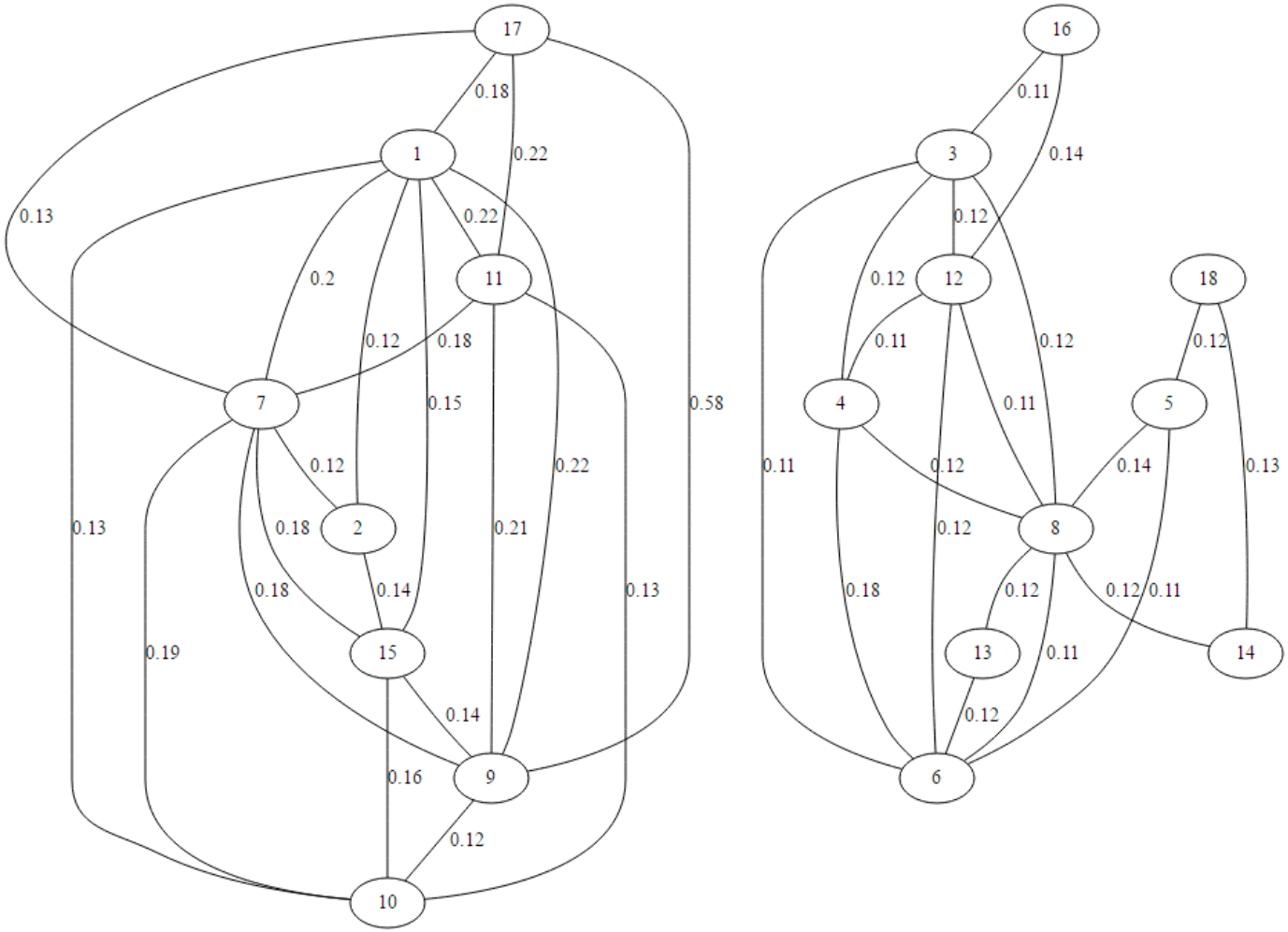
- Количество сообществ на миноритарном языке не зависит линейно он количества пользователей, «говорящих» на этом языке;
- Тексты на миноритарном языке короче текстов на русском: в среднем текст на миноритарном языке состоит из 5 слов, а на русском – из 13;
- Треть всех текстов сообщества написана на миноритарном языке;

Длина текстов

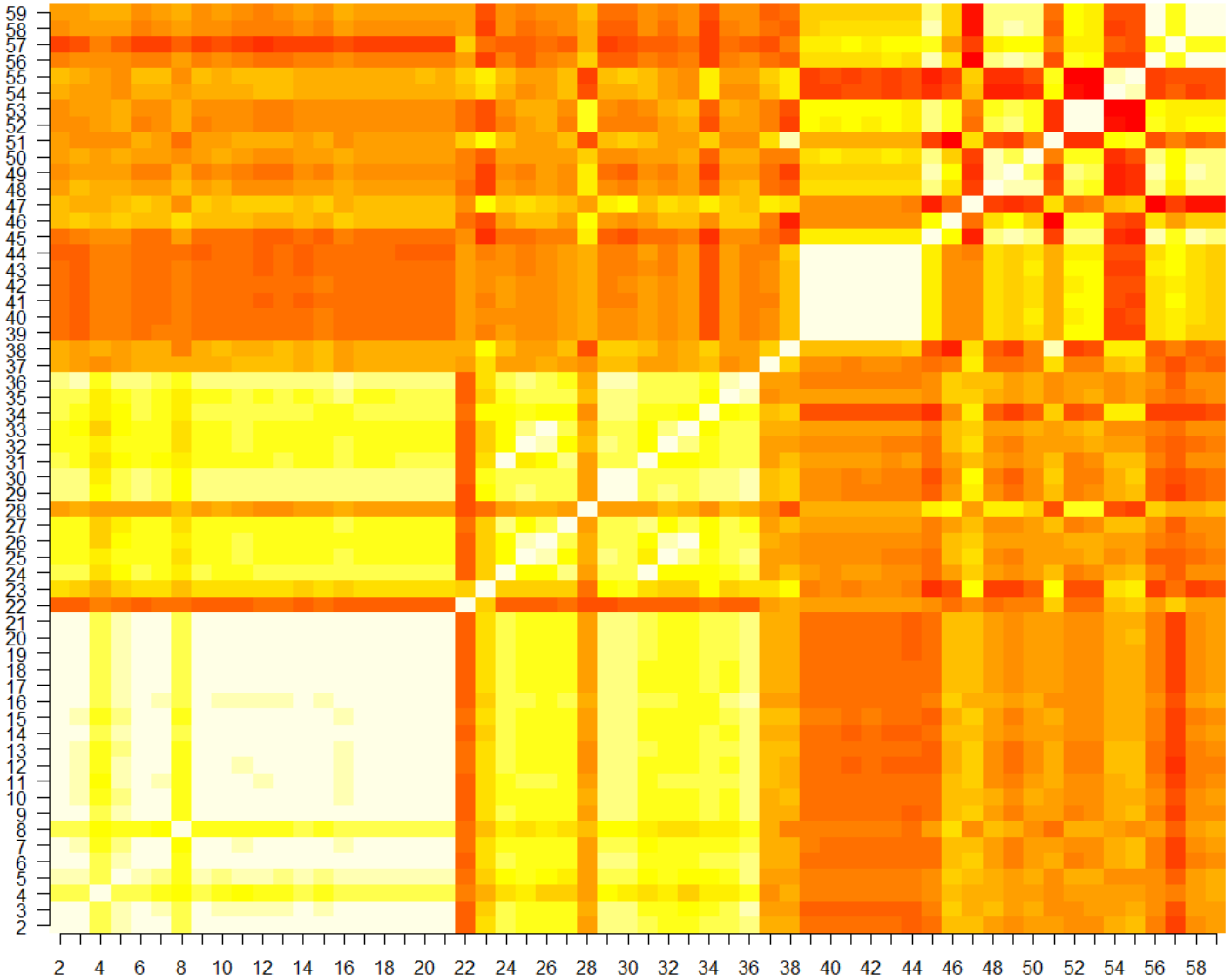


Тематическое моделирование. На примере чеченских сообществ

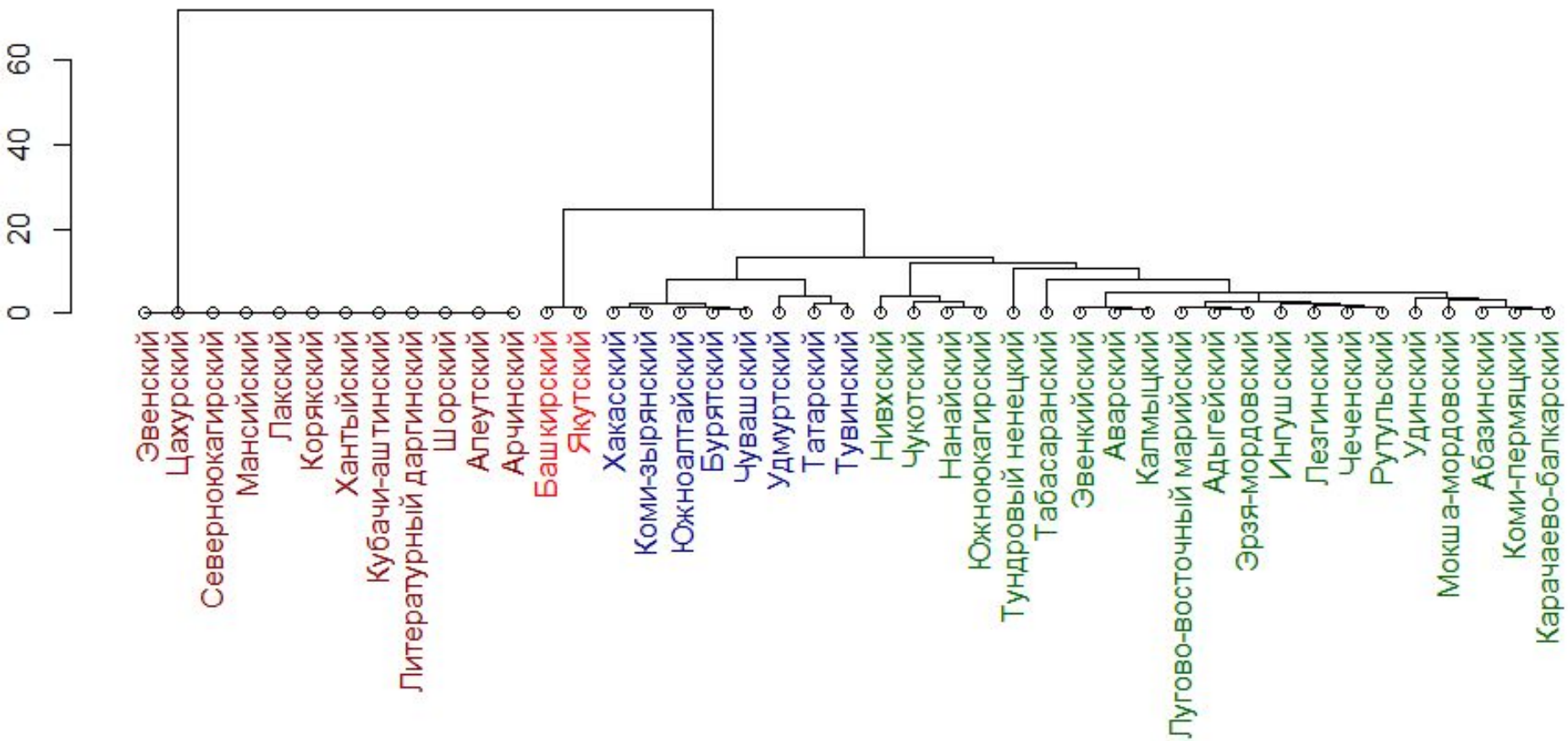
- Узлы – сообщества, «говорящие» на чеченском языке;
- Близость двух сообществ;
- Степень близости.



Зависимости между признаками



Иерархическая кластеризация языков



Портрет типичного носителя

Типичный пользователь-носитель миноритарного языка:

- Возраст: 19-31;
- Место проживания: титульный регион;
- Являясь носителем миноритарного языка, в основном «говорит» по-русски;
- Вероятнее всего, принимает участие в жизни только одного сообщества.

Типичное сообщество, «говорящее» на миноритарном языке:

- Треть всех постов написана на миноритарном языке;
- Длина текста на миноритарном языке в среднем около пяти слов.