Кластеризация корпуса анекдотов

Тамара Здорова, БКЛ153,
2017 год
Научный руководитель:
Б. В. Орехов

Цель работы:

Сравнительный анализ интуитивной категориализации по персонажам и кластеризации, произведенной при помощи компьютера.

Этапы работы:

1. Изучение тем, которые поднимаются авторами статей про анекдоты
2. Создание корпуса анекдотов, имеющего разметку по персонажам
3. Лемматизация текстов
4. Векторизация
5. Кластеризация разными методами:
- KMeans
- MeanShift
- DBSCAN
6. Визуализация
7. Анализ полученных результатов

Наблюдения:

1. Анекдоты про таких персонажей, как Путин, Медведев и Обама, часто оказывались в одном кластере. Во-первых, все они - политики. Во-вторых, нередко эти имена упоминаются вместе, например:

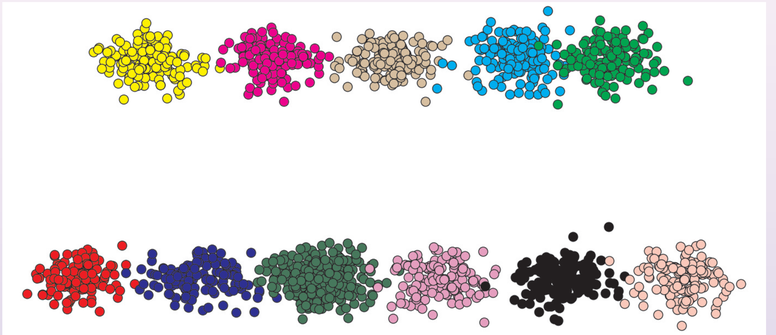
Пришел Обама в Кремль на переговоры с Медведевым. Ждет-ждет, а Медведева
все нет. Вдруг открываются двери, вбегает запыхавшийся Медведев.
Обама:
- Дмитрий, что случилось?
- Извини - в пробке стоял! Путин утром ехал на работу, - дорогу перекрыли!

2. Очень хорошо программа выделяла в отдельные кластеры Штирлица и Ржевского. Скорее всего, это связано с особой лексикой в таких анекдотах. Например, у Штирлица:

У Штирлица вблизи здания Рейхсканцелярии сломалась машина, он вышел и стал копаться в моторе.
- Штирлиц, вы - русский разведчик, - сказал проходивший мимо Мюллер. - Истинный ариец непременно обратился бы в автосервис.

Итоги:

Лучше всего с задачей справился алгоритм KMeans - кластеризация, полученная в результате его использования, довольно близка к исходной категоризации. Другие методы показали плохой результат.



Литература:

Д. Д. Амоголонова. Современный бурятский анекдот в конструировании этносферы. // Этнографическое обозрение. М., 2008. С. 159-170.
А. А. Демичев. Юридический анекдот XVIII - первой половины XIX как источник по истории российской ментальности. // Российская история. М., 2010, С. 137-153.
Л. Сюэфен. Языковая игра в русском анекдоте. // Русская речь. М., 2007. С. 56-60.
N. Thielemann. Arguing by 'Anekdot' - Humorous Accounts in Russian Media Interaction. // Zeitschrift für Slavische Philologie. 2010. P. 185-215.